Fraunhofer
IAIS

FRAUNHOFER-INSTITUT FÜR INTELLIGENTE ANALYSE- UND INFORMATIONSSYSTEME IAIS

STUDIE ZUM TECHNISCHEN JUGENDMEDIENSCHUTZ:

MÖGLICHKEITEN UND GRENZEN VON VERFAHREN ZUR DETEKTION JUGENDSCHUTZRELEVANTER WEB-INHALTE



Kontaktadresse

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven
53757 Sankt Augustin
Telefon 02241 14-3000
Telefax 02241 14-4-3000
E-Mail info@iais.fraunhofer.de
URL http://www.iais.fraunhofer.de

Druck und Weiterverarbeitung

IRB Mediendienstleistungen
Fraunhofer-Informationszentrum Raum und Bau IRB, Stuttgart

Für den Druck des Buches wurde chlor- und säurefreies Papier verwendet.

Die in der vorliegenden Studie enthaltenen Abbildungen dienen ausschließlich der Verdeutlichung von wissenschaftlichen Ergebnissen.

Autoren in alphabetischer Reihenfolge

Dr. Rolf Bardeli
Sven Becker
Dr. André Bergholz
Inke Kolb
Hannes Korte
Wout Maaskant
Dr. Gerhard Paaß
Dr. Daniel Schneider
Dr. Daniel Stein
Sebastian Tschöpel

Projektbeirat

Der Beauftragte der Bundesregierung für Kultur und Medien (BKM)
Bundesministerium für Familie, Senioren, Frauen und Jugend (BMFSFJ)
Kommission für Jugendmedienschutz (KJM)
jugendschutz.net
Stefan Schellenberg, Online Management Kontor

Verschiedene Institutionen haben mit ihren Beiträgen und Kommentaren zum Gelingen der vorliegenden Studie beigetragen. Hierzu zählen insbesondere: Centre for Research and Technology Hellas (CERTH), Deutsche Telekom AG, Institut national de recherche en informatique et en automatique (INRIA), International Audio Laboratories Erlangen, iRights.info sowie die Universität des Saarlandes.

Gefördert durch:



Der Beauftragte der Bundesregierung
für Kultur und Medien

VORWORT

Jugendschutzfilter nehmen eine wichtige ergänzende Rolle im technischen Jugendmedienschutz ein. Durch ihren Einsatz sollen Kinder seltener mit einem für ihre Altersklasse ungeeigneten Inhalt konfrontiert werden – insbesondere dann, wenn ein Kind nicht mit einer solchen Konfrontation rechnet. Äußerst drastische Inhalte wie reale Selbstmord- oder Enthauptungsvideos, die auch Erwachsene nachhaltig verstören können, stehen dabei besonders im Fokus.

Ziel der Studie

Angesichts stetig wachsender Datenmengen im Internet sind rein redaktionell gepflegte Filterlisten nicht für die effektive und umfassende Filterung von unerwünschten Inhalten durch Jugendschutzfilter geeignet. Ziel dieser Studie ist es, Verfahren zur automatischen Detektion von jugendschutzrelevanten Web-Inhalten zu bewerten, die zur automatischen Erstellung und Pflege solcher Filterlisten eingesetzt werden können.

Dabei soll überprüft werden, ob ein Verfahren solche Inhalte detektieren kann, die nach den aktuellen gesetzlichen Bestimmungen für bestimmte Altersklassen nicht geeignet sind. Die Auswahl und Bewertung der Verfahren orientiert sich an den Gesetzesvorgaben und bestehenden Kriterienkatalogen zur redaktionellen Klassifikation von Web-Inhalten.

Hierzu wurden dem Fraunhofer IAIS im Rahmen der vorliegenden Studie ausgewählte jugendschutzrelevante Web-Inhalte aus den Bereichen Gewalt, Extremismus und Selbstgefährdung zur Verfügung gestellt. Diese Bereiche werden bisher von Jugendschutzprogrammen nur unzureichend berücksichtigt [jugendschutz.net, 2011]. Die Daten stammen direkt aus der täglichen Prüfpraxis von jugendschutz.net. Mit Hilfe der Daten wurden dann vielversprechende Verfahren experimentell untersucht und bewertet.

Neben der messbaren Genauigkeit der Verfahren spielt auch ihre Umsetzbarkeit in die Praxis eine wichtige Rolle. Hierzu wurden notwendige Maßnahmen zur Integration der Verfahren in die bestehenden Strukturen betrachtet.

Konkrete Handlungsempfehlungen, die auf der Basis dieser Untersuchungen formuliert wurden, sollen zur Weiterentwicklung des technischen Jugendmedienschutzes in Deutschland beitragen.

Aufbau der Studie

Die Studie ist in drei Abschnitte unterteilt, die im Folgenden kurz beschrieben werden.

Abschnitt I: Rahmenbedingungen für technische Verfahren im Jugendmedienschutz (Kapitel 1 bis 3)

Kapitel 1 gibt einen detaillierten Überblick über den gegenwärtigen Status Quo des technischen Jugendmedienschutzes in Deutschland. Kapitel 2 beschreibt Anforderungen an technische Verfahren zur Detektion jugendschutzrelevanter Web-Inhalte. Ziel von Kapitel 3 ist eine möglichst umfassende Charakterisierung von jugendschutzrelevanten Web-Inhalten, die potentiell detektiert werden müssen.

Abschnitt II: Technische Verfahren im Jugendmedienschutz (Kapitel 4 und 5)

In Kapitel 4 erfolgt zunächst eine begründete Auswahl von Verfahren, die potentiell in der Lage sind, die in Kapitel 3 beschriebenen Charakteristika zu detektieren. Dann wird der Stand der Technik der entsprechenden Verfahren im Kontext des Jugendmedienschutzes beschrieben. Kapitel 5 beinhaltet eine umfassende experimentelle Bewertung von ausgewählten Verfahren.

Abschnitt III: Wege in die praktische Umsetzung (Kapitel 6 und 7)

In Kapitel 6 werden mögliche Wege zur Integration der vielversprechendsten Verfahren in bestehende Strukturen und Programme aufgezeigt. Dabei stehen praktische Aspekte der Realisierung im Vordergrund. Kapitel 7 schließt mit konkreten Handlungsempfehlungen, die sich direkt aus den Beobachtungen in den einzelnen Abschnitten ergeben.

INHALTSVERZEICHNIS

Vorwort **i**

I Rahmenbedingungen für technische Verfahren im Jugendmedienschutz 1

1 Technischer Jugendmedienschutz in Deutschland 3
1.1 Rechtliche Rahmenbedingungen . . . 4
1.1.1 Verfassungsrechtliche Grundlagen des Jugendschutzes . . . 4
1.1.2 Strafgesetzbuch (StGB) . . . 4
1.1.3 Jugendschutzgesetz (JuSchG) . . . 5
1.1.4 Jugendmedienschutz-Staatsvertrag (JMStV) . . . 5
1.1.5 Jugendschutzrichtlinien (JuSchRiL) . . . 9
1.1.6 Netzregeln der KJM . . . 10
1.1.7 UN-Kinderrechtskonvention . . . 10
1.2 Organisationen und Beteiligte . . . 11
1.2.1 Behörden . . . 11
1.2.2 Institutionen des Jugendmedienschutzes . . . 15
1.2.3 Initiativen für den Jugendmedienschutz . . . 22
1.2.4 Initiativen von Unternehmen der freien Wirtschaft im Bereich Selbstkontrolle 25
1.3 Praktische Umsetzung des Jugendmedienschutzes in Deutschland . . . 27
1.3.1 Beurteilung von Angeboten . . . 28
1.3.2 Technische Zugangssysteme im Sinne des Jugendmedienschutzes . . . 29
1.4 Aktuelle Jugendschutzfilter . . . 30
1.4.1 KJM-anerkannte Jugendschutzprogramme . . . 32
1.4.2 Jugendschutzfilter mit SIP-Bench Bewertung . . . 37
1.4.3 Weitere Jugendschutzfilter . . . 40
1.5 Kinder- und Jugendmedienschutz aus Sicht der Eltern . . . 42
1.6 Jugendmedienschutz aus Sicht von Kindern und Jugendlichen . . . 44
1.7 Zusammenfassung . . . 45

2 Anforderungen an technische Verfahren zur Detektion jugendschutzrelevanter Web-Inhalte 47
2.1 Funktionale Anforderungen an Klassifikationsverfahren . . . 49
2.1.1 Arten und Kategorien von Web-Inhalten . . . 49
2.1.2 Altersdifferenzierte Klassifikation . . . 52
2.1.3 Besondere Aspekte bei der Klassifikation von Web-Inhalten . . . 61
2.1.4 Zusammenfassung: funktionale Anforderungen an Klassifikationsverfahren . . . 64
2.2 Nicht-funktionale Anforderungen an Klassifikationsverfahren . . . 64
2.2.1 Zuverlässigkeit . . . 65
2.2.2 Korrektheit . . . 66
2.2.3 Benutzbarkeit und Gebrauchstauglichkeit . . . 66

2.2.4 Leistung und Effizienz . . . 67
2.2.5 Betrieb und Umgebungsbedingungen . . . 67
2.2.6 Wartbarkeit . . . 71
2.2.7 Portierbarkeit . . . 72
2.2.8 Flexibilität . . . 74
2.2.9 Skalierbarkeit . . . 75
2.2.10 Sicherheit der technischen Verfahren vor Manipulation durch Dritte . . . 75
2.2.11 Randbedingungen . . . 76
2.2.12 Zusammenfassung: nicht-funktionale Anforderungen an Klassifikationsverfahren . . . 76
2.3 Exkurs: Anforderungen an Jugendschutzsoftware . . . 77
2.3.1 Quellen für Anforderungen . . . 77
2.3.2 Systematisierung der Anforderungen . . . 78
2.3.3 Zusammenfassung: Anforderungen an technische Verfahren im Kontext von Jugendschutzsoftware . . . 83
2.4 Exkurs: Anforderungen an technische Jugendschutzstrukturen . . . 84
2.4.1 SafeSearch-Funktionen . . . 85
2.4.2 Kinder- und Jugendschutzfunktionen für einzelne Angebote und Anwendungen im WWW . . . 85
2.4.3 Bestehende Altersklassifikationssysteme im WWW . . . 86
2.4.4 Beschwerdestellen für Inhalte im WWW . . . 87
2.4.5 Zusammenfassung: Anforderungen an technische Verfahren im Kontext der technischen Jugendschutzstrukturen . . . 88
2.5 Zusammenfassung . . . 88

3 Charakterisierung jugendschutzrelevanter Web-Inhalte 89
3.1 Bestehende Kriterienkataloge zur redaktionellen Inhalteklassifikation . . . 89
3.1.1 Kommission für Jugendmedienschutz der Landesmedienanstalten (KJM) . . . 89
3.1.2 Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e. V. (FSM) . . . 90
3.1.3 Liesching: Schutzgrade im Jugendmedienschutz . . . 92
3.1.4 Kriterienkataloge vergleichbarer Medien . . . 92
3.2 Charakterisierung von Web-Inhalten im Bereich Gewalt . . . 94
3.2.1 Definitionen . . . 94
3.2.2 Kriterien und Altersklassifikation . . . 96
3.2.3 Beschreibung der exemplarischen Daten im Bereich Gewalt . . . 98
3.2.4 Zusammenfassung: Charakterisierung von Web-Inhalten im Bereich Gewalt . 100
3.3 Charakterisierung von Web-Inhalten im Bereich Extremismus . . . 100
3.3.1 Definitionen . . . 100
3.3.2 Kriterien und Altersklassifikation . . . 103
3.3.3 Beschreibung der exemplarischen Daten im Bereich Extremismus . . . 104
3.3.4 Zusammenfassung: Charakterisierung von Web-Inhalten im Bereich Extremismus . . . 107
3.4 Charakterisierung von Web-Inhalten im Bereich Selbstgefährdung . . . 107
3.4.1 Definitionen . . . 107
3.4.2 Kriterien und Altersklassifikation . . . 108
3.4.3 Beschreibung der exemplarischen Daten im Bereich Selbstgefährdung . . . 111
3.4.4 Zusammenfassung: Charakterisierung von Web-Inhalten im Bereich Selbstgefährdung . . . 113
3.5 Zusammenfassung . . . 114

II Technische Verfahren im Jugendmedienschutz 117

4 Analyse technischer Verfahren zum Jugendmedienschutz 119
4.1 Auswahl technischer Verfahren auf Grundlage von redaktionell klassifizierten Daten . 120
4.1.1 Unterscheidung: Detektion identischer und ähnlicher Dokumente 120
4.1.2 Auswahl von Verfahren zur Detektion von Web-Inhalten im Bereich Gewalt . . 121
4.1.3 Auswahl von Verfahren zur Detektion von Web-Inhalten im Bereich Extremismus 123
4.1.4 Auswahl von Verfahren zur Detektion von Web-Inhalten im Bereich Selbstgefährdung 125
4.2 Ziel, Vorgehensweise und Qualitätsbewertung eines Klassifikators 128
4.2.1 Bewertung eines Klassifikators 129
4.2.2 Überanpassung 130
4.2.3 Kreuzvalidierung 132
4.3 Vorverarbeitung von Webseiten: Identifikation zu klassifizierender Objekte 132
4.3.1 Datennormalisierung 133
4.3.2 Extraktion von HTML-Merkmalen 134
4.4 Technische Verfahren für die Textanalyse 135
4.4.1 Vorverarbeitung von Texten 135
4.4.2 Wichtige Verfahren zur Klassifikation von Textdokumenten 136
4.4.3 Unterstützende Modelle 145
4.4.4 Frei verfügbare Verfahren zur Textanalyse 148
4.4.5 Vergleich von Textklassifikatoren 151
4.4.6 Aktives Lernen 153
4.5 Technische Verfahren für die Multimedia-Analyse 154
4.5.1 Fingerprinting zur Wiedererkennung bekannter Inhalte 156
4.5.2 Automatische Spracherkennung 166
4.5.3 Optische Texterkennung 171
4.5.4 Logodetektion 175
4.5.5 Konzepterkennung 178
4.6 Umfeldanalyse: Klassifikation einer Webseite im Kontext ihres Umfelds 182
4.7 Fusion von Klassifikationsergebnissen 184
4.8 Zusammenfassung 185

5 Bewertung ausgewählter, verfügbarer Verfahren 187
5.1 Zielsetzung und allgemeines Vorgehen 187
5.2 Kategorieübergreifende Textanalyse-Verfahren 190
5.2.1 Keyword-Extraktion 190
5.2.2 Themenmodelle 194
5.2.3 Zusammenfassung 196
5.3 Betrachtung von Verfahren zur Klassifikation im Bereich Gewalt 197
5.3.1 Am Beispiel drastische Gewalt 197
5.3.2 Fingerprinting bei Videoschnitten: Braindead 201
5.3.3 Am Beispiel von Gewaltspielen 208
5.3.4 Spracherkennung: Let's play-Videos 211
5.3.5 Brutalitätserkennung: Robbenjagd 213
5.3.6 Textklassifikation 215
5.3.7 Zusammenfassung: Betrachtung von Verfahren im Bereich Gewalt 219
5.4 Betrachtung von Verfahren zur Klassifikation im Bereich Extremismus 220
5.4.1 Am Beispiel Geschichtsrevisionismus 221
5.4.2 Textanalyse: Rechtsextremes Wiki 222
5.4.3 Textklassifikation 224
5.4.4 Analyse eingehender Links 226
5.4.5 Musikklassifikation: Identifikation bekannter extremistischer Lieder 226

5.4.6 Musikklassifikation: Live-Versionen, Cover und Parodien . . . 228
5.4.7 Logoerkennung: Objekte und Symbole . . . 229
5.4.8 Spracherkennung: Holocaust-Leugnung . . . 232
5.4.9 Zusammenfassung: Betrachtung von Verfahren im Bereich Extremismus . . . 235
5.5 Betrachtung von Verfahren zur Klassifikation im Bereich Selbstgefährdung . . . 237
5.5.1 Am Beispiel Ritzen und Pro-Ana . . . 237
5.5.2 Musikcluster: Ritzvideos . . . 238
5.5.3 Erkennung von Text in Bildern . . . 240
5.5.4 Konzepterkennung Ritzbild . . . 240
5.5.5 Fingerprinting auf Bildsammlungen im Anorexie-Kontext . . . 246
5.5.6 Textklassifikation Pro-Ana . . . 247
5.5.7 Zusammenfassung: Betrachtung von Verfahren im Bereich Selbstgefährdung . 252
5.6 Abschließende Bewertung: Fragen und Antworten . . . 252
5.6.1 Fragestellungen aus dem Bereich Text . . . 252
5.6.2 Fragestellungen aus dem Bereich Multimedia . . . 254
5.7 Zusammenfassung . . . 256

III Wege in die praktische Umsetzung 259

6 Integration technischer Verfahren 261
6.1 Betriebsumgebungen für technische Verfahren . . . 261
6.1.1 Betrieb beim Nutzer . . . 263
6.1.2 Betrieb als Proxy- oder Cloud-Dienst in einer skalierbaren Serverumgebung . . 264
6.1.3 Offline-Betrieb . . . 267
6.2 Konzeption und Implementierung von Filterlisten . . . 268
6.2.1 Status Quo . . . 268
6.2.2 Erhöhung der Filtergranularität . . . 269
6.2.3 Filterung von sicheren Verbindungen (HTTPS) . . . 270
6.2.4 Anwendungsumgebungen für Filterlisten . . . 272
6.3 Crowdsourcing im technischen Jugendmedienschutz . . . 274
6.3.1 Definition von Crowdsourcing . . . 275
6.3.2 Anwendungsszenarien . . . 276
6.3.3 Risiken . . . 279
6.4 Teilautomatische Qualitätssicherung und Priorisierung im Bewertungsprozess . . . 282
6.4.1 Teilautomatische Qualitätssicherung . . . 282
6.4.2 Priorisierung von Bewertungsgegenständen . . . 284
6.5 Zusammenfassung . . . 286

7 Handlungsempfehlungen 289
7.1 Kurzfristig einsetzbare Verfahren . . . 289
7.2 Mittelfristig einsetzbare Verfahren . . . 290
7.3 Empfehlungen für den Einsatz von Filterlisten . . . 290
7.4 Empfehlungen für die technische Begleitforschung . . . 291

Glossar 293

Abkürzungen 301

Literaturverzeichnis 303

ABBILDUNGSVERZEICHNIS

1.1 Services der USK.online . . . 21
1.2 Screenshot JusProg Jugendschutzprogramm . . . 34
1.3 Screenshot Telekom Kinderschutz Software . . . 36

2.1 Jugendschutztechnologien und Schnittstellen . . . 48
2.2 Altersfreigabekennzeichnung nach FSK . . . 53
2.3 Altersfreigabekennzeichnung nach USK . . . 54
2.4 Altersfreigabekennzeichnung nach PEGI . . . 55
2.5 Inhaltekennzeichnung nach PEGI . . . 55
2.6 Altersfreigabekennzeichnung nach BBFC . . . 56
2.7 Altersfreigabekennzeichnung nach ESRB . . . 57
2.8 Altersfreigabekennzeichnung der MPAA . . . 58
2.9 Altersfreigabekennzeichnung nach CERO . . . 59
2.10 Inhaltekennzeichnung nach CERO . . . 59
2.11 Altersfreigabekennzeichnung nach OFLC . . . 59
2.12 Jugendschutzmechanismen in Infrastrukturkomponenten . . . 69
2.13 Personenidentifikation auf öffentlichen Veranstaltungen . . . 81

4.1 Informationsfluss bei der Anwendung eines Klassifikators . . . 128
4.2 Präzision und Recall der Klassifikation in Abhängigkeit von der Klassifikationsschwelle 130
4.3 Fehlerwerte im Verlauf des Trainings eines Klassifikators . . . 131
4.4 Kreuzvalidierung . . . 133
4.5 Der logistische Klassifikator . . . 137
4.6 Das Neuronale Netz . . . 139
4.7 Einfacher Entscheidungsbaum zur Klassifikation von »jugendschutzrelevant« . . . 140
4.8 Support Vektor Maschine mit einer trennenden Hyperebene . . . 142
4.9 Von AdaBoost erzeugter Klassifikator . . . 144
4.10 Verbindungsstruktur eines Conditional Random Fields mit zwei Zuständen . . . 147
4.11 Klassifikationsergebnisse für den 20-Newsgroups-Korpus mit zwölf verschiedenen Klassifikatoren . . . 149
4.12 Informationsfluss beim Aktiven Lernen. . . . 154
4.13 Klassifikation von Spam- und Phishing-E-Mails . . . 155
4.14 Fingerprint-Extraktion im Frequenzbereich . . . 161
4.15 TRECVID-Evaluationsergebnisse aus dem Jahr 2011 zur Detektion von Videokopien . 165
4.16 Zusammenfassung der NIST-Evaluationsergebnisse zur automatischen Spracherkennung seit 1987 . . . 169
4.17 Ergebnisse einer größeren Evaluation zur automatischen Texterkennung von Adressen auf Postsendungen . . . 174
4.18 Beispiel zur Logodetektion anhand des BelgaLogos-Korpus . . . 177
4.19 Vergleich zweier Verfahren zur Logodetektion auf dem BelgaLogos-Korpus . . . 178
4.20 TRECVID-Evaluationsergebnisse aus dem Jahr 2011 zur Multimedia-Event-Erkennung 181

4.21 Nachbarseiten einer Zielseite gemäß der Struktur der gerichteten Weblinks 183

5.1 Trennschärfe von gewichteten Keywords für »ab 6« und »ab 18« Inhalten 192
5.2 Verteilung der Textbewertung für »ab 6« und »ab 18« Inhalte, durch eine SVM mit RBF-Kernfunktion 193
5.3 Hürden bei der automatischen Erkennung von Grausamkeit, am Beispiel des Struwwelpeter 200
5.4 Anhand von Audio-Fingerprints aus der deutschen ungeschnittenen Fassung gefundene Abschnitte. 204
5.5 Abgleich der FSK 16 Fassung von »Braindead« mit der ungeschnittenen Fassung . . . 205
5.6 Unterschiedliche Aspekte in den Trailern für den Film »Braindead«, die in den jeweils anderen Sprachen nicht vorkommen. 206
5.7 Unterschiedliche Bildqualität in den Trailern für den Film »Braindead« 207
5.8 Gefundene Treffer von Ausschnitten aus 200 verschiedenen Trailern, in der ungeschnittenen Version 209
5.9 Ergebnisse der Video Copy Detection bei einem Fan-erstellten Video zum Thema »Braindead«. 209
5.10 Unterschiedliche Gewaltdarstellung bei Webseiten zum Thema Robbenjagd 214
5.11 Ergebnisse der Klassifikation bei Coverliedern. 230
5.12 Bewertung von Ähnlichkeiten für 10 Lieder im rechtsextremen Kontext, verglichen mit 102 unpolitischen Liedern 231
5.13 Erkennungsergebnisse auf dem Blood & Honour Logo 233
5.14 Erkennungsergebnisse für den Hakenkreuz-Detektor 234
5.15 Ähnlichkeitsmatrix für Ritzlieder, unter Benutzung von Audio-Fingerprints von pHash. 239
5.16 Screenshots mit Text aus Videos zum Thema Ritzen. 241
5.17 Screenshots mit Text über Bild aus Videos zum Thema Ritzen. 242
5.18 Vergleich von Bildern aus der Kategorie Selbstverletzung und Tattoos, jeweils am Unterarm. 244
5.19 Übersicht über die ausgegebene Wahrscheinlichkeit des Konzepterkenners »Ritzen / Tattoo« nach Kategorie 245
5.20 Archetypische Beispiele von verschiedenen Thinspo-Bildern. 246
5.21 Verschiedene Veränderungen am Bild »Der Maschinist«. 248

6.1 Varianten für den Einsatz von technischen Verfahren über Proxy- oder Cloud-Lösungen.265
6.2 Varianten für den Einsatz von offline betriebenen technischen Verfahren. 268
6.3 Crowdsourcing-Konfiguration: Bewertung von selbstgewählten Webangeboten oder Webseiten. 277
6.4 Crowdsourcing-Konfiguration: Bewertung von nicht klassifizierten Webangeboten oder Webseiten. 278
6.5 Crowdsourcing-Konfiguration: Bewertung von unsicher automatisch klassifizierten Webangeboten oder Webseiten. 278

TABELLENVERZEICHNIS

1.1 Sendezeitschienen nach § 5 des Jugendmedienschutz-Staatsvertrag (JMStV) 22
1.2 Jugendschutzfilter im Testfeld der SIP-BENCH II 4th Cycle Studie 39
1.3 Jugendschutzfilter für Mobilgeräte im Testfeld der SIP-BENCH II 4th Cycle Studie . . . 40

2.1 Altersfreigaben im internationalen Vergleich [FSK, 2011a]. 61

3.1 Bestehende Kriterienkataloge zur redaktionellen Inhalteklassifikation. 90
3.2 Prüfmatrix der FSM 91
3.3 Schutzstufen nach Liesching 92

4.1 Technische Verfahren zur Erkennung von Multimedia-Beiträgen mit Gewalt. 123
4.2 Technische Verfahren zur Erkennung von extremistischen Multimedia-Inhalten 125
4.3 Technische Verfahren zur Erkennung von Multimedia-Inhalten im Bereich Selbstgefährdung 127
4.4 Vergleich von Textklassifikatoren auf zwei Dokumentensammlungen. 151
4.5 Klassifikationsergebnisse für einen öffentlich verfügbaren Spam Corpus bei unterschiedlichen Merkmalen. 152
4.6 Audio-Fingerprinting-Lösungen und ihre Verfügbarkeit 159
4.7 Auswahl von Video-Fingerprinting-Lösungen und ihre Verfügbarkeit 164
4.8 Spracherkennungslösungen und ihre Verfügbarkeit. 171
4.9 Lösungen zur optischen Texterkennung und ihre Verfügbarkeit 173
4.10 Robustheit von Verfahren zur Logo-Erkennung 176

5.1 Übersicht der experimentell ausgewerteten Techniken, nach Themen 189
5.2 Einige beispielhaft ausgewählte Themen aus 32 028 Webseiten der Kategorie »Allgemein, ab 18« bei 100 Themen ohne Sprachfilterung 195
5.3 Einige beispielhaft ausgewählte Themen aus 32 028 Webseiten der Kategorie »Allgemein, ab 18« bei 200 Themen gefiltert nach deutscher Sprache 196
5.4 Einzelbegründungen zum Beschlagnahmungsbeschluss des Filmes »Braindead« . . . 202
5.5 Die 75 ersten Treffer auf die Anfrage »Braindead«, in einem großen Videoportal. . . . 203
5.6 Ergebnisse der Spracherkennung auf den Let's Play Videos. 213
5.7 Die relevantesten Merkmale für harmlose und drastische Gewaltdarstellungen zum Thema »Robbenjagd« 214
5.8 Die relevantesten Sprachmerkmale für Unterkategorien von Gewalt. 216
5.9 Die Klassifikationsergebnisse für das Thema Gewalt, getrennt nach Kategorien und Merkmalsgruppen 217
5.10 Ein Vergleich der Wortverteilungen über die Artikel aus einem rechtsextremen Wiki (R) gegenüber Artikeln zum Thema »Nationalsozialismus« der Wikipedia (W). 223
5.11 Die Klassifikationsergebnisse im Bereich Rechtsextremismus, für die Zielkategorien gegenüber *Allgemein ab 6*, getrennt nach Merkmalsgruppen. 225
5.12 Eine Stichprobe zur Frage, wie viele nach »ChillingEffects.org« als illegal bekannte Inhalte auf gesperrte rechtsextreme Inhalte verweisen. 227

5.13 Audio-Fingerprint Experiment auf rechtsextremen Liedern. 228
5.14 Ausgewählte Coverlieder für die Musikklassifikation. 228
5.15 Durchschnittliche Präzision und deren Mittelwert auf alle Anfragen für die 10 Liederanfragen bezüglich der drei Datensätze 231
5.16 Ergebnisse der Spracherkennung auf Holocaust-Leugnungsvideos. 236
5.17 Schwierigkeiten der optischen Texterkennung bei verschiedenen Farbtypen, Schriftarten und unerwarteter Orthographie. 243
5.18 Ergebnisse von pHash Bilder Fingerprinting auf Christian Bale »Der Maschinist« . . . 247
5.19 Die Klassifikationsergebnisse für die Zielkategorie *Pro-Ana* gegenüber *Allgemein ab 6* und *Pro-Ana-Information* getrennt nach Merkmalsgruppen. 251

6.1 Technische Verfahren und ihr Potenzial für Echtzeitanalyse 263

TEIL I

RAHMENBEDINGUNGEN FÜR TECHNISCHE VERFAHREN IM JUGENDMEDIENSCHUTZ

KAPITEL 1

TECHNISCHER JUGENDMEDIENSCHUTZ IN DEUTSCHLAND

> »Der Jugendschutz umfasst die Gesamtheit der staatlichen Maßnahmen zum Schutz von Kindern und Jugendlichen vor Gefährdungen für ihr körperliches, geistiges und seelisches Wohl. Als Teilbereich davon greift der Jugendmedienschutz das besondere Gefährdungspotential auf, das für Kinder und Jugendliche bei der Nutzung von Medien besteht.« [Eifler, 2011]

Den Auftrag zum Schutz der Heranwachsenden hat der deutsche Staat in seinem föderalistischen System durch eine Reihe von gesetzlichen Grundlagen konkretisiert und verschiedenen Stellen und Einrichtungen die Verantwortung für die Umsetzung des Jugendmedienschutzes übertragen.

Dabei setzt der Jugendmedienschutz auf zwei wesentliche Komponenten:

1. eine ordnungspolitisch-schützende Komponente – aufbauend auf der freiwilligen Selbstkontrolle der Anbieter von Medien – mit gesetzlich festgelegten Verboten, der Möglichkeit der Sanktionierung von Verstößen und technischen Mitteln zur Durchsetzung der gesetzlichen Regelungen und
2. eine pädagogisch-fördernde Komponente, die auf die Förderung der Medienkompetenz der Heranwachsenden (und deren Eltern) durch Bildung und Aufklärung setzt.

Dieses Kapitel gibt einen detaillierten Überblick über den gegenwärtigen Status Quo des technischen Jugendmedienschutzes in Deutschland. Die Darstellung umfasst dabei eine Reihe von sowohl rechtlichen, organisatorischen als auch praktischen Aspekten.

Zunächst werden in Abschnitt 1.1 die in Deutschland zugrunde liegenden rechtlichen Rahmenbedingungen des technischen Jugendmedienschutzes dargestellt. Im darauffolgenden Abschnitt 1.2 werden die am technischen Jugendmedienschutz beteiligten organisatorischen Einrichtungen vorgestellt. Der Abschnitt umfasst dabei sowohl offizielle Stellen und Behörden des Bundes und der Länder, die an der Durchführung des Jugendmedienschutzes operativ beteiligt sind, als auch andere Institutionen, die sich dezidiert mit dem Jugendmedienschutz beschäftigen. Es folgt eine Darstellung von weiteren Initiativen, die sich in Deutschland für den Jugendmedienschutz engagieren. Anschließend werden in Abschnitt 1.3 einige Aspekte der technischen Umsetzung des Jugendmedienschutzes betrachtet, die für die vorliegende Studie von besonderer Bedeutung sind. Der Abschnitt 1.4 gibt einen Überblick über aktuell verfügbare Jugendschutzfilter. Ein Augenmerk wird dabei sowohl auf existierende Möglichkeiten der Zertifizierung beziehungsweise Anerkennung durch offizielle Stellen als auch auf Studien zur Bewertungen von Jugendschutzfiltern gelegt. Abschließend geben Abschnitt 1.5 einen Einblick in die Sicht der Eltern und Abschnitt 1.6 in die Sicht der Kinder und Jugendlichen auf den Jugendmedienschutz.

1.1 Rechtliche Rahmenbedingungen

Der Jugendmedienschutz stützt sich auf die staatlich geschaffenen rechtlichen Rahmenbedingungen. Diese werden nachfolgend näher betrachtet und deren jeweilige Bedeutung für den Jugendmedienschutz dargestellt.

1.1.1 Verfassungsrechtliche Grundlagen des Jugendschutzes

Seine Rechtsgrundlagen findet der Jugendschutz im Grundgesetz (GG) [Grundgesetz, 2012]. Das Bundesverfassungsgericht hat dem Jugendschutz Verfassungsrang zuerkannt[1] [Eifler, 2011]. Dieser wird abgeleitet aus dem allgemeinen Persönlichkeitsrecht von Kindern und Jugendlichen, das auch die ungestörte Entwicklung der Persönlichkeit umfasst (Artikel 2 Absatz 1[2] i. V. m. Artikel 1 Absatz 1[3] GG). Hieraus leitet sich ein verfassungsrechtlicher Auftrag an den Staat ab, »Kinder und Jugendliche so zu schützen, dass sie sich zu eigenständigen, sozial verantwortlichen Persönlichkeiten entwickeln können« [Schulz, 2004]. Darüber hinaus muss der Staat den primär bei den Eltern liegenden Erziehungsauftrag im Sinne des Artikel 6 Absatz 2 GG[4] stützen.

Es ist also Verpflichtung des Staates, die Entwicklung der Persönlichkeit Minderjähriger zu schützen. »Wie der Staat diese Aufgabe bewältigt, bleibt grundsätzlich seiner Einschätzung überlassen […]. Aus diesem Grund konnten sich zum Jugendmedienschutz drei verschiedene Konzepte in Bezug auf drei Medientypen herausbilden, mit denen der Staat den Gefährdungen beim Konsum von Medien begegnen will. Normiert wurden sie von zwei Normgebern in zwei Regelwerken - im Jugendschutzgesetz des Bundes sowie im Jugendmedienschutz-Staatsvertrag der Bundesländer« [Eifler, 2011], denn im Sinne des Grundgesetzes unterliegt der Jugendschutz der konkurrierenden Gesetzgebungskompetenz von Bund und Ländern (Artikel 74 Absatz 1 Nummer 7 GG[5]) [Schulz, 2004]. So gelten differenziert nach den Verbreitungswegen für Medien und Medieninhalte heute das Bundesrecht in Form des Jugendschutzgesetz des Bundes (JuSchG) für Trägermedien und Länderrecht in Form des JMStV für die Verbreitung von Medien über Rundfunk und Telemedien.

1.1.2 Strafgesetzbuch (StGB)

Einen ersten Beitrag zum Jugendschutz leisten bereits Verbreitungsverbote nach dem Strafgesetzbuch (StGB) [StGB, 2012] durch allgemeine strafrechtliche Vorschriften, die nicht nach unterschiedlichen Verbreitungswegen differenzieren. Im Rahmen des Jugendmedienschutzes werden entsprechende Angebote als *absolut unzulässig* bezeichnet. Im Kontext des StGB beziehen sich die absolut unzulässigen Medienangebote auf:

- § 86 Verbreiten von Propagandamitteln verfassungswidriger Organisationen
- § 86a Verwenden von Kennzeichen verfassungswidriger Organisationen
- § 126 Störung des öffentlichen Friedens durch Androhung von Straftaten (Absatz 1)

[1] BVerfG v. 27.11.1990, BVerfGE 83, 130 (139 f.)

[2] Artikel 2 (1) GG: »Jeder hat das Recht auf die freie Entfaltung seiner Persönlichkeit, soweit er nicht die Rechte anderer verletzt und nicht gegen die verfassungsmäßige Ordnung oder das Sittengesetz verstößt.« [Grundgesetz, 2012]

[3] Artikel 1 (1) GG: »Die Würde des Menschen ist unantastbar. Sie zu achten und zu schützen ist Verpflichtung aller staatlichen Gewalt.« [Grundgesetz, 2012]

[4] Artikel 6 (2) GG: »Pflege und Erziehung der Kinder sind das natürliche Recht der Eltern und die zuvörderst ihnen obliegende Pflicht. Über ihre Betätigung wacht die staatliche Gemeinschaft.« [Grundgesetz, 2012]

[5] Artikel 74 (1) Nummer 7 GG: »Die konkurrierende Gesetzgebung erstreckt sich auf folgende Gebiete: […] die öffentliche Fürsorge (ohne das Heimrecht)« [Grundgesetz, 2012]

- § 130 Volksverhetzung
- § 130a Anleitung zu Straftaten
- § 131 Gewaltdarstellung
- § 184a Verbreitung gewalt- oder tierpornographischer Schriften
- § 184b Verbreitung, Erwerb und Besitz kinderpornographischer Schriften
- § 184c Verbreitung, Erwerb und Besitz jugendpornographischer Schriften

1.1.3 Jugendschutzgesetz (JuSchG)

Das JuSchG [JuSchG, 2008] regelt im Wesentlichen den Jugendschutz in der Öffentlichkeit und Verbreitungsbeschränkungen bei jugendgefährdenden Trägermedien (Printmedien, Videos, CD-ROMs, DVDs usw.). Das Gesetz umfasst unter anderem eine verbindliche Verpflichtung zur Prüfung und Kennzeichnung der Altersbeschränkung für Filme, Computer- und Videospiele. Diese müssen, wenn sie an Minderjährige verkauft werden sollen, der obersten Landesbehörde oder einer Organisation der freiwilligen Selbstkontrolle vorgelegt werden, die in einem Prüfverfahren die Altersfreigabe festlegen. Ebenfalls definiert das JuSchG den Auftrag der Bundesprüfstelle für jugendgefährdende Medien (BPjM) (§§ 17–25 JuSchG), der in der *Verordnung zur Durchführung des Jugendschutzgesetzes (DVO-JuSchG)* [DVO JuSchG, 2003] konkretisiert wird.

1.1.4 Jugendmedienschutz-Staatsvertrag (JMStV)

»Mit dem Jugendmedienschutz-Staatsvertrag (JMStV) der Länder wurde eine einheitliche Rechtsgrundlage für den Jugendschutz in den elektronischen Medien (z. B. Internet, Fernsehen, Hörfunk) geschaffen. Ziel des Jugendmedienschutz-Staatsvertrag ist der einheitliche Schutz der Kinder und Jugendlichen vor Angeboten in elektronischen Informations- und Kommunikationsmedien, die deren Entwicklung oder Erziehung beeinträchtigen oder gefährden, sowie der Schutz vor solchen Angeboten in elektronischen Informations- und Kommunikationsmedien, die die Menschenwürde oder sonstige durch das Strafgesetzbuch geschützte Rechtsgüter verletzen.« [KJM, 2012a]

Der JMStV folgt dem Gedanken der Selbstkontrolle der Medien und richtet sich daher an die deutschen Anbieter von Webseiten. Wesentliche Inhalte des JMStV sind unter anderem:

- Regelungen zu unzulässigen Angeboten (§ 4 JMStV)
- Regelungen zu entwicklungsbeeinträchtigenden Angeboten (§ 5 JMStV)
- Jugendschutz in Werbung und Teleshopping (§ 6 JMStV)
- Jugendschutzbeauftragte (§ 7 JMStV)
- Festlegung der Sendezeit, Programmankündigungen und Kenntlichmachung von Sendungen im Rundfunk (§§ 8 ff. JMStV)
- Jugendschutzprogramme und Kennzeichnungspflichten bei Telemedien (§§ 11 f. JMStV)
- Sperrverfügungen (§ 20 Absatz 4 JMStV i. V. m. § 59 Absatz 4 RStV)

Die Einhaltung des JMStV wird durch die zuständige Landesmedienanstalt beziehungsweise durch die *Kommission für Jugendmedienschutz (KJM)* überprüft (vgl. §§ 14 ff. JMStV). Dabei wird sie durch *jugendschutz.net* unterstützt (vgl. § 18 JMStV). Zusätzlich überprüfen *Einrichtungen der Freien Selbstkontrolle* (z. B. die Freiwillige Selbstkontrolle Multimedia-Diensteanbieter (FSM)) die Einhaltung der staatsvertraglichen Bestimmungen (vgl. § 19 JMStV).

1.1.4.1 Kategorien von Angeboten

Durch das im JMStV festgelegte Prüfverfahren (§§ 14 ff. JMStV) werden Angebote nach dem Grad ihrer Eignung zur Jugendgefährdung in drei verschiedene Kategorien eingeteilt:

1. absolut unzulässige Angebote (§ 4 Absatz 1 JMStV),
2. relativ unzulässige Angebote (§ 4 Absatz 2 JMStV) und
3. entwicklungsbeeinträchtigende Angebote (§ 5 JMStV).

Absolut unzulässige Angebote dürfen Kindern und Jugendlichen nicht zugänglich gemacht werden. Relativ unzulässige Angebote dürfen Erwachsenen in geschlossenen Benutzergruppen zugänglich gemacht werden. Entwicklungsbeeinträchtigende Angebote sind in der Regel altersdifferenziert zu betrachten und sind Kindern oder Jugendlichen unterhalb einer bestimmten Altersgrenze nicht zugänglich zu machen.

Im Kontext der vorliegenden Studie werden Inhalte und Angebote als jugendschutzrelevant bezeichnet, wenn sie im Sinne des JMStV einer der drei vorgestellten Kategorien genügen.

Unzulässige Angebote Absolut unzulässige Angebote unterliegen stets einem Verbreitungsverbot in Rundfunk und Telemedien. Unzulässige Angebote im Sinne des JMStV sind:

- Propaganda (i. S. d. JMStV § 4 (1) 1.)
- verfassungswidrige Organisationen (i. S. d. JMStV § 4 (1) 2.)
- Hass (i. S. d. JMStV § 4 (1) 3.)
- Nationalsozialismus (i. S. d. JMStV § 4 (1) 4.)
- Gewalttätigkeit (i. S. d. JMStV § 4 (1) 5.)
- rechtswidrige Taten (i. S. d. JMStV § 4 (1) 6.)
- Kriegsverherrlichung (i. S. d. JMStV § 4 (1) 7.)
- Verstöße gegen die Menschenwürde (i. S. d. JMStV § 4 (1) 8.)
- unnatürliche Abbildungen von Minderjährigen in geschlechtsbetonter Körperhaltung (i. S. d. JMStV § 4 (1) 9.)
- Kinderpornografie (i. S. d. JMStV § 4 (1) 10.)
- Liste jugendgefährdender Medien gem. § 18 JuschG (indiziert durch BPjM)
  - JuSchG § 18 Teil B: in Teil B (Öffentliche Liste der Trägermedien mit absolutem Verbreitungsverbot) sind, soweit sie nicht Teil D zuzuordnen sind, Trägermedien aufzunehmen, die nach Einschätzung der Bundesprüfstelle für jugendgefährdende Medien einen in § 86, § 130, § 130a, § 131, § 184a, § 184b oder § 184c des Strafgesetzbuches bezeichneten Inhalt haben;
  - JuSchG § 18 Teil D: in Teil D (Nichtöffentliche Liste der Medien mit absolutem Verbreitungsverbot) sind diejenigen Trägermedien, die nur deshalb nicht in Teil B aufzunehmen sind, weil bei ihnen von einer Bekanntmachung der Aufnahme in die Liste gemäß § 24 Absatz 3 Satz 2 abzusehen ist, sowie diejenigen Telemedien aufzunehmen, die nach Einschätzung der Bundesprüfstelle für jugendgefährdende Medien einen in § 86, § 130, § 130a, § 131, § 184a, § 184b oder § 184c des Strafgesetzbuches bezeichneten Inhalt haben.

Bei den relativ unzulässigen Angeboten im Sinne des JMStV (§ 4 Absatz 2) handelt es sich um Angebote, die Kindern und Jugendlichen ebenfalls nicht zugänglich gemacht werden dürfen. Als relativ unzulässig werden Angebote betrachtet, wenn sie

1. in sonstiger Weise pornografisch sind,

2. in den Teilen A und C der Liste nach § 18 des Jugendschutzgesetzes aufgenommen sind oder mit einem in diese Liste aufgenommenen Werk ganz oder im Wesentlichen inhaltsgleich sind, oder
3. offensichtlich geeignet sind, die Entwicklung von Kindern und Jugendlichen oder ihre Erziehung zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit unter Berücksichtigung der besonderen Wirkungsform des Verbreitungsmediums schwer zu gefährden.

Eine besondere Regelung sieht der JMStV für pornografische Angebote im Sinne von 1. vor, die auch als *einfache Pornografie* bezeichnet werden. Deren Verbreitung über Telemedien ist zulässig, wenn sicher gestellt wird, dass das Angebot ausschließlich für Erwachsene innerhalb einer »geschlossenen Benutzergruppe« erreichbar ist. Es handelt sich also um eine Ausnahme für die Zulässigkeit unzulässiger Angebote innerhalb geschlossener Benutzergruppen mit vorgeschalteter Identifikation und Authentifizierung der Benutzer.

Entwicklungsbeeinträchtigende Angebote Angebote, die geeignet sind, die Entwicklung von Kindern oder Jugendlichen zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit zu beeinträchtigen, dürfen verbreitet werden, sofern der Anbieter dafür Sorge trägt, dass Kinder oder Jugendliche der betroffenen Altersstufen sie üblicherweise nicht wahrnehmen [JMStV, 2003]. Der JMStV sieht drei Möglichkeiten vor, wie ein Anbieter dieser Sorgfaltspflicht genügen kann:

1. indem der Anbieter durch technische oder sonstige Mittel die Wahrnehmung des Angebots durch Kinder oder Jugendliche der betroffenen Altersstufe unmöglich macht oder wesentlich erschwert,
2. indem der Anbieter die Zeit, in der die Angebote verbreitet oder zugänglich gemacht werden, so wählt, dass Kinder oder Jugendliche der betroffenen Altersstufe üblicherweise die Angebote nicht wahrnehmen[6] oder
3. indem der Anbieter das Angebot für ein als geeignet anerkanntes Jugendschutzprogramm[7] programmiert oder dieses dem Angebot vorgeschaltet wird.

Ist eine entwicklungsbeeinträchtigende Wirkung nur auf Kinder zu befürchten, genügt der Anbieter der Sorgfaltspflicht entsprechend § 5 Absatz 1 JMStV, wenn das Angebot getrennt von Angeboten für Kinder verbreitet wird (§ 5 Absatz 5 JMStV).

1.1.4.2 Altersdifferenzierung entwicklungsbeeinträchtigender Angebote

Nach dem JMStV müssen Angebote hinsichtlich ihrer Eignung (bzw. Gefahr) zur Entwicklungsbeeinträchtigung von Kindern oder Jugendlichen beurteilt werden. Der Begriff *Entwicklungsbeeinträchtigung* ist dabei eine neue Wortschöpfung durch den JMStV und ein unbestimmter Rechtsbegriff, der erst durch die Spruchpraxis mit Leben gefüllt wird. Durch den JMStV ist aber ersichtlich, dass zur Annahme einer Entwicklungsbeeinträchtigung eine niedrigere Schwelle angesetzt werden muss als zur Annahme einer Jugendgefährdung[8] und dass zur Beurteilung allgemeingültige Maßstäbe von Kindern oder Jugendlichen einer gewissen Altersstufe heranzuziehen sind [FSM, 2004a]. Zur Differenzierung der Altersstufen definiert der JMStV, dass ein Kind noch nicht 14 Jahre und ein Jugendlicher mindestens 14 aber noch nicht 18 Jahre alt ist (§ 3 Absatz 1 JMStV) und fordert von Jugendschutzprogrammen die Umsetzung eines nach Altersstufen differenzierten Zugangs, ohne dabei konkrete Altersstufen zu benennen. Dem gegenüber erfolgt die Empfehlung zur altersdiffe-

[6] Dies erfolgt in Anlehnung an das Sendezeitschema für FSK-klassifizierte Filme im Rundfunk (vgl. § 5 Absatz 4 JMStV).
[7] Das Jugendschutzprogramm ist ein technisches Mittel i. S. d. § 5 Absatz 3 Nummer 1 JMStV, siehe hierzu § 11 Absatz 1 JMStV.
[8] i. S. d. § 18 Absatz 1 JuSchG

renzierten Sendezeitbeschränkung in Telemedien in Anlehnung an die Sendezeitregelungen für FSK klassifizierte Filme im Rundfunk entlang der Altersstufen 12, 16 und 18 Jahre (§ 5 Absatz 4 JMStV).

In den *Prüfgrundsätzen der FSM* ([FSM, 2011d], vgl. Abschnitt 9.1.1) betrachtet die FSM systematische Vorüberlegungen für die Definition des Anwendungsbereichs. Hier wird die Definition der *Entwicklungsbeeinträchtigung* durch eine Abgrenzung zu den Bereichen *einfache Entwicklungsgefährdung*, *schwere Entwicklungsgefährdung* und *unzulässige Inhalte* versucht. Aus der Gesetzessystematik ergibt sich, dass diese vier Kategorien in der hier wiedergegebenen Reihenfolge hinsichtlich der möglichen Schädlichkeit eines Angebotes aufeinander aufbauen. Der JMStV hingegen unterscheidet die *entwicklungsbeeinträchtigenden Angebote* (§ 5 JMStV) von *unzulässigen Angeboten* (§ 4 JMStV), die wiederum strafrechtlicht relevant, von der BPjM als jugendgefährdend indiziert oder schwer entwicklungsgefährdend sein können.

Die nähere Betrachtung der im § 5 JMStV formulierten Voraussetzung »Angebote, die geeignet sind, die Entwicklung von Kindern oder Jugendlichen zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit zu beeinträchtigen« [JMStV, 2003] wirft die Frage auf, welche Kriterien für die Eigenverantwortlichkeit und der Gemeinschaftsfähigkeit einer Persönlichkeit zu bewerten sind. Da im JMStV entsprechende Entwicklungsziele einer Persönlichkeit nicht formuliert sind, können allgemein (gesellschaftlich) anerkannte Kriterien zur Definition der Eigenschaften herangezogen werden, deren juristische Relevanz jedoch in Frage zu stellen ist. Hierzu »bedarf es zusätzlich eines Rückgriffs auf die im höherrangigen Recht formulierten Werte. In streng juristischer Hinsicht wird man daher als entwicklungsbeeinträchtigend und/oder entwicklungsgefährdend solche Inhalte ansehen können, die mit im Grundgesetz zum Ausdruck kommenden Werten […] nicht in Einklang zu bringen sind« [FSM, 2011d]. Jedoch erkennt die FSM, dass das Grundgesetz längst nicht alle Arten von Inhalten als gesellschaftliche Werte erfasst und daher als Definitionsgrundlage zu kurz greift.

Aus sozialwissenschaftlicher Sicht liegt eine Entwicklungsbeeinträchtigung vor, wenn die natürliche Entwicklung und Erziehung eines Minderjährigen unterbrochen oder gehemmt wird [FSM, 2011d]. Dies ist in den Grundsätzen der Freiwilligen Selbstkontrolle der Filmwirtschaft (FSK) näher definiert: »Unter Beeinträchtigungen sind Hemmungen, Störungen oder Schädigungen zu verstehen. […] Die Entwicklung von Kindern und Jugendlichen oder ihre Erziehung zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit können insbesondere Filme oder Trägermedien beeinträchtigen, welche die Nerven überreizen, übermäßige Belastungen hervorrufen, die Phantasie über Gebühr erregen, die charakterliche, sittliche (einschl. religiöse) oder geistige Erziehung hemmen, stören oder schädigen oder zu falschen und abträglichen Lebenserwartungen verführen« (vgl. [FSK, 2010] § 18 Absatz 2).

Die Beurteilung, aufgrund welcher Ursache bei einer Person eines bestimmten Alters eine entsprechende Wirkung eintritt, ist aus rechtlicher Sicht nicht zweifelsfrei zu klären. Der JMStV sieht daher ein Angebot bereits als entwicklungsbeeinträchtigend an, wenn aus der Eignung eines Angebots zur Beeinträchtigung der Persönlichkeitsentwicklung eine Gefährdungslage entstehen könnte (§ 5 Absatz 1 JMStV). Für die dementsprechende Bewertung von Angeboten hat die FSM in ihren Prüfgrundsätzen einen umfangreichen Kriterienkatalog aufgestellt (vgl. [FSM, 2011d], Abschnitt 9.2.2), der jedoch nur abschnittsweise themenspezifisch eine Altersdifferenzierung enthält. Der Kriterienkatalog der FSM wird von der vorliegenden Studie in Abschnitt 3.1.2.1 näher betrachtet.

Zur Interpretation des Begriffs *Entwicklungsbeeinträchtigung* im Kontext des JMStV liegen nun einige Quellen und Anwendungsbeispiele vor. Eine gesetzliche und damit juristisch eindeutige Definition existiert nicht. Offensichtlich ist aber, dass eine Beurteilung der Eignung eines Angebots zur Entwicklungsgefährdung grundsätzlich nur im Hinblick auf einen angenommenen Rezipienten eines bestimmten Alters vorgenommen werden kann. Doch zur altersdifferenzierten Beurteilung können

die Jugendschützer bislang nicht auf normative Quellen, sondern zumeist nur auf Erfahrung und etablierte Wertvorstellungen zurückgreifen.

1.1.4.3 Jugendschutzprogramme

Der JMStV zeigt Möglichkeiten auf, wie ein Anbieter jugendschutzrelevante Inhalte, solange sie nicht absolut unzulässig sind, diese Inhalte verbreiten darf und dabei dem Jugendschutz genügt. Neben Sendezeitbeschränkungen oder Altersverifikationssystemen sind dies anerkannte Jugendschutzprogramme. »Existiert ein solches anerkanntes Jugendschutzprogramm, kommt ein Telemedienanbieter eines Angebots, das entwicklungsbeeinträchtigend für Kinder und Jugendliche unter 14 Jahren oder unter 16 Jahren[9] ist, seiner gesetzlichen Pflicht bereits dann nach, wenn er sein Angebot entsprechend labelt,[10] so dass das anerkannte Jugendschutzprogramm diese Anbieterkennzeichnung auslesen und – entsprechend seiner Konfiguration – ggf. ausfiltern kann.« [HBI, 2011]

Die bloße Existenz eines anerkannten Jugendschutzprogramms und die Programmierung eines Angebots für ein solches Jugendschutzprogramm reicht also aus, um den gesetzlichen Anforderungen zu genügen. Anders als beispielsweise bei dem Erwerb jugendschutzrelevanter Genussmittel, Trägermedien oder bei Kinobesuchen muss keine weitergehende Kontrolle stattfinden, ob beim Besuch eines entsprechenden Angebots ein Jugendschutzprogramm aktiviert ist. »Jugendschutzprogramme sind somit in Deutschland einer der gesetzlich vorgesehenen Stützpfeiler von Jugendschutz im Internet. Die Effektivität dieses Schutzmechanismus hängt aber nicht nur von der (Erkennungs-)Qualität der anerkannten Jugendschutzprogramme ab, sondern auch, wenn nicht gar vor allem, von dem Einsatz der Software auf Elternseite« [HBI, 2011], sowie der notwendigen Verbreitung der Jugendschutzprogramme in Deutschland. Die Kommission für Jugendmedienschutz (KJM) bezeichnet dies als eine »Privilegierung des Anbieters. […] Die Verschiebung der Pflichten des Anbieters hin zu der Elternverantwortung bereits mit Anerkennung eines Jugendschutzprogramms, aber vor wirksamer Verbreitung der Jugendschutzprogramme, könnte demzufolge zunächst eine Verschlechterung des Jugendschutzes im Internet bewirken.« [Weigand und Braml, 2012]

1.1.5 Jugendschutzrichtlinien (JuSchRiL)

Die *Gemeinsamen Richtlinien der Landesmedienanstalten zur Gewährleistung des Schutzes der Menschenwürde und des Jugendschutzes* wurden von den Landesmedienanstalten vereinbart und »konkretisieren die gesetzlichen Anforderungen des JMStV und liefern entsprechende Vorgaben und Handlungsanweisungen, soweit Regelungsbedarf gesehen wird« [JuSchRiL, 2005]. Die Jugendschutzrichtlinien (JuSchRiL) unterstreichen sowohl die Eigenverantwortlichkeit der Anbieter zur Berücksichtigung der gesetzlichen Vorgaben zum Jugendmedienschutz (wozu diese sich der Unterstützung anerkannter Einrichtungen der freiwilligen Selbstkontrolle bedienen können), als auch die Rolle der KJM als Organ für die Aufsicht und Klärung von Einzelfällen. Die JuSchRiL konkretisieren die im JMStV definierten Tatbestände und liefern somit Interpretationshilfen für die Beurteilung von Einzelfällen (vgl. [JuSchRiL, 2005] Abschnitt 2). Ebenfalls werden die im JMStV zum Begriff *Entwicklungsbeeinträchtigung* verwendeten Formulierungen weiter konkretisiert. Hinsichtlich der altersdifferenzierten Bewertung und Verfügbarmachung von Angeboten liefern die JuSchRiL jedoch keine weitergehenden Erläuterungen oder Festlegungen. Es wird aber festgestellt: »Die Beurteilung der Beeinträchtigung hat an den schwächeren und noch nicht so entwickelten Mitgliedern der Altersgruppe zu erfolgen. Die mögliche Wirkung auf bereits gefährdungsgeneigte Kinder und Jugendliche

[9] Dies ist ebenfalls zutreffend für Angebote, die für Kinder und Jugendliche unter 18 Jahren entwicklungsbeeinträchtigend sein können.

[10] Unter dem Begriff *labeln* wird hier die technische Auszeichnung eines Angebots mit einer Altersfreigabe verstanden.

ist angemessen zu berücksichtigen.« [JuSchRiL, 2005] Daraus lässt sich ableiten, dass die Zuordnung einer größeren Altersspanne zu einer Gruppe zu Lasten der Bedürfnisse der älteren Mitglieder einer Altersgruppe geht. Abschnitt 2.1.2 der vorliegenden Studie betrachtet die in Deutschland Anwendung findenden Altersgruppen näher und vergleicht diese mit dem internationalen Umfeld.

1.1.6 Netzregeln der KJM

Die KJM hat 2004 zehn Netzregeln formuliert, die aus Sicht der KJM wünschenswerte Regelungen an die Eigenverantwortung und Eigeninitiative der Anbieter im Modell der regulierten Selbstregulierung skizzieren und damit geltende gesetzliche Regelungen und Richtlinien ergänzen. Sie gehen dabei über die gesetzlichen Vorschriften hinaus und sollen den Unternehmen als Orientierung zur Beachtung des Jugendschutzes im Internet dienen. Sie lauten [KJM, 2004]:

1. »Der Anbieter ist für alle eigenen Inhalte selbst verantwortlich. Darüber hinaus wählt er verlinkte Angebote sorgfältig aus und überprüft regelmäßig ihre Jugendschutz-Konformität.«
2. »Der Anbieter macht pornographische und sonstige jugendgefährdende Angebote und Werbung für solche Angebote nur Erwachsenen in geschlossenen Benutzergruppen zugänglich. Darüber hinaus verfährt er mit grenzwertigen Angeboten und grenzwertiger Werbung in vergleichbarer Weise.«
3. »Der Anbieter trägt dafür Sorge, dass entwicklungsbeeinträchtigende Inhalte durch Minderjährige üblicherweise nicht wahrgenommen werden. Darüber hinaus nimmt er bei der Gestaltung seines Angebots auf minderjährige Nutzer besondere Rücksicht.«
4. »Der Anbieter macht Angebote, die nur für Kinder beeinträchtigend sind, getrennt von Kinderangeboten zugänglich. Darüber hinaus unternimmt er Anstrengungen, sein eigenes Kinderangebot – soweit vorhanden – bekannt zu machen.«
5. »Der Anbieter trifft Vorkehrungen, um Gefährdungen und Beeinträchtigungen für minderjährige Nutzer bei seinen allgemein zugänglichen Angeboten mit Kontakt- und Austauschmöglichkeiten zu minimieren.«
6. »Der Anbieter trifft bei Angeboten mit Kontakt- und Austauschmöglichkeiten, die sich gezielt an Kinder richten, besonders hohe Sicherheitsvorkehrungen.«
7. »Der Anbieter trennt Inhalt und Werbung/Sponsoring. Darüber hinaus gestaltet er Werbebotschaften so, dass sie auch als solche für Kinder und Jugendliche erkennbar sind.«
8. »Der Anbieter, dessen Angebot sich überwiegend an Kinder und Jugendliche richtet, wirbt dort nicht für Alkoholika und Tabak. Darüber hinaus zeigt er im Rahmen dieses Angebotes eine besondere Sorgfalt und Sensibilität für die Inhalte und die Gestaltung von Werbung und verzichtet auf die Bewerbung von für Jugendliche entwicklungsbeeinträchtigenden Angeboten.«
9. »Der Anbieter, der personenbezogene Daten erhebt, trägt dafür Sorge, dass den Interessen von Kindern und Jugendlichen nicht geschadet oder deren Unerfahrenheit ausgenutzt wird.«
10. »Der Anbieter unterstützt Kinder, Jugendliche und Eltern in ihrem Verständnis und in der Nutzung des Internets. Er integriert nach Möglichkeit internetadäquate Maßnahmen zur Förderung von Medienkompetenz in seinen Webauftritt oder verlinkt entsprechende Angebote.«

1.1.7 UN-Kinderrechtskonvention

Das *Übereinkommen über die Rechte des Kindes*, wie die UN-Kinderrechtskonvention im vollen Wortlaut heißt, enthält zahlreiche Aspekte hinsichtlich der Rechte von Heranwachsenden, die auch auf die Nutzung des Internets anzuwenden sind. Die UN-Kinderrechtskonvention wurde 1990 von der Bundesrepublik Deutschland unterzeichnet und ist 1992 für Deutschland in Kraft getreten.

»Die in dem Dokument niedergelegten Grundsätze machen über die vorrangige Elternverantwortung hinaus die Verpflichtung der Vertragsstaaten deutlich, positive Rahmenbedingungen für die Entwicklung von Kindern und Jugendlichen zu schaffen. Die Kinderrechtskonvention ist somit ein Zeichen von Achtung und Verantwortlichkeit der internationalen Staatengemeinschaft gegenüber Kindern in aller Welt. Die Bundesrepublik Deutschland hat die Forderungen der Konvention nahezu lückenlos im nationalen Recht festgeschrieben«, schrieb die ehemalige Bundesministerin für Familie, Senioren, Frauen und Jugend, Dr. Christine Bergmann, im Vorwort zur deutschen Übersetzung der UN-Kinderrechtskonvention [Vereinte Nationen, 1989]. Zwar steckte die Entwicklung des Internets zum Zeitpunkt der Verabschiedung der Kinderrechtskonvention noch in den Kinderschuhen, die in dem Dokument niedergelegten Grundsätze sind aber durchaus auf den Zugang zum und die Nutzung des Internets übertragbar.

Basierend auf dem Recht des Kindes, sich eine eigene Meinung zu bilden und äußern zu dürfen (Artikel 12), enthalten weitere Artikel Konkretisierungen der Rechte – und anzuwendende Grenzen, die für den Zugang und die Nutzung des Internets von Bedeutung sind. Dies sind zum Beispiel

- das Recht auf Informationsfreiheit (Artikel 13),
- das Recht auf Gedanken-, Gewissens- und Religionsfreiheit (Artikel 13),
- das Recht auf den Schutz der Privatsphäre (Artikel 16),
- das Recht auf Zugang zu den Medien und Kinder- und Jugendschutz (Artikel 17),
- das Recht auf den Schutz vor geistiger Gewaltanwendung (Artikel 19),
- das Recht auf Bildung (Artikel 28) und Bildungsziele (Artikel 29) sowie
- das Recht auf freie Teilnahme am kulturellen und künstlerischen Leben (Artikel 31).

1.2 Organisationen und Beteiligte

Wie in Abschnitt 1.1.1 dargelegt wurde, ist der Schutz von Kindern und Jugendlichen eine verfassungsrechtlich verankerte Aufgabe des Staates. Der deutsche Staat mit seinem föderalen System hat in Erfüllung dieser Aufgabe die Verantwortung an verschiedene Stellen übertragen, die in diesem Abschnitt näher betrachtet werden.

1.2.1 Behörden

Nachfolgend werden Behörden des Bundes und der Länder vorgestellt, die mit dem Thema Jugendmedienschutz befasst sind.

1.2.1.1 Der Beauftragte der Bundesregierung für Kultur und Medien (BKM)

Der Beauftragte der Bundesregierung für Kultur und Medien (BKM) unterstützt Initiativen, die die Entwicklung der Medienkompetenz von Kindern und Jugendlichen fördern. »Kinder und Jugendliche sind heute unzähligen Einflüssen der Medien ausgesetzt. Es ist wichtig, sie für das Leben mit Medien stark zu machen. Zu lernen, mit allen Medienformen selbstbewusst und reflektiert umzugehen, schützt Kinder vor den negativen Aspekten der Medienkonsums. Und: Medienkompetenz ist notwendig, um verantwortlich an gesellschaftlichen Prozessen in unserer Demokratie mitwirken zu können.« [BKM, 2012c]

Für Staatsminister Bernd Neumann ist die Vermittlung von Medienkompetenz in der Informationsgesellschaft »von herausragender Bedeutung, um Kindern und Jugendlichen den Zugang zu Bildung, Kultur und Information zu ermöglichen« [BKM, 2012a].

Mit der 2007 von Bundeskanzlerin Dr. Angela Merkel gestarteten Initiative »Ein Netz für Kinder«, die vom BKM gemeinsam mit dem Bundesministerium für Familie, Senioren, Frauen und Jugend (BMFSFJ) umgesetzt wird, soll ein attraktiver und sicherer Surfraum für Kinder von acht bis zwölf Jahren geschaffen werden, in dem die Kinder nicht nur einzelne gute Angebote vorfinden, sondern in einem großen Surfraum durch selbstständiges Navigieren den Umgang mit dem Medium Internet erlernen können, ohne der Gefahr ausgesetzt zu sein, mit für sie ungeeigneten Inhalten konfrontiert zu werden.

Die Initiative besteht aus der medienpädagogisch geprüften Positivliste von für Kinder unbedenklichen Angeboten, die von den Wirtschaftsbeteiligten der Initiative im Verein fragFINN e. V. gepflegt wird und über die Kindersuchmaschine fragFINN[11] zugänglich ist, sowie aus einem Förderprogramm, für das der BKM jährlich eine Million und BMFSFJ eine halbe Millionen Euro zur Verfügung stellen, um innovative und qualitativ hochwertige Kinderangebote zu fördern.[12]

Zu den weiteren vom BKM geförderten Initiativen gehören hinsichtlich der Internetnutzung das in Abschnitt 1.2.1.2 beschriebene Projekt *sicher online gehen - Kinderschutz im Internet* sowie der *Deutsche Computerspielpreis*[13].

Daneben unterstützt der BKM weitere Initiativen zur Förderung der Medienkompetenz, zum Beispiel im Bereich Film oder Print. Mit der Förderung der Medienkompetenz Heranwachsender und der technischen Begleitforschung des Jugendmedienschutzes im Internet durch die vorliegende Studie leistet der BKM – häufig in Partnerschaft mit dem BMFSFJ – einen Beitrag zum Jugendmedienschutz in Deutschland.

1.2.1.2 Bundesministerium für Familie, Senioren, Frauen und Jugend (BMFSFJ)

Das BMFSFJ setzt sich naturgemäß für den Schutz von Kindern und Jugendlichen ein. Im Kontext des Jugendschutzes wird gerade der Bereich *Online* von der Ministerin Dr. Kristina Schröder als verbesserungswürdig betrachtet: »Eltern brauchen bessere Rahmenbedingungen zum Schutz ihrer Kinder bei Online-Filmen und Online-Spielen […] Offline und Online brauchen wir ein soweit wie möglich vergleichbares Niveau des Jugendschutzes. Mir ist wichtig, dass Eltern, die im Internet auf Jugendschutzprogramme setzen wollen, optimale Rahmenbedingungen zum Schutz ihrer Kinder vorfinden. Das betrifft in besonderem Maße Filme und Spiele im Internet.« [BMFSFJ, 2012b] Das BMFSFJ fördert zahlreiche Projekte und Initiativen zur Förderung der Medienkompetenz und des Jugendmedienschutzes. »Eine Kinder- und Jugendnetzpolitik heißt konkret mit dem Netz zu arbeiten statt abstrakt dagegen anzukämpfen. Drei Dinge gehören dabei untrennbar zusammen: Medienkompetenz, Elternverantwortung und Schutz vor Gefahren im Netz. Das Ziel ist, Kinder und Jugendliche für den Umgang mit dem Netz stark zu machen.« [BMFSFJ, 2011] Exemplarisch für das Engagement des BMFSFJ werden vier Initiativen kurz vorgestellt:

[11] siehe http://www.fragFINN.de/, Stand 26.11.2012

[12] siehe http://www.ein-Netz-fuer-Kinder.de, Stand 26.11.2012

[13] »Mit dem Deutschen Computerspielpreis werden kulturell und pädagogisch wertvolle Computerspiele ausgezeichnet. 2009 haben der BKM und Branchenverbände den Preis ins Leben gerufen. In sieben Kategorien werden herausragende deutsche Spieleproduktionen prämiert« [BKM, 2012a]. Zu den prämierten Spielen gehören auch Browser-/Onlinespiele [BKM, 2012b].

Dialog Internet Bereits im Jahr 2010 hat das BMFSFJ den *Dialog Internet*[14] gestartet, der im Dialog mit Vertretern der Wirtschaft, Wissenschaft, Medien, Politik, Eltern, Pädagogen, Kindern und Jugendlichen Empfehlungen für eine zeitgemäße Kinder- und Jugendnetzpolitik entwickeln soll. Die Handlungsempfehlungen des Dialog Internet wurden bei den in 2011 angekündigten Initiativen für eine Kinder- und Jugendnetzpolitik berücksichtigt [BMFSFJ, 2011], bei denen gleichrangig auf die Förderung der Medienkompetenz von Kindern und Jugendlichen, der Erziehungskompetenz der Eltern sowie des technischen und gesetzlichen Schutzes gesetzt wird [Dialog Internet, 2011].

Initiative »sicher online gehen - Kinderschutz im Internet« Mit der Initiative setzen sich Bund, Länder und Wirtschaft unter anderem für die Information der Eltern über die vorhandenen Schutzmechanismen und die Verbreitung und Weiterentwicklung von Jugendschutzprogrammen ein. »Mit dem Start der gemeinsamen Initiative folgt Bundesfamilienministerin Dr. Kristina Schröder einer Handlungsempfehlung des Dialog Internet. Einer Forsa-Umfrage zufolge halten 95 Prozent der befragten Eltern Jugendschutzprogramme für wichtig, aber nur 20 Prozent von ihnen setzen sie zu Hause ein […]. Hier setzt die Initiative ›sicher online gehen‹ an, indem sie Eltern über technische Schutzmöglichkeiten im Internet informiert und ihnen dadurch die Begleitung ihrer Kinder beim Aufwachsen mit dem Netz erleichtert.« [Dialog Internet, 2012b]

Die Förderung eines sicheren Surfraums für Kinder steht in ähnlicher Weise auch bei der im Abschnitt 1.2.1.1 dargestellten, gemeinsam mit dem Beauftragten der Bundesregierung für Kultur und Medien getragene Initiative »Ein Netz für Kinder« sowie bei der vom BMFSFJ geförderten Kindersuchmaschine *Blinde Kuh*[15] und einer kindersicheren Startseite *www.meine-startseite.de*[16] im Projekt *Surfen Ohne Risiko* im Vordergrund.

I-KiZ – Zentrum für Kinderschutz im Internet »Zentrale Erkenntnis des Dialog Internet war, dass wirksamer Jugendschutz im Internet ein intelligentes Risikomanagement und eine Gesamtstrategie erfordert, die Vorsorge von Anbietern, technischen Schutz und Kompetenzvermittlung sinnvoll verbinden. Mit dem I-KiZ – Zentrum für Kinderschutz im Internet wird ein kinder- und jugendpolitisches Forum auf Bundesebene geschaffen, das den Jugendschutz im Internet angesichts der vielfältigen, neuen Herausforderungen des Web 2.0 in den Mittelpunkt rückt und hierzu ein dauerhaftes und starkes Bündnis staatlicher Stellen mit zivilgesellschaftlichen Partnern, Unternehmen und Verbänden schafft.« [Dialog Internet, 2012a]

Das I-KiZ wurde im September 2012 von der Familienministerin Dr. Kristina Schröder eröffnet und soll in der Diskussion mit allen Beteiligten »Lösungen anstoßen, die Risiken für junge Nutzer (des Internets) vorausschauend in den Blick nehmen. Im I-KiZ sollen Wissen gebündelt, Forschung initiiert und gemeinsam zukunftstaugliche Strategien entwickelt werden.« [BMFSFJ, 2012a]

1.2.1.3 Bundesprüfstelle für jugendgefährdende Medien (BPjM)

Die BPjM ist eine dem BMFSFJ nachgeordnete Bundesoberbehörde. »Aufgaben der BPjM sind die Indizierung jugendgefährdender Medien auf Antrag oder Anregung (›gesetzlicher Jugendmedienschutz‹) sowie die Förderung wertorientierter Medienerziehung und die Sensibilisierung der Öffentlichkeit für die Belange des Jugendmedienschutzes (›Jugendmedienschutz: Medienerziehung‹).« [BPjM, 2012a]. Die BPjM entscheidet basierend auf den gesetzlichen Arbeitsgrundlagen JuSchG, Verordnung zur Durchführung des Jugendschutzgesetzes (DVO JuSchG) und JMStV über die Jugendgefährdung von Medien. Die Zuständigkeit der BPjM umfasst dabei sowohl Trägermedien als auch Telemedien. Die BPjM definiert den Tatbestand der Jugendgefährdung wie folgt: »Träger- und Tele-

[14] siehe http://dialog-internet.de/, Stand 25.9.2012
[15] siehe http://www.blinde-kuh.de, Stand 30.10.2012
[16] siehe http://www.meine-startseite.de, Stand 30.10.2012

medien, die geeignet sind, die Entwicklung von Kindern oder Jugendlichen oder ihre Erziehung zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit zu gefährden, sind als jugendgefährdend zu bewerten. Dazu zählen vor allem unsittliche, verrohend wirkende, zu Gewalttätigkeit, Verbrechen oder Rassenhass anreizende Medien sowie Medien, in denen Gewalthandlungen wie Mord- und Metzelszenen selbstzweckhaft und detailliert dargestellt werden oder Selbstjustiz als einzig bewährtes Mittel zur Durchsetzung der vermeintlichen Gerechtigkeit nahe gelegt wird.« [BPjM, 2012a]

Die als jugendgefährdend eingestuften Medien werden auf der Liste der jugendgefährdenden Medien eingetragen (Indizierung). Die Liste wird gemäß § 18 Absatz 2 JuSchG in vier Teilen geführt, die sich hinsichtlich des Verbreitungsweges (Telemedien oder Trägermedien), der strafrechtlichen Relevanz (gem. §§ 86, 130, 130a, 131, 184a, 184b und 184c StGB) und der Veröffentlichung beziehungsweise Nicht-Veröffentlichung der Liste – beziehungsweise des Listenteils – der indizierten Medien unterscheiden [JuSchG, 2008]. Die Liste umfasste im Jahr 2012 etwa 2 870 Webangebote (Uniform Resource Locators (URLs)) [Wendland, 2012]. Die BPjM überprüft im Rahmen ihrer Tätigkeit auch Angebote des World Wide Web (WWW) auf Inhalte, die nicht außerhalb einer geschlossenen Benutzergruppe verbreitet werden dürfen (z. B. einfache Pornographie) sowie Angebote, deren Verbreitung auch in geschlossenen Benutzergruppen absolut unzulässig ist [FSM, 2011a]. Entsprechende Webseiten (URLs) werden ebenfalls auf der Liste indiziert.

Die BPjM hat zusammen mit der *Selbstkontrolle Suchmaschinen in der FSM* ein technisches Verfahren entwickelt, »durch das verhindert wird, dass Internetadressen (URLs), die von der BPjM in einem gesetzlich geregelten Verfahren indiziert und auf den Index jugendgefährdender Medien gesetzt wurden, in den Ergebnislisten der Suchmaschinen angezeigt werden. Das so genannte BPjM-Modul selbst enthält von der BPjM zur automatisierten Verarbeitung aufbereitete Datensätze, die von der BPjM verschlüsselt an die Suchmaschinenanbieter in der *Selbstkontrolle Suchmaschinen* übermittelt werden, so dass die indizierten Internetangebote aus den Ergebnislisten der beteiligten Suchmaschinen ausgeblendet werden« [FSM, 2011a]. Die Verwendung des BPjM-Moduls ist auch Voraussetzung für die offizielle Anerkennung von Jugendschutzprogrammen durch die KJM (vgl. Abschnitt 2.3.1.1).

1.2.1.4 Oberste Landesjugendbehörden

Die »Oberste Landesjugendbehörde ist die nach den Organisationsgesetzen des jeweiligen Bundeslandes für die Aufgaben […] des Jugendschutzes zuständige Behörde, also das zuständige Ministerium beziehungsweise die Senatsbehörde. Im Bereich der Kontrollaufgaben des Jugendschutzes ist sie unter anderem Antragsbehörde für die Aufnahme von Medien in die Liste der jugendgefährdenden Medien und zuständig für die Ernennung eines Teils der Prüfer und Sachverständigen bei den Prüfinstitutionen (BPjM, FSK, Unterhaltungssoftware Selbstkontrolle (USK), KJM).« [Nikles, 2011] Beispielsweise definiert das Landesjugendamt des Freistaats Bayern sein Aufgabenspektrum hinsichtlich des Jugendmedienschutzes in der Sichtung und Bewertung von Medien, zum Beispiel durch die Mitarbeit in den Ausschüssen der Freiwillige Selbstkontrolle der Filmwirtschaft (FSK), USK, Freiwillige Selbstkontrolle Fernsehen (FSF) und KJM oder der Geschäftsführung des Bayerischen Mediengutachterausschusses[17], sowie durch Vernetzung mit Jugendschutzreferenten der Medienanbieter, Kontrollinstanzen und Jugendschutzinstitutionen im Medienbereich [ZBFS, 2012].

[17] »Der Bayerische Mediengutachterausschuss berät und unterstützt die oberste Landesjugendbehörde bei ihren Entscheidungen, die die inhaltliche oder grundsätzliche Bewertung von Medien betreffen, insbesondere im Hinblick auf einschlägige Paragraphen der Jugendschutzgesetze.« [ZBFS, 2012]

1.2.1.5 Landesmedienanstalten

Die Landesmedienanstalten sind in Deutschland die Aufsichtsbehörden für private Radio- und Fernsehprogramme. Ebenfalls haben die Landesmedienanstalten die Einhaltung von Anforderungen an Telemedien, etwa im Bereich des Jugendmedienschutzes, zu prüfen. Entsprechend der grundgesetzlich verankerten Rundfunkzuständigkeit der Länder ist jedem Bundesland in Deutschland eine Landesmedienanstalt zugeordnet. Die Länder Berlin und Brandenburg sowie Hamburg und Schleswig-Holstein haben per Staatsvertrag gemeinsame Landesmedienanstalten vereinbart.[18]

Die Landesmedienanstalten sind für die Einhaltung der Vorschriften des JMStV zuständig. Für diese länderübergreifende Aufgabe ist die KJM als zentrales Organ zuständig (vgl. Abschnitt 1.2.2.1). Zudem haben die Landesmedienanstalten auch eigene Abteilungen, die sich insbesondere um Angebote aus dem Zuständigkeitsbereich der entsprechenden Landesmedienanstalt kümmern.

1.2.1.6 Bundesamt für Sicherheit in der Informationstechnik (BSI)

Das Bundesamt für Sicherheit in der Informationstechnik (BSI) ist eine zivile obere Bundesbehörde im Geschäftsbereich des Bundesministeriums des Innern (BMI). Als nationale Sicherheitsbehörde ist es das Ziel des BSI, die IT-Sicherheit in Deutschland voran zu bringen. Dabei ist das BSI in erster Linie der zentrale IT-Sicherheitsdienstleister des Bundes. Mit seinem Angebot wendet es sich aber auch an die Hersteller sowie die privaten und gewerblichen Nutzer von Informationstechnik [BSI, 2012].

Das BSI informiert private Internetnutzer über Gefahren und Schutzmöglichkeiten im Internet und erläutert in einer eigenen Rubrik die Probleme und mögliche Gefahren bei der Nutzung von Online-Technologien und -Diensten durch Kinder, stellt einen Leitfaden für Gespräche mit Kindern zur Verfügung und erläutert knapp die technischen Grundlagen für Kinderschutz-Systeme im Netz [DJI, 2011].

1.2.2 Institutionen des Jugendmedienschutzes

Neben den staatlichen Behörden, zu deren Aufgabenspektrum auch der Jugendmedienschutz zu zählen ist, existieren in Deutschland einige Institutionen, deren hauptamtliche Aufgabe die Überwachung und/oder Umsetzung des Jugendmedienschutzes und der Medienkompetenzförderung ist.

1.2.2.1 Kommission für Jugendmedienschutz (KJM)

Die Kommission für Jugendmedienschutz (KJM) ist nach dem Jugendmedienschutz-Staatsvertrag (JMStV) für die Aufsicht der Bereiche privater Rundfunk und Telemedien zuständig. Sie ist ein Organ der Landesmedienanstalten, durch das diese zentral die Einhaltung der Vorschriften des JMStV kontrollieren. »Die KJM prüft, ob Verstöße gegen den JMStV vorliegen und entscheidet über die Maßnahmen gegen den Medienanbieter. Die KJM versteht sich aber nicht nur als Aufsichtsinstanz, sondern will auch gesellschaftspolitische Prozesse anstoßen.« [KJM, 2012c] Die von der KJM gegenüber den Anbietern im Falle eines Verstoßes gegen den JMStV beschlossenen Maßnahmen werden von den Landesmedienanstalten vollzogen. (vgl. §§ 14 – 17 JMStV)

[18] siehe http://www.die-medienanstalten.de/ueber-uns.html, Stand 16.11.2012

Nach dem Leitprinzip der Eigenverantwortung des Anbieters stärkt der Staatsvertrag die Einrichtungen der Freiwilligen Selbstkontrolle der Anbieter durch das Aufsichtsmodell der *regulierten Selbstregulierung*. Dies bedeutet, dass die Anbieter für die Gewährleistung des Schutzes von Kindern und Jugendlichen bei der Gestaltung ihres Angebotes verantwortlich sind. Sie haben daher vor der Verbreitung beziehungsweise dem Zugänglichmachen die mögliche entwicklungsbeeinträchtigende oder jugendgefährdende Wirkung ihres Angebotes auf Kinder und Jugendliche in eigener Verantwortung zu prüfen und entsprechende Schutzmaßnahmen zu ergreifen. Der Anbieter kann sich für einen effektiven Jugendschutz im Sinne einer *regulierten Selbstregulierung* unter Beibehaltung der hoheitlichen Regulierungskompetenz zur Erfüllung seiner Verantwortung der Einrichtungen der Freiwilligen Selbstkontrolle bedienen. Halten sich die Anbieter an die Vorgaben der anerkannten Selbstkontrolleinrichtungen und bewegen sich die Entscheidungen der Selbstkontrolleinrichtungen im Rahmen des ihnen übertragenen Beurteilungsspielraums, ist ein aufsichtsrechtliches Einschreiten gegenüber dem Anbieter durch die KJM beziehungsweise die zuständige Landesmedienanstalt gemäß § 20 Absatz 3 bzw. 5 JMStV ausgeschlossen. Aufgabe der KJM ist es daher auch, Einrichtungen der Freiwilligen Selbstkontrolle anzuerkennen [KJM, 2012c].

Zu den weiteren Tätigkeitsfeldern der KJM gehören unter anderem die Festlegung von Sendezeiten, die Prüfung und Genehmigung einer Verschlüsselungs- und Vorsperrungstechnik und die Stellungnahme zu Indizierungsanträgen für Angebote in Telemedien (Internet) (vgl. auch [KJM, 2011b]).

Auch für die Anerkennung der Eignung von Jugendschutzprogrammen ist die KJM zuständig. Gemäß § 11 Absatz 3 JMStV ist die Anerkennung – befristet auf fünf Jahre – zu erteilen, wenn die Jugendschutzprogramme einen »nach Altersstufen differenzierten Zugang ermöglichen oder vergleichbar geeignet sind«. Weitere Anforderungen an Jugendschutzprogramme sind in den normkonkretisierenden Gemeinsamen Richtlinien der Landesmedienanstalten zur Gewährleistung des Schutzes der Menschenwürde und des Jugendschutzes (Jugendschutzrichtlinien - JuSchRiL) vom 8./9. März 2005 festgelegt. Daneben hat die KJM *Informationen für Betreiber und Anbieter von Jugendschutzprogrammen vom 11.05.2011* [KJM, 2011a] zusammengestellt, in denen die Kriterien der KJM für die Anerkennung von Jugendschutzprogrammen im Bereich des World Wide Web zusammengefasst sind.

Organisation Die KJM besteht nach den Vorgaben des JMStV aus zwölf Sachverständigen:

- sechs Direktoren der Landesmedienanstalten
- vier von den für den Jugendschutz zuständigen Obersten Landesbehörden benannte Mitglieder
- zwei von der für den Jugendschutz zuständigen Obersten Bundesbehörde benannte Mitglieder

Die KJM verfügt über eine Stabsstelle und eine Geschäftsstelle. Die Stabsstelle ist für die Bearbeitung inhaltlicher Fragen, die Vorbereitung von Grundsatzangelegenheiten und die Öffentlichkeitsarbeit zuständig, die Geschäftsstelle übernimmt organisierende und koordinierende Aufgaben.

Die »durch die obersten Landesjugendbehörden eingerichtete gemeinsame Stelle Jugendschutz aller Länder (›jugendschutz.net‹) ist organisatorisch an die KJM angebunden«, so sieht es bereits der JMStV (vgl. § 18 JMStV) vor (vgl. Abschnitt 1.2.2.2). jugendschutz.net unterstützt die KJM bei der Telemedien-Aufsicht, indem sie Anbieter auf Verstöße gegen die Bestimmungen des JMStV hinweist und die KJM oder die zuständige Einrichtung der freiwilligen Selbstkontrolle informiert.

Bearbeitung der Aufgaben Die KJM entscheidet – als Organ der Landesmedienanstalten – abschließend über einen möglichen Verstoß gegen den JMStV. Dies erfolgt in der Regel durch den Prüfausschuss, der aus drei KJM-Mitgliedern besteht, im schriftlichen Verfahren. Besteht keine Einstimmigkeit der Mitglieder, so werden die zwölf Mitglieder der KJM im schriftlichen Verfahren oder

einer ihrer monatlichen Sitzungen mit dem Prüffall befasst. Die Arbeit der KJM wird durch die inhaltliche Dokumentation und eine Vorbewertung der Prüffälle durch die zuständige Landesmedienanstalt oder jugendschutz.net sowie der Prüfgruppen der KJM unterstützt.[19] Die Umsetzung und der Vollzug der Entscheidungen erfolgt im Anschluss durch die zuständige Landesmedienanstalt.

Ein Großteil der Arbeit der KJM besteht nicht in der Regulierung von Angeboten, sondern in der Beförderung des austarierten Systems des Jugendmedienschutzes. Für solche Themenkomplexe hat die KJM verschiedene Arbeitsgruppen etabliert, die sich aus Mitgliedern der KJM, Mitarbeitern der KJM-Stabsstelle, der KJM-Geschäftsstelle, der Landesmedienanstalten und jugendschutz.net zusammensetzen. So gibt es zwischenzeitlich neun Arbeitsgruppen, die sich mit den unterschiedlichsten Themen beschäftigen und diese für die KJM aufbereiten:

- AG Berichtswesen
- AG Geschäfts- und Verfahrensordnung
- AG Jugendschutzrichtlinien
- AG Kriterien
- AG Öffentlichkeitsarbeit
- AG Selbstkontrolleinrichtungen
- AG Spiele
- AG Telemedien
- AG Verfahren

Vernetzung Im Bereich der Telemedien steht die KJM mit der BPjM (vgl. Abschnitt 1.2.1.3) bezüglich möglicher zu indizierender Angebote in einem engen Austausch. So gibt der Vorsitzende der KJM Stellungnahmen zu Indizierungsanträgen für Telemedien ab, die von der BPjM maßgeblich bei der Entscheidung für oder gegen eine Indizierung berücksichtigt werden müssen. Zudem kann die KJM selbst Indizierungsanträge stellen. Außerdem steht die KJM zur Förderung eines besseren Jugendmedienschutzes im Austausch mit zahlreichen Einrichtungen und Vertretern von Politik, Gesellschaft, Bildung und der Medienbranche.[KJM, 2011b]

1.2.2.2 jugendschutz.net

jugendschutz.net ist die Zentralstelle der Länder, die die KJM bei der Telemedien-Aufsicht (Internet) und die obersten Landesjugendbehörden (Jugendministerien bzw. Senatsbehörden der Länder) bei der Durchsetzung des Jugendschutzes im Internet [jugendschutz.net, 2012] unterstützt. Die Aufgaben von jugendschutz.net sind im JMStV (siehe § 18 JMStV) definiert:

- Die durch die obersten Landesjugendbehörden eingerichtete gemeinsame Stelle Jugendschutz aller Länder (jugendschutz.net) ist organisatorisch an die KJM angebunden. Die Stelle jugendschutz.net wird von den Landesmedienanstalten und den Ländern gemeinsam finanziert.
- jugendschutz.net unterstützt die KJM und die obersten Landesjugendbehörden bei deren Aufgaben.
- jugendschutz.net überprüft die Angebote der Telemedien. Daneben nimmt jugendschutz.net auch Aufgaben der Beratung und Schulung bei Telemedien wahr.
- Bei Verstößen gegen Bestimmungen dieses Staatsvertrages weist jugendschutz.net den Anbieter hierauf hin und informiert die anerkannten Einrichtungen der Freiwilligen Selbstkontrolle und die KJM hierüber.

[19] Eine detaillierte Dokumentation des Ablaufs der KJM-Prüfverfahren ist zu finden in [KJM, 2011b, Anlage 14 (Seite 136)].

Der Auftrag von jugendschutz.net umfasst neben der Überwachung von öffentlich verfügbaren Internetangeboten (z. B. Webseiten)[20] auch die Überwachung interaktiver und kommunikativer Angebote (z. B. Chat, Instant Messaging, File-Sharing),[21] die eine besondere Anziehungskraft auf Jugendliche ausüben. Außerdem übernimmt jugendschutz.net Aufgaben der Beratung und Schulung bei Telemedien und informiert über jugendschutzrelevante Entwicklungen und Probleme in Internet-Diensten [jugendschutz.net, 2012].

jugendschutz.net hat eine Hotline eingerichtet, über die Beschwerden von Nutzern entgegen genommen und anschließend überprüft werden. Außerdem suchen Mitarbeiterinnen und Mitarbeiter von jugendschutz.net selbst nach jugendschutzrelevanten Inhalten im Internet. Sofern es sich bei jugendschutzrelevanten Funden nicht um absolut unzulässige Inhalte im Sinne des JMStV (vgl. Abschnitt 1.1.4.1) handelt – diese werden den zuständigen Strafverfolgungsbehörden übergeben – weist jugendschutz.net den Anbieter auf den Verstoß gegen geltendes Jugendmedienschutzrecht hin und sucht den Dialog, um Anbieter für die Fragen des Jugendmedienschutzes zu sensibilisieren.

jugendschutz.net führt auch umfassende, kontinuierliche und standardisierte Tests von Jugendschutzprogrammen durch. »Das Testszenario ist von der KJM beschlossen und wird mit der FSM abgestimmt, um eine möglichst hohe Akzeptanz zu erreichen.« [jugendschutz.net, 2007] Auf Grundlage der Prüfergebnisse erteilt die KJM die offizielle Anerkennung von Jugendschutzprogrammen durch die Landesmedienanstalten (vgl. hierzu auch Abschnitt 1.4.1).

1.2.2.3 Einrichtungen der freiwilligen Selbstkontrolle

»Der Jugendmedienschutz-Staatsvertrag sieht laut § 19 Absatz 3 die Anerkennung von Einrichtungen der Freiwilligen Selbstkontrolle vor. Einrichtungen der Freiwilligen Selbstkontrolle können nach § 19 Absatz 1 JMStV für Rundfunk und Telemedien gebildet werden. Anerkannte Einrichtungen der Freiwilligen Selbstkontrolle überprüfen – im Rahmen ihres satzungsgemäßen Aufgabenbereichs nach § 19 Absatz 2 JMStV – die Einhaltung der Bestimmungen des JMStV sowie der zu diesem Zweck erlassenen Satzungen und Richtlinien bei ihnen angeschlossenen Anbietern.« [KJM, 2011b]

§ 19 des JMStV lautet [JMStV, 2003]:

1. Einrichtungen Freiwilliger Selbstkontrolle können für Rundfunk und Telemedien gebildet werden.
2. Anerkannte Einrichtungen der Freiwilligen Selbstkontrolle überprüfen im Rahmen ihres satzungsgemäßen Aufgabenbereichs die Einhaltung der Bestimmungen dieses Staatsvertrages sowie der hierzu erlassenen Satzungen und Richtlinien bei ihnen angeschlossenen Anbietern.
3. Eine Einrichtung ist als Einrichtung der Freiwilligen Selbstkontrolle im Sinne dieses Staatsvertrages anzuerkennen, wenn
   (a) die Unabhängigkeit und Sachkunde ihrer benannten Prüfer gewährleistet ist und dabei auch Vertreter aus gesellschaftlichen Gruppen berücksichtigt sind, die sich in besonderer Weise mit Fragen des Jugendschutzes befassen,
   (b) eine sachgerechte Ausstattung durch eine Vielzahl von Anbietern sichergestellt ist,
   (c) Vorgaben für die Entscheidungen der Prüfer bestehen, die in der Spruchpraxis einen wirksamen Kinder- und Jugendschutz zu gewährleisten geeignet sind,
   (d) eine Verfahrensordnung besteht, die den Umfang der Überprüfung, bei Veranstaltern auch die Vorlagepflicht, sowie mögliche Sanktionen regelt und eine Möglichkeit der Über-

[20] Diese Angebote wurden bis zum Neunten Rundfunkänderungsstaatsvertrag als Mediendienste i. S. d. 2007 abgelösten Mediendienste-Staatsvertrags (MDStV) bezeichnet.

[21] Seit 2007 sind diese den Telemedien i. S. d. Telemediengesetzes (TMG) zuzuordnen, bis dato wurden sie den Telediensten i. S. d. 2007 abgelösten Teledienstegesetzes (TDG) zugerechnet.

prüfung der Entscheidungen auch auf Antrag von landesrechtlich bestimmten Trägern der Jugendhilfe vorsieht,

(e) gewährleistet ist, dass die betroffenen Anbieter vor einer Entscheidung gehört werden, die Entscheidung schriftlich begründet und den Beteiligten mitgeteilt wird und

(f) eine Beschwerdestelle eingerichtet ist.

4. Die zuständige Landesmedienanstalt trifft die Entscheidung durch die KJM. Zuständig ist die Landesmedienanstalt des Landes, in dem die Einrichtung der Freiwilligen Selbstkontrolle ihren Sitz hat. Ergibt sich danach keine Zuständigkeit, so ist diejenige Landesmedienanstalt zuständig, bei der der Antrag auf Anerkennung gestellt wurde. Die Einrichtung legt der KJM die für die Prüfung der Anerkennungsvoraussetzungen erforderlichen Unterlagen vor. Die Anerkennung ist auf vier Jahre befristet. Verlängerung ist möglich.
5. Die Anerkennung kann widerrufen werden, wenn Voraussetzungen für die Anerkennung nachträglich entfallen sind oder sich die Spruchpraxis der Einrichtung nicht im Einklang mit dem geltenden Jugendschutzrecht befindet. Eine Entschädigung für Vermögensnachteile durch den Widerruf der Anerkennung wird nicht gewährt.
6. Die anerkannten Einrichtungen der Freiwilligen Selbstkontrolle sollen sich über die Anwendung dieses Staatsvertrages abstimmen.

Mit den Einrichtungen der Freiwilligen Selbstkontrolle führt der JMStV das Prinzip der *regulierten Selbstregulierung* ein. Dieses hat zum Ziel, die Eigenverantwortung der Rundfunk- und Internetanbieter zu stärken. Die Einrichtungen der Freiwilligen Selbstkontrolle, die zum Beispiel von Verbänden der Wirtschaft getragen werden, übernehmen die Überprüfung der Einhaltung des Jugendmedienschutzes. Den Einrichtungen der freiwilligen Selbstkontrolle wird ein gesetzlich festgeschriebener Entscheidungsrahmen zugebilligt, den die Medienaufsicht nur begrenzt überprüfen darf. Im Bereich Rundfunk und Telemedien obliegt der KJM die Anerkennung einer entsprechenden Einrichtung [KJM, 2012e]. Bislang hat die KJM vier Einrichtungen der Freiwilligen Selbstkontrolle für den Bereich der Telemedien anerkannt: die FSM, die Freiwillige Selbstkontrolle der Filmwirtschaft Abteilung FSK.online (FSK.online), die Unterhaltungssoftware Selbstkontrolle für den Bereich der Telemedien (USK.online) sowie die FSF.

Freiwillige Selbstkontrolle Multimedia-Diensteanbieter (FSM) »Die FSM ist ein eingetragener Verein, der 1997 von Medienverbänden und Unternehmen der Online-Wirtschaft gegründet wurde. Die Selbstkontrollorganisation bietet jedermann die Möglichkeit, sich im Bereich des Jugendmedienschutzes über strafbare oder jugendgefährdende Inhalte im Netz zu beschweren oder Fragen zum Thema Jugendschutz im Internet zu stellen. […] Der Verein bietet ordentlichen Mitgliedern die Möglichkeit, sich dem im JMStV vorgesehenen Modell der regulierten Selbstregulierung anzuschließen und die FSM bei Streitigkeiten mit der KJM einzuschalten. Die entsprechenden Unternehmen genießen somit die im JMStV vorgesehene Privilegierung für Mitglieder einer anerkannten Selbstkontrolle.« [FSM, 2011c]

Die FSM bietet ein System zur Unterstützung der Altersklassifikation von Webseiten. Mittels eines Fragebogens[22] kann jeder Anbieter die geeignete Altersklassifikation für seine Webseiten bestimmen und sein Webangebot mit der entsprechenden Alterskennzeichnung (age-de.xml[23]) auszeichnen.

Als Grundlage für die Altersklassifizierung dient der Verhaltenskodex der FSM [FSM, 2010], der sich hinsichtlich der Beurteilung von Angeboten eng am JMStV orientiert. Verstöße gegen den Verhaltenskodex kann die FSM gegenüber ihren Mitgliedern gemäß der FSM Beschwerdeordnung mit ei-

[22] Der Fragebogen war zum Zeitpunkt der Erstellung der vorliegenden Studie in einer kostenlosen Beta-Phase verfügbar. Die FSM kündigte an: »Nach Ende dieser Phase wird die Verwendung des Bewertungssystems nur für Privatpersonen oder für Testzwecke kostenfrei möglich sein« (siehe https://www.altersklassifizierung.de/, Stand 28.11.2012).

[23] Weitere Informationen zum technischen Auszeichnungsstandard siehe [OMK, 2012] und [Schellenberg, 2011].

nem Hinweis mit Abhilfeaufforderung, einer Rüge, einer Vereinsstrafe oder Vereinsausschluss sanktionieren (vgl. § 11 Absatz 4 [FSM, 2009a]).

FSK.online Aufgabe der FSK ist die »freiwillige Altersfreigabeprüfung von Filmen und anderen Trägermedien, die in Deutschland für die öffentliche Vorführung und Verbreitung vorgesehen sind. […] Die FSK engagiert sich auch im Online-Bereich und bietet für den Jugendschutz im Internet ein umfangreiches Angebot von Informations- und Serviceleistungen. Seit September 2011 ist sie als FSK.online eine anerkannte Einrichtung der Freiwilligen Selbstkontrolle für Webangebote.« [FSK, 2012b]

Die Grundlage der Arbeit der FSK und das Prüfverfahren ist in den *Grundsätzen der freiwilligen Selbstkontrolle der Filmwirtschaft GmbH* [FSK, 2010] definiert. Das Prüfverfahren [FSK, 2011b] sieht eine enge Zusammenarbeit der Filmwirtschaft mit der öffentlichen Hand vor, indem ständige Vertreter der obersten Landesjugendbehörden Mitglieder der Bewertungsausschüsse sind [FSK, 2012e].

Die Abteilung FSK.online bietet ihren Mitgliedern die Überprüfung von Webangeboten hinsichtlich Jugendschutzkonformität, die Altersbewertung von Webangeboten sowie Beratung und die Übernahme der Funktion des Jugendschutzbeauftragten an [FSK, 2012f]. Außerdem bietet FSK.online einen frei zugänglichen Label-Generator an, mit dem jeder Anbieter in wenigen Schritten ein age-de.xml Label (siehe [Schellenberg, 2011]) für die Altersfreigabeauszeichnung seines Webangebots generieren lassen kann [FSK, 2012f]. Die Beschwerdestelle von FSK.online »bietet allen Internetnutzern die Möglichkeit sich über Angebote im Internet zu beschweren.« [FSK, 2012c] Dabei ist FSK.online aber nur für Angebote von Mitgliedern zuständig.

Ergänzend zu den Altersfreigaben für Filme durch die FSK gibt es auch noch juristische Gutachten zu Filmen, die nicht von der FSK geprüft wurden oder deren Freigabe von der FSK – abhängig vom Trägermedium – wegen einfacher beziehungsweise schwerer Jugendgefährdung oder strafrechtlicher Bedenken abgelehnt wurde. Dabei prüft eine Juristen-Kommission (JK) der Spitzenorganisation der Filmwirtschaft (SPIO) ein Medium darauf, ob ein Verstoß gegen das Jugendschutzgesetz oder strafrechtliche Bedenken bestehen. Es existieren zwei SPIO/JK-Freigaben [SPIO, 2007]:

- »SPIO/JK geprüft: keine schwere Jugendgefährdung«: Der begutachtete Film oder Bildträger verstößt nach Auffassung der Juristenkommission weder gegen Vorschriften des Strafgesetzbuchs noch wird der Film oder Bildträger als schwer jugendgefährdend im Sinne des § 15 Absatz 2 JuSchG beurteilt. Ein so gekennzeichnetes Medium darf analog zu Titeln mit der Freigabe FSK ab 18 bis zu einer möglichen Indizierung durch die BPjM im Handel offen ausgestellt werden; der Verkauf ist nur an Personen ab 18 Jahren erlaubt.
- »SPIO/JK geprüft: strafrechtlich unbedenklich«: Die Juristenkommission kommt zu der Auffassung, dass der Film in der vorgelegten Fassung nicht gegen Vorschriften des Strafgesetzbuchs verstößt, aber als schwer jugendgefährdend im Sinne des § 15 Absatz 2 JuSchG zu beurteilen ist. Ein derartig gekennzeichnetes Medium unterliegt automatisch den Vertriebs- und Werbebeschränkungen nach § 15 Absatz 1 Nummer 1-7 JuSchG. Daher darf ein solches Medium nicht offen im Handel ausgestellt werden. Nur auf gezielte Nachfrage volljähriger Personen darf das Medium *unter der Ladentheke* verkauft werden. Es ist damit indizierten Titeln gleichgestellt.

USK.online »Die Unterhaltungssoftware Selbstkontrolle (USK) ist eine freiwillige Einrichtung der Computerspielewirtschaft. Sie ist zuständig für die Prüfung von Computerspielen in Deutschland. Am Ende eines gemeinsamen Verfahrens vergeben staatliche Vertreter die Alterskennzeichen. Das System stellt sicher, dass Computerspiele nur an Kinder und Jugendliche abgegeben werden, wenn die Inhalte für ihre Altersgruppe freigegeben sind. Diese Einstufungen richten sich nach dem Jugendschutzgesetz.« [USK, 2012a]

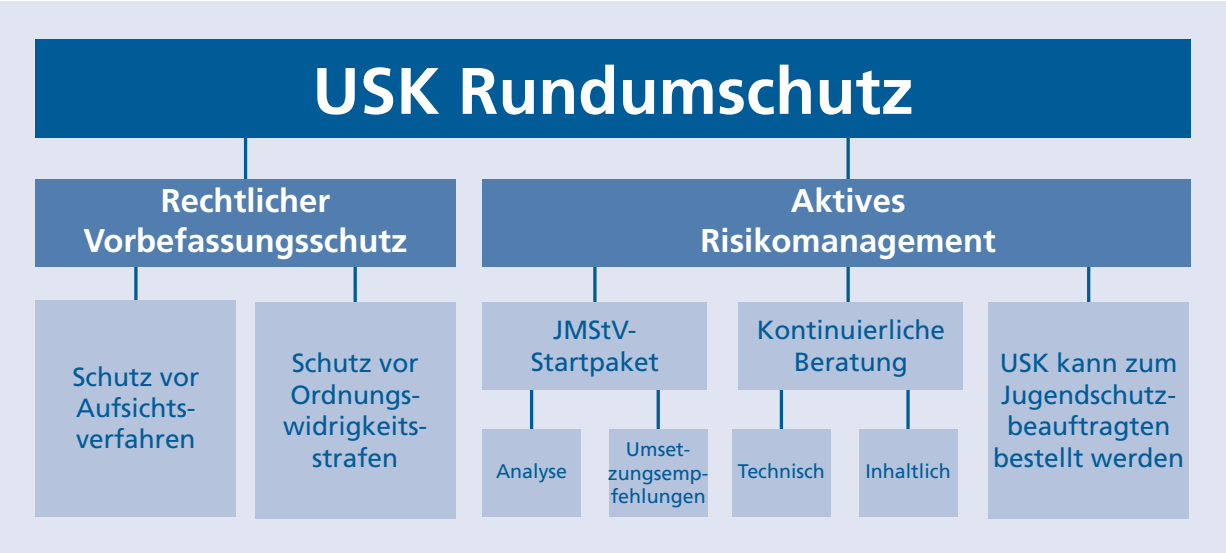


Abbildung 1.1
Services der USK.online [USK, 2012b]

Die Grundlagen der Arbeit der USK ist in den *Grundsätzen der Unterhaltungssoftware Selbstkontrolle (USK)* [USK, 2011a] definiert. Beraten wird die USK dabei von einem Beirat, der die Arbeit der Sichter und Jugendschutzsachverständigen bestätigt. »Die Beiratsmitglieder sind u. a. Vertreterinnen und Vertreter der Kirchen, der Medienpädagogik, des BMFSFJ, der Landesjugendbehörden, der BPjM, der Jugendorganisationen, der KJM sowie der Computerspielewirtschaft.« [USK, 2011b] Ähnlich der durch die FSK vorgenommenen Altersfreigabe stellen die vergebenen Alterskennzeichen der USK eine rechtliche Verbindlichkeit für den Vertrieb dar und sind im juristischen Sinne als *Verwaltungsakt* anzusehen.

»Die USK ist […] seit September 2011 auch für den Bereich der Online-Medien nach dem JMStV anerkannt. Durch diese umfassende staatlich übertragene Kompetenz kann die USK als einheitlicher Ansprechpartner bei allen Fragen des Jugendschutzes wirken.« [USK, 2012b] Für den Bereich der Telemedien hat die USK ein eigenes Grundsatzpapier *Statut USK.online* [USK, 2011c] verabschiedet, das die Arbeit und Strukturen der Abteilung USK.online definiert.

Die Serviceleistungen der USK.online (vgl. Abbildung 1.1) umfassen die Überprüfung und Altersbewertung von Spielen, Einzelinhalten oder ganzen Webangeboten, die Übernahme der Funktion des Jugendschutzbeauftragten und Beratung hinsichtlich der Gestaltung und Jugendschutz-konformer technischer Umsetzung von Webangeboten. Auch USK.online bietet einen Label-Generator zur einfachen Erstellung einer age-de.xml Altersauszeichnung für eine Webseite [USK, 2012b]. Ebenfalls führt die USK.online eine Beschwerdestelle, die für die Bearbeitung von Beschwerden über Angebote von Mitgliedern zuständig ist.

Freiwillige Selbstkontrolle Fernsehen (FSF) Die FSF ist ein gemeinnütziger Verein privater Fernsehanbieter in Deutschland. Ziel der FSF ist es, einerseits durch eine Programmbegutachtung den Jugendschutzbelangen im Fernsehen gerecht zu werden und andererseits durch medienpädagogische Aktivitäten, Publikationen und Unterstützung von Forschungsarbeiten den bewussteren Umgang mit dem Medium Fernsehen zu fördern.

Aufgabe der Prüfungen ist es, zu bewerten, ob Fernsehsendungen Wirkungsrisiken für bestimmte Altersgruppen beinhalten. Ergebnis der Prüfungen ist die Freigabe für eine bestimmte Sendezeit, die Erteilung von Schnittauflagen oder die Ablehnung. Nach § 5 JMStV gelten die in Tabelle 1.1 dargestellten Sendezeitschienen für die verschiedenen Altersfreigaben.

Tabelle 1.1
Sendezeitschienen nach § 5 des JMStV

Sendezeitschiene:	Uhrzeit:	Altersgruppe:
Tagesprogramm	6–20 Uhr	Kinder bis 12 Jahren, auch ohne Aufsicht
Hauptabendprogramm	20–22 Uhr	ältere Kinder ab 12 Jahren, meist zusammen mit den Eltern
Spätabendprogramm	22–23 Uhr	Jugendliche ab 16 Jahren
Nachtprogramm	23–6 Uhr	Erwachsene ab 18 Jahren

Seit März 2012 ist die FSF auch als Selbstkontrolleinrichtung für Inhalte im Internet anerkannt. Im Zuge der erweiterten Anerkennung und im Sinne einer einheitlichen Information der Öffentlichkeit über Jugendschutzentscheidungen hat die FSF ihre Prüfentscheidungen erweitert: Neben den bisher üblichen Sendezeitbeschränkungen werden Sendungen nun auch mit Altersfreigaben versehen. Die Altersstufen sind: ab 0 Jahren, ab 6 Jahren, ab 12 Jahren, ab 16 Jahren und ab 18 Jahren [FSF, 2012a].

1.2.3 Initiativen für den Jugendmedienschutz

Neben den in Abschnitt 1.2.2 beschriebenen Institutionen engagieren sich auch zahlreiche Initiativen und Vereine für den Jugendmedienschutz, die sich zum Beispiel aus Vertretern der Politik, der Wirtschaft, der Jugendschutzinstitutionen, der Kirche und anderen Einrichtungen der Gesellschaft zusammensetzen. Einige der Initiativen werden im Folgenden kurz vorgestellt.

Deutschland sicher im Netz Deutschland sicher im Netz e. V. (DsiN) ist ein Verein unter der Schirmherrschaft des Bundesministeriums des Innern (BMI) mit 18 Mitgliedern[24] aus den Bereichen Unternehmen, Verbände sowie gemeinnützige Organisationen. Diese leisten mit ihren konkreten Handlungsversprechen einen praktischen Beitrag für mehr IT-Sicherheit [DsiN, 2012b].

»Zum Konzept gehört auch, speziell Eltern und Lehrer/Pädagogen dabei zu unterstützen, das ganze Gefahrenpotenzial im Web zu erkennen, ihr individuelles Verhalten darauf abzustimmen, Fehler und Nachlässigkeiten zu vermeiden sowie aktuelle Sicherheitsfunktionen und -technologien richtig nutzen zu können.« [DJI, 2011]

Initiative »SCHAU HIN! Was Deine Kinder machen« *SCHAU HIN! Was Deine Kinder machen* ist eine 2003 gegründete Initiative des BMFSFJ, des Telekommunikationsunternehmens Vodafone, der öffentlich-rechtlichen Rundfunkanstalten ARD und ZDF sowie der Fernseh-Programmzeitschrift TV Spielfilm. SCHAU HIN! hilft Eltern und Erziehenden mit zahlreichen alltagstauglichen Tipps und Empfehlungen, ihre Kinder im Umgang mit Medien zu stärken [SCHAU HIN!, 2012].

»Vermittlung von fundiertem Wissen und konkreter Rat, zum Beispiel zu Filtersoftware oder zu Onlinespielen, sollen dabei unterstützen, die (Medien-)Erziehungskompetenz zu stärken. Mit Hilfe eines interaktiven *Medienpasses* können Eltern und Pädagogen überprüfen, ob sie selbst *fit für kindgerechte Medienerziehung* sind.« [DJI, 2011]

Bundesarbeitsgemeinschaft Kinder- und Jugendschutz (BAJ) Die Bundesarbeitsgemeinschaft Kinder- und Jugendschutz e. V. (BAJ) ist ein Zusammenschluss von freien Trägern der Jugendhilfe,

[24] Mitglieder sind: BITKOM, Bundesdruckerei, DATEV, Deutsche Telekom, Deutsches Kinderhilfswerk, eBay, eco, FSM, Google, IDW, it-sa Benefiz, Microsoft, OpenLimit, PayPal, poolworks, SAP, Sophos und TeleTrusT. [DsiN, 2012a]

Fachorganisationen und Landesarbeitsgemeinschaften/Landesstellen für Kinder- und Jugendschutz sowie interessierten Einzelpersönlichkeiten, die sich für den Jugendschutz engagieren.

Das Thema Medien wird in der Aktion Jugendschutz von den Landesarbeitsstellen Baden Württemberg,[25] Bayern (Aktion Jugendschutz, Landesarbeitsstelle Bayern (aj)),[26] Brandenburg (Aktion Kinder- und Jugendschutz, Landesarbeitsstelle Brandenburg e. V. (AKJS)),[27] Niedersachsen,[28] Nordrhein-Westfalen (Arbeitsgemeinschaft Kinder- und Jugendschutz (AJS)),[29] Sachsen,[30] Sachsen-Anhalt,[31] Schleswig Holstein,[32] und Thüringen[33] [BAJ, 2012] berücksichtigt.

klicksafe.de – Die EU-Initiative für mehr Sicherheit im Netz *klicksafe* ist eine Sensibilisierungskampagne zur Förderung der Medienkompetenz im Umgang mit dem Internet und neuen Medien im Auftrag der Europäischen Kommission (*Safer Internet Programme*). Aktionsfelder des Projektes sind Inhalte/Qualifikation, Kampagne/Marketing sowie die bundesweite Vernetzung mit Partnern und Akteuren [klicksafe, 2012]. klicksafe erarbeitet vielfältige Informationsmaterialien zum Thema Jugendmedienschutz insbesondere für Eltern und Pädagogen und stellt diese online und in gedruckter Form zur Verfügung.

Europäische Kommission – Safer Internet Programme sowie Mitteilung Europäische Strategie für ein besseres Internet für Kinder Zur Förderung der sicheren Nutzung moderner Kommunikationsmittel und dem Schutz von Verbrauchern vor unerwünschten Inhalten startete die EU 1999 das *Safer Internet Programme*, einen Aktionsplan, dessen Schwerpunkte die Bekämpfung von illegalen, unerwünschten oder schädlichen Inhalten sowie die Förderung eines sicheren Umfelds darstellen. Das beinhaltet auch die Aufklärung der Bevölkerung und Pflege der Zusammenarbeit in einer breit angelegten Partnerschaft von gesellschaftlichen Gruppen und Einrichtungen. Mit der Verabschiedung des *Safer Internet Programme 2009-2013* durch das Europäische Parlament und den Rat Ende 2008 ging das Programm in die mittlerweile dritte Runde und wurde um weitere vier Jahre verlängert [klicksafe, 2012].[34]

Im Mai 2012 hat die Europäische Kommission eine neue *EU-Strategie für ein besseres Internet für Kinder*[35] verabschiedet, damit der besonderen Schutzbedürftigkeit der Kinder Rechnung getragen wird (vgl. [EC, 2012a] und [EC, 2012c]). Wesentliche Ziele sind die Förderung hochwertiger Online-Inhalte für junge Leute, eine verstärkte Sensibilisierung und Befähigung sowie die Schaffung eines sicheren Online-Umfelds für Kinder. Die Mitteilung [EC, 2012b] bezieht sich ausdrücklich auf die deutsche Initiative, hochwertige Inhalte zu fördern sowie durch Positivlisten geschaffene Surfräume wie fragFINN bereit zu stellen. Sie weist zudem auf die in Deutschland zertifizierte Jugendschutzsoftware hin, die benutzt wird, um zu verhindern, dass Kinder auf Websites mit schädlichen Inhalten gelangen und fordert die Entwicklung wirksamer Schutzinstrumente und -strategien für alle Endgeräte, um die Kinder und Jugendlichen vor schädlichen Inhalten zu schützen sowie breitere Nutzung von Alterseinstufungs- und Inhaltsklassifizierungssystemen.

fragFINN fragFINN ist ein Gemeinschaftsprojekt von Staat und Wirtschaft, durch das ein gesicherter Surfraum für Kinder im Alter von etwa acht bis zwölf Jahren geschaffen wurde, in dem Kinder durch

[25] siehe http://www.ajs-bw.de, Stand 30.10.2012
[26] siehe http://www.bayern.jugendschutz.de, Stand 30.10.2012
[27] siehe http://www.jugendschutz-brandenburg.de, Stand 30.10.2012
[28] siehe http://www.jugendschutz-niedersachsen.de, Stand 30.10.2012
[29] siehe http://www.ajs.nrw.de, Stand 30.10.2012
[30] siehe http://www.jugendschutz-sachsen.de, Stand 30.10.2012
[31] siehe http://www.jugend-lsa.de/jugendschutz, Stand 30.10.2012
[32] siehe http://www.akjs-sh.de, Stand 30.10.2012
[33] siehe http://www.jugendschutz-thueringen.de, Stand 30.10.2012
[34] Informationen des Safer Internet Programme: http://ec.europa.eu/information_society/activities/sip/index_en.htm, Stand 11.10.2012
[35] siehe http://ec.europa.eu/information_society/activities/sip/policy/index_en.htm, Stand 13.11.2012

selbstständiges Navigieren den Umgang mit dem Medium Internet erlernen können. Hinter fragFINN steht eine Datenbank (die Positivliste), die von Medienpädagogen in redaktioneller Unabhängigkeit aufgebaut, beständig kontrolliert und ergänzt wird. Maßstab für die Aufnahme in die Liste ist ganz bewusst die Unbedenklichkeit der Angebote für Kinder. Dadurch kann die Whitelist so umfangreich ausgebaut werden, dass Kinder die Grenzen des Surfraums nicht spüren; sie enthält zur Zeit über 11 000 Domains und wächst ständig. Durch verschiedene technische Schutzmechanismen kann die Suche auf die geprüften Seiten beschränkt werden. Die Positivliste von fragFINN ist integraler Bestandteil der von der KJM anerkannten Jugendschutzprogramme (vgl. Abschnitt 1.4.1). Für die mobile Nutzung auf Smartphones und Tablet PCs steht seit November 2012 die fragFINN App zur Verfügung (vgl. [fragFINN, 2012a]).

Der fragFINN e. V. wird von namhaften Unternehmen und Verbänden der Telekommunikations-, Internet- und Medienbranche finanziert. fragFINN startete Ende 2007 im Rahmen der Initiative »Ein Netz für Kinder« des BKM und des BMFSFJ ([fragFINN, 2012b]).

Blinde Kuh Die Kindersuchmaschine *Blinde Kuh*[36] enthält ein großes Angebot altersgerechter Kinderseiten, die von einer Redaktion händisch aufgenommen werden und über eine Schlagwortsuche zugänglich sind. Das Angebot für 6- bis 12-Jährige ist sicher, da ausschließlich geprüfte Seiten aufgenommen werden. Suchergebnisse können für ältere und jüngere Kinder differenziert angezeigt werden. Es gibt auf der Seite keine Werbung und keine kostenpflichtigen Angebote. Betreiber der *Blinden Kuh* ist der gemeinnützige Verein *Blinde Kuh e. V.*, der vom BMFSFJ gefördert wird.

Helles Köpfchen *Helles Köpfchen*[37] ist ein Internetportal für Kinder und Jugendliche, das neben einem redaktionellen Wissensportal für Kinder und Jugendliche zu politischen, gesellschaftlichen, kulturellen und naturwissenschaftlichen Themen einen betreuten Community-Bereich und eine kindgerechte Suchmaschine umfasst. Betreiber von *Helles Köpfchen* ist die seit 2009 von *Ein Netz für Kinder* geförderte Cosmos Media UG [Cosmos Media, 2012].

Seitenstark – Arbeitsgemeinschaft vernetzter Kinderseiten *Seitenstark*[38] ist der Zusammenschluss von 48 miteinander vernetzten Internetseiten für Kinder, die strenge Seitenstark-Qualitätskriterien erfüllen und Sicherheit vor Datenmissbrauch und Kommerz garantieren [Seitenstark, 2012]. *Seitenstark* wird vom Trägerverein Seitenstark e. V. und seit 2008 von *Ein Netz für Kinder* getragen.

Klick-Tipps.net klick-tipps.net[39] ist ein gemeinsames Projekt der Stiftung MedienKompetenz Forum Südwest und jugendschutz.net. Angeboten wird ein Service für Kinder und Jugendliche, der wöchentlich zu aktuellen Themen kindgeeignete Webangebote empfiehlt und eine Top 100 Liste bereitstellt, welche aus Surftipps besteht, die von Kindern bewertet und von Medienpädagogen überprüft wurden [klick-tipps.net, 2012].

Internet ABC Das *Internet-ABC*[40] ist eine werbefreie, spielerisch-informative Plattform für einen sicheren Interneteinstieg von Kindern. »Als Ratgeber im Netz bietet es konkrete Hilfestellung und Informationen über den verantwortungsvollen Umgang mit dem World Wide Web. Die werbefreie Plattform richtet sich mit Erklärungen, Tipps und Tricks an Kinder von fünf bis zwölf Jahren, Eltern und Pädagogen – ob Anfänger oder Fortgeschrittene. Hinter dem Projekt steht der gemeinnützige Verein Internet-ABC, dem 13 Landesmedienanstalten angehören. Zentrales Ziel der Vereinsarbeit ist

[36] siehe http://www.blinde-kuh.de/, Stand 26.11.2012
[37] siehe http://www.helles-koepfchen.de/, Stand 24.10.2012
[38] siehe http://seitenstark.de/, Stand 26.11.2012
[39] siehe http://www.klick-tipps.net, Stand 1.11.2012
[40] siehe http://www.internet-abc.de/kinder/, Stand 11.20.2012

es, Kinder und Erwachsene beim Erwerb und der Vermittlung von Internetkompetenz zu unterstützen.« [internet abc, 2012]

Erfurter Netcode – Qualitätssiegel für Kinderseiten Der *Erfurter Netcode*[41] ist eine Initiative der Landesmedienanstalten, der evangelischen und katholischen Kirche, der Karl-Kübel-Stiftung für Kind und Familie sowie weiterer Unterstützer,[42] die das anerkannte Qualitätssiegel *Erfurter Netcode* für herausragende Internetangebote für Kinder vergibt und Qualitätsstandards für Kinderseiten entwickelt.

1.2.4 Initiativen von Unternehmen der freien Wirtschaft im Bereich Selbstkontrolle

Für Unternehmen der freien Wirtschaft gewinnt der Jugendmedienschutz zunehmend an Bedeutung: Zum einen sind die Unternehmen durch gesetzliche Vorgaben angehalten, ihre Angebote Jugendschutz-konform zu gestalten, zum anderen sehen immer mehr Unternehmen die Jugendschutz-Konformität ihrer Angebote und Produkte als positives Merkmal im Wettbewerb an. Unter dem Dach der FSM haben sich Unternehmen verschiedener Sparten der Onlinewirtschaft organisiert und Verhaltenskodizes zur freiwilligen Selbstkontrolle verabschiedet.

Selbstkontrolle Mobilfunk »Im gemeinsamen ›Verhaltenskodex der Mobilfunkanbieter in Deutschland zum Jugendschutz im Mobilfunk‹ [FSM, 2005] haben sich diese nachdrücklich zum Schutz von Kindern und Jugendlichen vor entwicklungsbeeinträchtigenden und jugendgefährdenden mobilen Informations- und Kommunikationsangeboten bekannt. Der Verhaltenskodex, den die Mobilfunkanbieter im Sommer 2005 unterzeichnet haben, beschreibt gemeinsame Standards, mit denen die Mobilfunkanbieter dafür Sorge tragen, dass solche Inhalte von Kindern und Jugendlichen üblicherweise nicht wahrgenommen werden. Er regelt außerdem die mobile Nutzung von Chatrooms und den Download von Filmen und Spielen im Hinblick auf den Jugendschutz sowie die Einsetzung von Jugendschutzbeauftragten.« [FSM, 2007a]

In der Fortschreibung des Verhaltenskodex haben die Mobilfunkanbieter gemeinsam Verbesserungsmöglichkeiten für den Jugendmedienschutz erarbeitet und 2007 eine gemeinsame *Selbstverpflichtung der Mobilfunkanbieter* unterzeichnet (vgl. [FSM, 2007a]). Danach verpflichten sich die Anbieter, selbständig folgende Maßnahmen umzusetzen:

1. Einrichtung einer Jugendschutz-Hotline,
2. Online-Beratung für Eltern und Pädagogen,
3. Eltern-Informationen bei Vertragsabschluss,
4. Online-Informationen zu Jugendschutzangeboten,
5. Einrichtung einer Internet-Sperre für Endgeräte durch die Eltern

und die Umsetzung der Maßnahmen laufend zu evaluieren und kontrollieren.

Betreiber von Social Communities Die reichweitenstärksten deutschen Social Community-Betreiber haben 2009 unter dem Dach der FSM einen Verhaltenskodex [FSM, 2009b] gezeichnet, der den Jugendschutz, Datenschutz und Verbraucherschutz in den deutschen Angeboten deutlich verbessern

[41] siehe http://www.erfurter-netcode.de/, Stand 26.11.2012

[42] Unterstützende Institutionen sind Evangelische Kirche Deutschland (EKD), Stadt Erfurt, Deutsche Bischofskonferenz, Thüringer Landesmedienanstalt (TLM), Karl-Kübel-Stiftung für Kind und Familie, Thüringer Staatskanzlei, Bundeszentrale für politische Bildung (bpb), Universität Erfurt, Evangelische Kirche in Mitteldeutschland (EKM), Bistum Erfurt, Bistum-Dresden-Meissen, Hessische Landesanstalt für privaten Rundfunk und neue Medien (LPR), Landesanstalt für Medien Nordrhein-Westfalen (LfM), Landeszentrale für Medien und Kommunikation Rheinland-Pfalz (LMK) und die Bayerische Landeszentrale für neue Medien (BLM) [Erfurter Netcode, 2012].

soll. Dabei verpflichten sich die Unternehmen, vor allem junge Nutzer durch technische Maßnahmen vor Missbrauchshandlungen Dritter wie beispielsweise Cyberbullying zu schützen und durch eine verstärkte Aufklärung von Minderjährigen, Eltern und Pädagogen gezielt darauf hinzuweisen, welche Schutzmöglichkeiten bestehen. Als konkrete Schutzmechanismen sieht der Kodex vor [FSM, 2011f]:

- deutlich sichtbare Hinweise zum Schutz der Privatsphäre auf Informationsseiten direkt nach dem Registrierungsprozess,
- standardmäßig voreingestellte strengere Privatsphäreneinstellungen bei unter 14-Jährigen: Profildaten grundsätzlich nur für Freunde sichtbar,
- keine Auffindbarkeit der Profile von unter 16-Jährigen durch externe Suchmaschinen sowie keine Möglichkeit der Aufhebung dieser Einstellung,
- Bereitstellung von Beschwerdemöglichkeiten bei Verstößen gegen die Regeln der Community oder bei Verlinkungen auf Fotos durch Dritte,
- Ignorierfunktion: die Möglichkeit, andere Nutzer von der community-internen Kommunikation zu sich selbst auszuschließen sowie
- einfach zu findende Möglichkeit, das Nutzerprofil zu löschen und damit Entfernung der vom Nutzer hochgeladenen Dateien.

Selbstkontrolle Suchmaschinen Die bekanntesten deutschen Suchmaschinenanbieter haben 2005 unter dem Dach der FSM die *Selbstkontrolle Suchmaschinen* gegründet. Die Gründungsmitglieder waren:

- AOL Deutschland mit seinen Online-Diensten AOL sowie der Browser-Technologie Netscape
- Google
- IAC Search & Media mit ask.de
- MSN Deutschland
- Searchteq GmbH (vormals t-online GmbH) mit dem Online-Dienst Suchen.de
- T-Online
- Yahoo! Deutschland mit seinem gleichnamigen Online-Dienst Yahoo.de

Ziel der Initiative ist die gemeinsame Verständigung über einheitliche Standards zur Gewährleistung von Verbraucher- und Jugendschutz bei gleichzeitiger Wahrung der Meinungsfreiheit und der Vermeidung von Zensur [FSM, 2011e]. Mit dem gemeinsam verabschiedeten Verhaltenssubkodex für Suchmaschinenanbieter der FSM (VK-S) [FSM, 2004b] verpflichten sich die Anbieter zu folgenden Maßnahmen:

- Aufklärung und Information über die Funktionsweise der Suchmaschinen,
- transparente Gestaltung der Suchergebnisse (u.a. Kennzeichnung von Werbung),
- Einsatz technischer Vorrichtungen zum Schutz von Kindern und Jugendlichen vor jugendgefährdenden Inhalten,
- Grundsatz der Datensparsamkeit mit Nutzerdaten,
- Verbesserung des Jugendmedienschutzes (insbesondere Schutz vor jugendgefährdenden Inhalten),
- Nicht-Anzeige von Internetadressen, die auf dem Index jugendgefährdender Medien der BPjM stehen.

Bei Verstößen gegen den Kodex können Beschwerden gemäß der festgelegten Verfahrensordnung (Verfahrensordnung Freiwillige Selbstkontrolle Multimedia (VO-S)) bei der FSM eingereicht werden, die die Beschwerden prüft und Verstöße sanktionieren kann.

Selbstkontrolle der Chatanbieter Im Jahr 2007 haben einige Chatanbieter in Deutschland den Verhaltenssubkodex für Chatanbieter der FSM (VK-C) verabschiedet.[43] Ziel des VK-C »ist es, den Kinder- und Jugendschutz bei der Nutzung von Chats in Deutschland zu verbessern. […] Die Sicherheit in Chats soll durch eine Kombination von unterschiedlichen Maßnahmen erhöht werden. Dabei soll es den Unternehmen und der unternehmerischen Verantwortung überlassen bleiben, durch welche Maßnahmenkombination die gemeinsam definierten Sicherheitsanforderungen erreicht werden sollen.« [FSM, 2007b]

»Das Regelwerk sieht vor, dass in allen für Kinder und Jugendliche zugänglichen Chats der beteiligten Unternehmen zwischen 10 und 22 Uhr ein Moderator verfügbar sein muss. Darüber hinaus pflegen die Unternehmen eine zentrale Liste mit unerwünschten Begriffen, sogenannten ›Badwords‹, und verpflichten sich, eine Ignorierfunktion anzubieten. Dazu soll es effektive Maßnahmen zur Sperrung von Teilnehmern geben, die gegen Regeln verstoßen haben.« [Heise Online, 2007]

Kompetenzgruppe Selbstregulierung und Jugendschutz beim eco – Verband der deutschen Internetwirtschaft e. V. eco wurde 1995 als Interessenvertretung der Internetwirtschaft in Deutschland gegründet und ist mit rund 600 Mitgliedern größter Verband der Internetindustrie in Europa. Ziel von eco ist die Entwicklung von Märkten, die Förderung von Technologien und die Gestaltung von Rahmenbedingungen [eco, 2012b].

Die *Kompetenzgruppe Selbstregulierung und Jugendschutz* fungiert als Sprachrohr und Interessenvertretung der Internetwirtschaft bezüglich der Verantwortlichkeit für Inhalte.

»Vorrangiges Ziel ist die Verhinderung der Verbreitung rechtswidriger und jugendgefährdender Inhalte im Internet. Schwerpunkt der hierzu notwendigen Prüfung und Bewertung von Inhalten im Internet ist das Sichten von Usenet Newsgroups sowie die Durchführung des Beschwerdemanagements im Auftrag der teilnehmenden Internet Service Provider (ISPs). Die Kompetenzgruppe Selbstregulierung und Jugendschutz überwacht aber auch die rechtliche und politische Entwicklung im Hinblick auf die Verantwortlichkeiten und Pflichten der Provider insbesondere im Bereich des Jugendschutzes und schaltet sich im Interesse der Provider bei Forderungen staatlicher Stellen nach restriktiveren Regeln im Vorfeld durch Abgabe von Stellungnahmen, Teilnahme an Anhörungen und Erstellung juristischer Gutachten ein, um zu einer sinnvollen Berücksichtigung der Interessen der Provider und denen des Gesetzgebers zu gelangen. Die Kompetenzgruppe betreibt zudem gemeinsam mit der FSM die Internet-Beschwerdestelle Deutschland zur Entgegennahme von Beschwerden von Internetnutzern über illegale Inhalte. Solche Fremdbeschwerden werden dann nach der Beschwerdeordnung der eco-Beschwerdehotline geprüft und bearbeitet.« [eco, 2012a]

1.3 Praktische Umsetzung des Jugendmedienschutzes in Deutschland

Der Schutz von Heranwachsenden vor Einflüssen, die die Entwicklung der eigenen Persönlichkeit beeinträchtigen können, ist eine verfassungsrechtlich verankerte Aufgabe des Staates (vgl. Abschnitt 1.1.1). Dies schließt auch den Schutz vor gefährdenden oder entwicklungsbeeinträchtigenden Medien ein.

Zur Durchsetzung von gesellschaftlichen Wertvorstellungen können Verabredungen getroffen und Gesetze erlassen werden, die Verstöße gegen die Werte der Gesellschaft sanktionieren. Dies wird in Deutschland für zahlreiche Medien praktiziert. Heute besteht ein umfangreiches Netz an Maßnahmen zum Schutz der Minderjährigen, zum Beispiel für Filmvorführungen (Kino), Rundfunksen-

[43] Zu den Erstunterzeichnern gehören Lycos Europe, RTL interactive und Knuddels. Außerdem unterstützte AOL Deutschland von Beginn an den Kodex [Heise Online, 2007].

dungen, Filme und Computerspiele auf Trägermedien. Für Druckerzeugnisse und (Gesellschafts-) Spiele[44] wird auf (pädagogische) Empfehlungen der Anbieter vertraut, ohne dass ein eigenes gesetzliches Regelwerk zu Grunde liegt.

Dies funktioniert gut für abgegrenzte Werteräume von Gesellschaften. »Im Onlinebereich ist das jedoch viel schwieriger, weil man sich hier in einem Terrain internationaler Vertriebswege und Verbreitungsmöglichkeiten befindet, also nicht mehr einen begrenzten Raum hat, in dem man gesellschaftliche Verabredungen zu 100 % umsetzen und rechtlich garantieren kann. Das ist letztendlich das Problem, was wir für den Onlinebereich haben«, sagt Thomas Krüger, Präsident der bpb und stellvertretender Vorsitzender der KJM [Krüger und von Gottberg, 2012].

Zur praktischen Umsetzung des Jugendmedienschutzes wurden vielfältige Einrichtungen und Stellen geschaffen, deren hauptamtliche Aufgabe die Beurteilung von Inhalten und die Aufsicht über die Medien ist (vgl. Abschnitt 1.2). Dies sind zum Beispiel die KJM, die Landesmedienanstalten, die obersten Landesjugendbehörden, die BPjM und die Einrichtungen der Freiwilligen Selbstkontrolle (FSM, USK, FSF und FSK).

Dabei besteht zunächst eine unterschiedliche Zuständigkeit in Abhängigkeit vom Verbreitungsweg (Trägermedien, Rundfunk oder Telemedien) basierend auf den beiden Regelungskreisen des Jugendschutzes im JuSchG und im JMStV, die die getrennten Zuständigkeiten von Bund und Ländern widerspiegeln, sich in der Praxis aber teilweise überschneiden und durch eine »relativ pragmatische Zusammenarbeit« ausprägen [Krüger und von Gottberg, 2012]. Das etablierte System reagierte beispielsweise mit der Anerkennung von FSK.online, USK.online und der Erweiterung des Tätigkeitsfeldes der FSF auf die fortschreitende Konvergenz der Medien und überträgt bereits vorhandene Kompetenzen auf den Bereich des WWW.

1.3.1 Beurteilung von Angeboten

Die Beurteilung von Angeboten hinsichtlich ihrer potentiellen Eignung zur Beeinträchtigung oder Gefährdung Minderjähriger ist eine schwierige Aufgabe, da diese auf der Grundlage der Werte einer Gesellschaft stattfinden, regional durchaus voneinander abweichen können und dem Wandel der Zeit unterliegen. Insbesondere religiöse und politisch/weltanschauliche Wertvorstellungen haben einen großen Einfluss auf die Verabredungen, die Grundlage einer Gesellschaft sind.

»Gesellschaftliche Gruppen werden in die Bildung von Wertemaßstäben und entsprechenden Einschätzungen einbezogen, man verständigt sich – auch vor dem Hintergrund sich verändernder Werte – darauf, Jugendschutz als gesellschaftliche Aufgabe zu erfassen, die verfassungsrechtliche Bezüge hat.«, sagt Krüger von der KJM [Krüger und von Gottberg, 2012].

Die Arbeit einer Einrichtung bei der Beurteilung von Angeboten basiert auf Richtlinien, Handlungsanweisungen und Auslegungsgrundsätzen, die die bestehenden Gesetze konkretisieren und praktisch anwendbar machen. Eine systematische Konsolidierung der jeweiligen Spruchpraxis dokumentierter Einzelfallentscheidungen ermöglicht ein schärferes Bild der Auslegungsgrundsätze (vgl. [Liesching, 2011]). Kapitel 3 betrachtet die Kriterien für die Beurteilung von – im Sinne der vorliegenden Studie relevanten – Angeboten ausführlicher.

[44] Gemeint sind hier alle Spiele, die nicht an einem Computer oder einer Spielekonsole gespielt werden wie zum Beispiel Brett- oder Kartenspiele.

1.3.2 Technische Zugangssysteme im Sinne des Jugendmedienschutzes

Neben dem Prinzip der regulierten Selbstregulierung, das durch die freiwilligen Selbstkontrollen und die Medienaufsicht etabliert wird, hat der Gesetzgeber auch technische Zugangssysteme zur Umsetzung des Jugendmedienschutzes vorgesehen, etwa die Einrichtung geschlossener Benutzergruppen, für deren Zugang ein Altersverifikationsverfahren vorausgesetzt wird oder Schutzprogramme, die einen altersdifferenzierten Zugang zu Angeboten ermöglichen.

Nach § 4 Absatz 2 JMStV dürfen pornografische (»einfache Pornografie«[45]), bestimmte indizierte und offensichtlich schwer jugendgefährdende Inhalte im Internet verbreitet werden, wenn der Anbieter durch geschlossene Benutzergruppen sicherstellt, dass nur Erwachsene Zugriff darauf haben. Hierzu ist der Einsatz von Altersverifikations- (AV-Systeme) oder Altersprüfsystemen vorgesehen. Durch den JMStV ist aber keine gesetzliche Grundlage für die Anerkennung solcher Systeme geschaffen. Daher haben KJM und FSM jeweils ein Verfahren für die Positivbewertung von AV-Systemen entwickelt, bei dem Unternehmen ihre Konzepte für AV-Systeme von der KJM bewerten lassen können. Die Bewertung erfolgt entlang eines veröffentlichten Katalogs von Kriterien [KJM, 2012f]. Bis zum November 2012 hat die KJM 27 Module oder Gesamtkonzepte für geschlossene Benutzergruppen positiv beurteilt.

Zu beachten ist, dass die von der KJM bewerteten technischen Mittel in ihrer Strenge variieren, denn AV-Systeme müssen im Sinne des § 4 Absatz 2 Satz 2 JMStV »sicherstellen«, dass ein Angebot nur Erwachsenen zugänglich gemacht wird und technische Mittel im Sinne des § 5 JMStV müssen gewährleisten, dass Kinder oder Jugendliche Angebote »üblicherweise nicht wahrnehmen«. »[Es] sind unterschiedliche Varianten technischer Mittel möglich. [...] Technische Mittel müssen nicht das strenge Schutzniveau geschlossener Benutzergruppen erfüllen, zu denen ausschließlich Erwachsene Zugang haben dürfen. Vergleichsmaßstab sind vielmehr die aus dem traditionellen Fernsehen bekannten Sendezeitgrenzen.«[KJM, 2012g]

Die gesetzlichen Anforderungen an technische Mittel und somit an Jugendschutzprogramme, die ein mögliches technisches Mittel darstellen, stellen einen definiertes Schutzniveau dar. Die Bedürfnisse und Wünsche von Heranwachsenden und Eltern sowie die Angebote der Wirtschaft gehen über diese Anforderungen zum Teil deutlich hinaus (vgl. Abschnitt 2.3). Die vielfältigen Möglichkeiten und Bedrohungen und die rasante technologische Weiterentwicklung des Internets bieten viel Potenzial für die Weiterentwicklung bestehender und die Entwicklung neuer Schutzmechanismen. »Unabdingbar ist die kontinuierliche Verbesserung der Effizienz der Software-Lösungen, beispielsweise für Angebote im Web 2.0. Ferner sollten Jugendschutzprogramme für möglichst alle relevanten Plattformen verfügbar gemacht werden (z. B. Smartphones, Tablets, Spielkonsolen). Auch die Vereinfachung der Handhabbarkeit der bislang anerkannten Jugendschutzprogramme spielt eine zentrale Rolle. So wird zurzeit auf der Grundlage eines Praxistests[46] mit Eltern die Benutzerfreundlichkeit (Usability) der Jugendschutzprogramme überprüft.« [Weigand und Braml, 2012]

Ansatzpunkte für die weitere Entwicklung von Schutzmechanismen sind zusätzliche Funktionen, Steigerung der Effizienz und die Berücksichtigung weiterer Anwendungsszenarien – hier insbesondere der mobile Zugang zum Internet, wie in den Abschnitten 2.1 ff. aufgezeigt wird.

[45] Der JMStV unterscheidet Angebote, die »pornografisch sind und Gewalttätigkeiten, den sexuellen Missbrauch von Kindern oder Jugendlichen oder sexuelle Handlungen von Menschen mit Tieren zum Gegenstand haben; dies gilt auch bei virtuellen Darstellungen« (§ 4 Absatz 1 Nummer 10.) von Angeboten, die »in sonstiger Weise pornografisch sind« (§ 4 Absatz 2 Nummer 1.). Letztere Angebote werden als »einfache Pornografie« bezeichnet.

[46] Zum Zeitpunkt der Erstellung der vorliegenden Studie waren die in den Zulassungsbescheiden zur Anerkennung geforderten Praxistests sowohl bei der Deutschen Telekom als auch bei JusProg abgeschlossen.

1.4 Aktuelle Jugendschutzfilter

Der JMStV sieht vor, dass ein Anbieter von Telemedien, die Kinder oder Jugendliche beeinträchtigen können, seine Jugendschutzpflichten durch die Programmierung einer Altersauszeichnung für sein Angebot erfüllen kann. Dabei dient die Altersauszeichnung der altersdifferenzierten Gewährung des Zugangs zu dem Angebot durch ein von der KJM anerkanntes *Jugendschutzprogramm*. Die KJM definiert Jugendschutzprogramme wie folgt (siehe [KJM, 2011a]):

- »Jugendschutzprogramme sind technische Mittel in Form nutzerautonomer Filterprogramme für Telemedien, die von Eltern oder sonstigen Betreuungspersonen zum Schutz minderjähriger Kinder und Jugendlicher installiert und aktiviert werden können.«
- »Jugendschutzprogramme müssen Inhalte, die Kinder und Jugendliche beeinträchtigen oder gefährden können […], so zuverlässig wie möglich blockieren.«
- »Jugendschutzprogramme sind technische Mittel, mit deren Hilfe Anbieter gewährleisten sollen, dass Kinder und Jugendliche auf ihre beeinträchtigenden Inhalte üblicherweise nicht zugreifen können. Sie sollen Minderjährigen zugleich auch einen altersdifferenzierten Zugang zu Telemedien ermöglichen (oder vergleichbar geeignet sein).«

Die Realisierung eines Jugendschutzprogramms, das der vorausgehenden Definition genügt, ist auf verschiedene Weisen möglich, zum Beispiel als

- Software, die auf dem Endgerät (z. B. Personal Computer (PC)) installiert werden kann,
- App für mobile Endgeräte, Spielekonsolen oder Smart TVs,
- buchbarer Online-Dienst,
- Bestandteil des Internetzugangs durch einen Internet Service Provider (ISP),[47]
- Bestandteil eines Routers, der zum Beispiel den Übergabepunkt im Heimnetzwerk realisiert,
- Bestandteil eines Betriebssystems,
- Teil einer Internet Security Suite,[48]
- oder als Kombination der vorgenannten Möglichkeiten.

Die Nutzung von Jugendschutzsoftware ist freiwillig. Grundsätzlich besteht für den Endnutzer Wahlfreiheit hinsichtlich der Auswahl eines Programms, sofern keine Abhängigkeit vom Anbieter (z. B. ISP) oder dem verwendeten Endgerät oder Betriebssystem besteht.

Eine Unterscheidung der Begriffe Jugendschutzprogramm und Jugendschutzfilter ist aus technischer Sicht nicht eindeutig möglich und bezeichnet umgangssprachlich zumeist ähnliche Produkte. Im Rahmen der vorliegenden Studie findet der Begriff *Jugendschutzprogramm* Anwendung, wenn es sich um eine im rechtlichen Sinne gemäß § 11 JMStV definierte Software handelt. *Jugendschutzfilter* bezeichnen im Rahmen der vorliegenden Studie Softwareanwendungen, die auf einem internetfähigen End- oder Netzwerkgerät den Zugriff auf Angebote im Internet unter Berücksichtigung jugendschutzrelevanter Aspekte technisch beschränken. Sie sind damit eine spezielle Form von Filterprogrammen. In der Regel arbeiten diese auf Basis von Filterlisten, die entweder einzelne Seiten aus dem Gesamtangebot sperren (Blacklist beziehungsweise Negativliste, vgl. Abschnitt 2.4), einzelne Seiten aus dem Gesamtangebot freischalten, wobei das Gesamtangebot nicht zugänglich ist (Whitelist bzw. Positivliste, vgl. Kapitel 2.4) oder einer Kombination aus beidem. Außerdem verfügen Jugendschutzfilter zum Teil über Möglichkeiten zur »On-The-Fly« Bewertung von Web-Inhalten. Das bedeutet, dass die Inhalte nicht über ihre Webadresse mit einer Filterliste abgeglichen, sondern beim Seitenaufruf selbst untersucht und je nach Bewertung freigeschaltet oder blockiert werden. Zum

[47] z. B. durch den Einsatz von Filtern auf Proxy-Servern

[48] Als Internet Security Suite werden Software-Pakete bezeichnet, die den Anwender vor verschiedenen Gefahren bei der Nutzung des Internets schützen sollen (z. B. Virenscanner, Spamfilter, Anti-Phishing, Inhaltefilter, …).

Beispiel kann dazu der textuelle Inhalt der Webseite vor der Darstellung im Browser auf verdächtige Wortkombinationen überprüft werden. Dieses Vorgehen hat den Vorteil, dass auch nicht bewertete Seiten blockiert werden können, allerdings bedeutet die »On-The-Fly« Analyse zwangsläufig eine Verzögerung des Seitenaufrufs, die gegebenenfalls die Qualität des Surferlebnisses beeinträchtigen kann.

Im Zusammenhang mit der Bewertung der Güte von Filterprogrammen fallen üblicherweise die Begriffe Overblocking und Underblocking. Im Rahmen der vorliegenden Studie bezeichnet dabei Overblocking das unerwünschte Blockieren von nicht jugendschutzrelevanten Web-Inhalten und, im Gegensatz dazu, das Underblocking die fehlende Blockierung eines eigentlich jugendschutzrelevanten Web-Inhalts für die jeweilige Altersstufe.

Jugendschutzfilter sind idealerweise nutzerautonom, das heißt sie können durch den Anwender selbst kontrolliert, konfiguriert und bei Bedarf erweitert oder deaktiviert werden. Darüber hinaus bieten einige Jugendschutzfilter die Möglichkeit, einen altersklassifizierten Zugriff zu konfigurieren, das heißt unterschiedliche Filtergrade für unterschiedliche Altersstufen zu definieren. Somit grenzen sich diese Filterprogramme insbesondere von verwandter Software zur Überwachung und Zeitkontrolle der Internetnutzung von Minderjährigen, Antivirus-, Spamschutz- oder Sicherheits-Software und Firewall-Software ab [HBI, 2011]. Teilweise sind die eigentlichen Jugendschutzfilter in Kombination mit den genannten Typen in größere Softwarepakete (Software-Suiten) integriert.

Die KJM, die für die offizielle Anerkennung von Jugendschutzprogrammen zuständig ist, definiert den Unterschied zwischen Jugendschutzfiltern und Jugendschutzprogrammen wie folgt: »Jugendschutzprogramme zielen gemäß der Systematik des JMStV darauf ab, dass deutsche Anbieter von entwicklungsbeeinträchtigenden Inhalten in Telemedien diese verbreiten können, ohne rechtsaufsichtliche Maßnahmen durch KJM und Landesmedienanstalten befürchten zu müssen. Jugendschutzprogramme gemäß § 11 JMStV stellen damit […] ein Instrument der Privilegierung von Inhalte-Anbietern dar. Sie sind deshalb nicht mit Jugendschutzfiltern zu verwechseln, wie sie von Filterherstellern entwickelt und als Schutzmaßnahme für Familien, Schulen, Internet-Cafes oder Jugendeinrichtungen angeboten werden.« [KJM, 2012b] Voraussetzung ist insbesondere eine altersdifferenzierte Kennzeichnung der verbreiteten Inhalte durch die Anbieter, die von den anerkannten Jugendschutzprogrammen unterstützt wird. Zum Einsatz kommt hierbei in Deutschland der age-de.xml Labeling-Standard.[49] Somit sind gemäß dieser Definition nur solche Filterprogramme auch Jugendschutzprogramme, wenn sie durch die KJM anerkannt sind und damit dem § 11 JMStV genügen. Die Kriterien zur Anerkennung werden in einem späteren Abschnitt im Detail dargelegt (vgl. Abschnitt 2.3.1.1). Die zum Zeitpunkt der Anfertigung der vorliegenden Studie anerkannten Jugendschutzprogramme werden in Abschnitt 1.4.1 vorgestellt.

Begriffsabgrenzung: Jugendschutzprogramm, Jugendschutzfilter, Jugendschutzsoftware
In der Praxis werden zahlreiche Begrifflichkeiten für Instrumente zum technischen Jugendmedienschutz verwendet. Neben dem regulatorisch gewählten Begriff *Jugendschutzprogramm* finden häufig Kombinationen aus den Begriffen »Jugend-«, »Kinder-«, »Elternkontroll-« oder »Filter-« und »-programm«, »-software« oder »-system« Anwendung, sowie »Inhaltsfilter«, »technische Vorkehrung zur (Inhalts-) Filterung«, »URL-/Internet-Filter«, »Kindersicherung«, »Schutz-Software für Eltern«, »Überwachungsprogramm« und weitere [HBI, 2011]. Aus manchen der Begrifflichkeiten lassen sich dahinter stehende Funktionen ableiten, eine exakte Definition der Begriffe und dem damit verbundenen Leistungsspektrum existiert jedoch nicht. In der vorliegenden Studie wird der Begriff *Jugendschutzsoftware* oder *Jugendschutzfilter* synonym für die Gruppe der unterschiedlichen Begrifflichkeiten verwendet, ohne dabei spezielle Funktionen vorauszusetzen oder auszuschließen. Der Begriff *Jugendschutzprogramm* wird dem gegenüber verwendet, um die in Deutschland offiziell anzuerkennenden Jugendschutzprogramme auch verbal differenziert bezeichnen zu können.

[49] Details siehe [OMK, 2012]

Explizit nicht Bestandteil von Jugendschutzsoftware im Sinne der vorliegenden Studie sind Funktionen, die vor Angriffen aus dem Internet schützen wie zum Beispiel Viren, Würmer, Trojaner, Hacking, Spamming, Spoofing, Phishing etc.

Darüber hinaus existiert eine Vielzahl sowohl frei als auch kommerziell verfügbarer Jugendschutzsoftware. Im englischen Sprachgebrauch findet man häufig die Begriffe *Parental Control Software* oder eben *Filter software* oder *Internet filter*. Sie unterscheiden sich im Wesentlichen durch den Betriebskontext (Betrieb in einem Netzwerkgerät oder im Endgerät), den unterstützten Betriebssystemen und dem allgemeinen Funktionsumfang. Ein Teil dieser Programme wurde aktuell durch den vierten Testzyklus des durch die Europäische Kommission geförderten Projektes *SIP-Benchmark II*[50] evaluiert. Nähere Details dazu befinden sich in Abschnitt 1.4.2.

1.4.1 KJM-anerkannte Jugendschutzprogramme

Die KJM hat einen Katalog von *Kriterien der KJM für die Anerkennung von Jugendschutzprogrammen im Bereich des World Wide Web* [KJM, 2011a] herausgegeben, der Anbietern von Jugendschutzprogrammen offen legt, welche technischen und organisatorischen Anforderungen für die offizielle (staatliche) Anerkennung gemäß der gesetzlichen Vorgabe überprüft werden. Darüber hinaus enthält das Dokument Empfehlungen der KJM für die Hersteller von Jugendschutzprogrammen. Die in den Anerkennungskriterien und Empfehlungen enthaltenen wichtigsten technischen Anforderungen[51] sind:

- Filterung von Inhalten: Ein Jugendschutzprogramm muss Minderjährigen einen altersdifferenzierten Zugang zum WWW ermöglichen (oder vergleichbar geeignet sein).
- Effektivität: Ein Jugendschutzprogramm muss mindestens 80 % der ungeeigneten Inhalte[52] blockieren.
- Die Zuverlässigkeit eines Jugendschutzprogramms muss dem aktuellen Stand der Technik genügen.[53]
- BPjM-Listen (Empfehlung der KJM): Ein Jugendschutzprogramm soll alle jugendgefährdenden Internetangebote, die von der BPjM indiziert und in die Liste der jugendgefährdenden Medien (BPjM-Liste) eingetragen wurden, blockieren.
- Altersdifferenzierung: Ein Jugendschutzprogramm muss über altersdifferenzierte Konfigurationsmöglichkeiten mindestens für die drei Altersgruppen »bis unter 12 Jahre«, »12 bis unter 16 Jahre« und »16 Jahre bis unter 18 Jahre« verfügen.
- »age-de.xml«-Auszeichnung: Ein Jugendschutzprogramm muss standardisierte maschinenlesbare Altersklassifizierungen (Label) auslesen, in allen technisch möglichen und zumutbaren Klassifizierungsvarianten richtig interpretieren und in altersdifferenzierte Blockaden umsetzen.
- Whitelist-Konzept für Kinder unter 12 Jahren (Empfehlung der KJM): Ein Jugendschutzprogramm soll mindestens die kindgeeigneten Adressen der fragFINN-Liste für alle Altersgruppen freischalten.
- Blacklist-Konzept für Jugendliche ab 12 Jahren (Empfehlung der KJM),
- Nutzerautonomie: Eltern sollen die Möglichkeit haben,
  - den Filter auszuschalten,
  - den Filter individuell zu konfigurieren,

[50] siehe http://www.yprt.eu/yprt/content/sections/, Stand 1.10.2012

[51] Nicht alle der im Kriterienkatalog der KJM erwähnten Anforderungen müssen erfüllt werden, damit eine Anerkennung ausgesprochen werden kann. Die nähere Betrachtung, welche der Anforderungen notwendige und welche hinreichende Kriterien darstellen, ist für die weitere Betrachtung im Rahmen der vorliegenden Studie jedoch unerheblich.

[52] Ungeeignete Inhalte sind in diesem Zusammenhang absolut und relativ unzulässige Inhalte i. S. d. § 4 JMStV sowie entwicklungsbeeinträchtigende Inhalte i. S. d. § 5 JMStV der Altersstufe »ab 18« [KJM, 2011a].

[53] Hierzu wird ein Vergleich mit verfügbaren Jugendschutzfiltern gezogen. Das betrachtete Jugendschutzprogramm muss sich im oberen Drittel des Leistungsspektrums von Jugendschutzfiltern bewegen [KJM, 2011a].

  - den Besuch von Webseiten explizit zu gestatten,
  - den Besuch von Webseiten explizit zu sperren und
  - alle Internetangebote, die nicht mit dem Labeling-Standard gekennzeichnet sind, zu blockieren.
- Plattform: Ein Jugendschutzprogramm soll für eine vom Anbieter definierte Menge von Endgeräten/Betriebssystemen funktionsfähig sein.
- Interoperabilität: Ein Jugendschutzprogramm darf keine Konflikten mit anderen Programmen hervorrufen.
- Aktualität: Die dem Jugendschutzprogramm zu Grunde liegenden Filterlisten müssen stets aktuell gehalten werden.
- Sicherheit: Ein Jugendschutzprogramm darf – mit einfachen Mitteln – nicht umgangen werden können.
- Usability: Ein Jugendschutzprogramm soll einfach zu installieren und zu bedienen sein.
- Grundeinstellungen: Ein Jugendschutzprogramm sollte über vorkonfigurierte Grundeinstellungen verfügen, die altersdifferenziert einen wirksamen Schutz realisieren.

Darüber hinaus betrachtet die KJM bei der Anerkennung von Jugendschutzprogrammen folgende nicht-technische Aspekte:

- Kontinuierliche Weiterentwicklung,
- Aktive Vermarktung und
- Finanzielle Ausstattung.

Im Februar 2012 hat die KJM erstmals zwei Jugendschutzfilter (*JusProg Jugendschutzprogramm* und *Kinderschutz Software der Deutsche Telekom AG*) unter Auflagen und mit einer Befristung von fünf Jahren als Jugendschutzprogramme anerkannt. Die KJM informiert dazu:

»Anbieter von entwicklungsbeeinträchtigenden Inhalten in Telemedien, die ihr Angebot für ein anerkanntes Jugendschutzprogramm programmiert haben, dürfen diese […] verbreiten, ohne weitere Jugendschutz-Maßnahmen – wie die Einhaltung bestimmter Zeitgrenzen, Vorschaltung eines technischen Mittels – ergreifen zu müssen (= Privilegierung).« [KJM, 2012d]

Außerdem gibt die KJM darüber Auskunft, in welchen Bereichen sie weiteres Verbesserungspotential sieht:

»[…] die Programme müssen regelmäßig überprüft, weiterentwickelt und an den neuesten Stand der Technik angepasst werden. Sie sollten beispielsweise auch auf Smartphones, auf Spielekonsolen und weiteren Plattformen verfügbar gemacht werden.« [KJM, 2012d]

Gegenwärtig sind beide Programme nur für Windows PCs verfügbar. Außen vor bleiben also beispielsweise auch Mac- und Linux-Computer oder Netzwerkgeräte wie Router. Darüber hinaus »müssen die Programme mittels eines Praxistests weiter auf ihre Benutzerfreundlichkeit überprüft werden« [KJM, 2012d].

1.4.1.1 JusProg Jugendschutzprogramm

Eins der beiden unter Auflagen anerkannten Jugendschutzprogramme ist das *JusProg Jugendschutzprogramm* des Vereins *JusProg e. V. – Verein zur Förderung des Kinder- und Jugendschutzes in den Telemedien e. V.*[54] Abbildung 1.2 zeigt das Hauptmenü des Programms. Zum Zeitpunkt der Un-

[54] siehe http://www.jugendschutzprogramm.de, Stand 1.10.2012



Abbildung 1.2
Screenshot JusProg Jugendschutzprogramm (Version 4.3.0.349).

tersuchung des Programms im Rahmen der vorliegenden Studie war es in der Version 4.3.0.349[55] verfügbar. Die Umsetzung der Software basiert auf der *Surf-Sitter* Software der *Cybits AG*.[56] Unterstützt werden nur aktuelle Windows-Betriebssysteme. Das Programm liegt in zwei Varianten vor. Zum einen eine kostenfreie Version mit Basis-Schutz, die den Anforderungen der KJM genügt, sowie eine Premiumversion, die außerdem die folgenden Funktionen beinhaltet:

- Surfzeit-Begrenzer (Begrenzung der maximalen Zeit zum Surfen im Internet, danach vollständiges Blockieren aller Webseitenzugriffe),
- Abzockschutz (spezielle Filterliste für Webseiten mit *besonders rüden Geschäftspraktiken*),[57]
- zusätzliche Filter (spezielle Filterlisten für Werbefallen, Filesharingportale und Proxyseiten),[58]
- zusätzliche Einstellungsmöglichkeiten, zum Beispiel Blockieren von Hypertext Transfer Protocol Secure (HTTPS) oder FTP Verbindungen, individuelle Konfigurationen für jedes Kind.

Der für die Premiumversion notwendige Lizenzschlüssel kann je nach Laufzeit (angegeben sind 5- und 12-monatig) und Anzahl der Installationen ab 9,90 Euro erworben werden. Das Programm nutzt die Filterlisten von fragFINN und das BPjM-Modul und bietet die Option zur Unterstützung der SafeSearch-Funktion von Google (vgl. Abschnitt 2.4.1). Außerdem unterstützt es eine Reihe von freiwilligen technischen Auszeichnungen für Web-Inhalte, namentlich die bereits erwähnte age-de.xml Kennzeichnung sowie die nicht fortgeführte Kennzeichnung der mittlerweile aufgelösten und im Family Online Safety Institute (FOSI) aufgegangenen Non-Profit Organisation Internet Content Rating Association (ICRA) [FOSI, 2012] und die *Restricted To Adults (RTA)* Auszeichnung der amerika-

[55] Das Programm wurde im Verlauf der Studie in 2012 kontinuierlich weiterentwickelt, Ende November 2012 war bereits die Version 4.3.0.376 verfügbar.
[56] siehe http://www.cybits.de/, Stand 1.10.2012
[57] vgl. http://www.jugendschutzprogramm.de/faq4.php#glossar_abzockseiten, Stand 1.10.2012
[58] vgl. http://www.jugendschutzprogramm.de/faq4.php#glossar_werbefallen, Stand 1.10.2012

nischen Non-Profit Organisation *Association of Sites Advocating Child Protection (ASACP)*.[59] Alle Kennzeichnungen werden durch das Programm automatisch ausgelesen und die Anzeige der Seite je nach Alterskonfiguration des aktiven Nutzers blockiert oder freigegeben. Im Gegensatz zu den altersdifferenzierten Labels age-de.xml und IRCA ist die RTA-Auszeichnung ausschließlich für »ab 18 Jahre« Webseiten vorgesehen. Die Software führt, abgesehen vom Auslesen der Kennzeichnungen, keine Echtzeitanalyse von Webseiten beim Benutzer durch. Die Entscheidungsgrundlage für das Blockieren oder Freischalten von Seiten sind die mitgelieferten Filterlisten, die individuellen Filterlisten der Nutzer sowie die genannten Alterskennzeichnungen. Bisher nicht bewertete oder eventuell falsch bewertete Webseiten können auf dem Webangebot des Anbieters gemeldet und zur Überprüfung markiert werden. Die während der Nutzung des Jugendschutzprogramms angesteuerten Webseiten können außerdem über eine optionale Funktion zur Produktverbesserung automatisch und anonymisiert an die JusProg-Server übertragen werden. Die Überprüfung von bisher nicht bewerteten Seiten wird dann zunächst technisch mit automatisierter Klassifizierung und anschließend bei relevanten Websites manuell durch Mitarbeiter bei JusProg durchgeführt. Nach eigenen Angaben bietet JusProg die eigenen, redaktionell gepflegten Filterlisten auch anderen Akteuren für Zwecke des Jugendmedienschutzes an.

1.4.1.2 Telekom Kinderschutz Software

Das andere unter Auflagen anerkannte Jugendschutzprogramm ist die *Kinderschutz Software* der *Deutsche Telekom AG*.[60] Abbildung 1.3 zeigt den Dialog, der Eltern zur Konfiguration der Software zur Verfügung steht. Zum Zeitpunkt der Studie war das Programm in der Version 3.01.0033 verfügbar. Die Kinderschutz Software wird von der Deutschen Telekom selbst vertrieben und entwickelt und ist auf allen aktuellen Windows-Betriebssystemen lauffähig. Die Software ist nur für Telekom-Kunden, allerdings dann kostenfrei, verfügbar. Diese Einschränkung soll laut Hersteller in 2013 aufgehoben werden. Das Programm verfügt über einen sehr ähnlichen Funktionsumfang wie das JusProg Jugendschutzprogramm und hat über die KJM-Bestimmungen hinaus folgende Möglichkeiten:

- Verwaltung von Kinderprofilen unabhängig von konfigurierten Windows-Benutzerprofilen,
- Konfiguration von Freigaben einzelner Kategorien (circa 60 Kategorien),
- Surfzeit-Steuerung (Eltern können zum einen Surfzeiten festlegen und bestimmen, wie lange Kinder innerhalb dieser Dauer maximal surfen dürfen),
- Beschränkung von herunterladbaren Inhalten (Musik-, Video-, Programmdateien, Dokumente, Tauschbörsen, Downloads),
- zusätzliche Einstellungsmöglichkeiten, zum Beispiel Blockieren von HTTPS oder FTP Verbindungen, individuelle Konfigurationen für jedes Kind.

Die Kinderschutz Software nutzt ebenfalls die Positivliste von fragFINN.de und die Negativliste des BPjM-Moduls sowie die der Firma IBM Internet Security Systems (ISS, ehemals Cobion[61]), die etwa 100 Millionen Webseiten in circa 60 Kategorien enthält. Auch hier werden die durch die KJM überprüften age-de.xml Labels ausgewertet und die Anzeige der Webseite je nach eingestelltem Kinderprofil mit Darstellung einer kindgerechten Hinweisseite blockiert. Des Weiteren sind die von Google, Bing, Yahoo und YouTube angebotenen SafeSearch-Filter sowie die *Restricted To Adults (RTA)* Auszeichnung integriert. Eltern haben die Möglichkeit, auf dem PC installierte Software zu blockieren. Die Kinderschutz Software führt ebenfalls keine Echtzeitanalyse von Webseiten auf dem Rechner des Nutzers durch. Im Unterschied zum JusProg Jugendschutzprogramm können Webseiten nicht an die Telekom bzw. einer anderen Stelle zur Überprüfung übermittelt werden.

[59] siehe http://www.rtalabel.org/ und http://www.asacp.org/, Stand 1.10.2012
[60] siehe http://www.t-online.de/kinderschutz, Stand 1.10.2012
[61] siehe http://www.iss.net/, Stand 4.10.2012



Abbildung 1.3
Screenshot Telekom Kinderschutz Software (Version 3.01.0033).

1.4.2 Jugendschutzfilter mit SIP-Bench Bewertung

Die Europäische Kommission hat im Rahmen des *Safer Internet Programme* (vgl. Abschnitt 1.2.3) mehrere Projekte zum Thema Jugendmedienschutz gefördert [European Commission, 2012b]. Im Rahmen der vorliegenden Status Quo Betrachtung sind dabei insbesondere das 2007 abgeschlossene *PEGI Online*, in dessen Rahmen eine Erweiterung für Online-Spiele für das EU-weit gültige System der Altersempfehlungen von Unterhaltungsmedien wie Filmen, Videos, DVDs und Computerspielen entstanden ist [European Commission, 2012a], und das Projekt *SIP-BENCH* [European Commission, 2012c] sowie dessen bis Ende 2012 laufenden Nachfolgers *SIP-BENCH II* [European Commission, 2012d] von Bedeutung. Das bereits veröffentlichte Ergebnis der Förderung ist die Studie *Benchmarking of parental control tools for the online protection of children - SIP-Bench II*, welche das Ziel verfolgt, »die wesentlichen Funktionalitäten, die Effektivität und die Benutzbarkeit (Usability) häufig genutzter Filterprodukte zum Jugendschutz in Europa vergleichend zu bewerten« [INNOVA Europe, 2012]. Partner der Studie sind das Consulting Unternehmen *Innova Europe*,[62] das italienische Unternehmen *Cybion Online Business Intelligence*[63] und die in Berlin ansässige *Stiftung Digitale Chancen*,[64] die unter anderem mit jugendschutz.net und der FSM (vgl. Abschnitt 1.2) Partner des beendeten EU-Projektes *Youth Protection Roundtable*[65] waren. Die Studie konzentriert sich auf Jugendschutzfilter für Endgeräte und umfasst im mittlerweile vierten von fünf Bewertungszyklen 22 PC- und 5 Mobil-Jugendschutzfilter, darunter auch viele namhafte Virenschutzprogramme mit eingebauter Jugendschutz-Funktion, zwei Betriebssystem-eigene Filterkomponenten sowie eine Filterlösung für Netzwerkgeräte. Die Bewertungskriterien konzentrieren sich auf vier Bereiche [Cybion et al., 2012]: Funktionalität, Effektivität, Usability und Sicherheit. Dabei werden die Bereiche folgendermaßen spezifiziert:

- Funktionalität: Bewertet den Funktionsumfang der Software, insbesondere die Unterstützung von altersdifferenzierten Konfigurationen, außerdem auch die Kompatibilität mit verschiedenen Plattformen, Mehrbenutzerunterstützung und weitere.
- Effektivität: Bewertet quantitativ die Filterleistung der Jugendschutzfilter, insbesondere in Hinblick auf Underblocking und Overblocking.
- Usability: Bewertet die Benutzerführung, die Konfigurations- und Updatemöglichkeiten durch den Nutzer und ähnliche Aspekte.
- Sicherheit: Bewertet die Widerstandsfähigkeit der Software gegenüber Versuchen, die Filterfunktion zu umgehen.

Es ist darauf hinzuweisen, dass die SIP-Bench Studie eine geringere Granularität bei der Altersklassifizierung aufweist. Während die age-de.xml Auszeichnung von einer fünfstufigen Einteilung ausgeht (»ab 0 Jahre«, »ab 6 Jahre«, »ab 12 Jahre«, »ab 16 Jahre«, »ab 18 Jahren«), wird in der Studie die Bewertung der Effektivität nur für die Stufen »bis 12 Jahre« und »ab 13 Jahre« unterschieden. Außerdem wurde die Filterleistung auf Basis von gesamteuropäischen, das heißt gemischtsprachigen Webangeboten untersucht. Insofern sind direkte Schlussfolgerungen über die Filterleistung bei deutschsprachigen Angeboten nicht direkt möglich. Die Kinderschutz Software der Deutschen Telekom war, im Gegensatz zum JusProg Jugendschutzprogramm, auch Teil der bewerteten Jugendschutzfilter – allerdings in einer Version 2.06.00.36. Laut Herstellerauskunft ist zwischen dieser und der aktuellen Version 3.01.0033 neben einer Neuumsetzung der Oberfläche auf Basis von Usability-Tests insbesondere die Anerkennung der KJM-Kriterien und damit eine Verbesserung der Filterleistung hinzugekommen. Einen Beitrag leisten dazu unter anderem die Unterstützung der age-de.xml Auszeichnung, eine Aktualisierung der Cobion-Liste sowie die Möglichkeit, auf dem Windows-PC installierte Software zu blockieren (vgl. Abschnitt 1.4.1.2). Insofern sind Implikationen

[62] siehe http://www.innova-europe.eu/, Stand 4.10.2012
[63] siehe http://www.cybion.it/, Stand 4.10.2012
[64] siehe http://www.digitale-chancen.de/, Stand 4.10.2012
[65] siehe http://www.yprt.eu, Stand 4.10.2012

beim Vergleich der gegenwärtig verfügbaren Software mit dem Testfeld nicht möglich. Ein Überblick über die 22 getesteten Jugendschutzfilter, die als PC-Programm zur Verfügung stehen, bietet Tabelle 1.2. Ausführliche Details und Bewertungen können der veröffentlichten Studie entnommen werden [Cybion et al., 2012].

An dieser Stelle sei noch auf einige besondere Funktionen bzw. Betriebskontexte hingewiesen. Die beiden Programme Mobicip und Davide sind die einzigen im Testfeld, die auch für Linux-Systeme zur Verfügung stehen. Zusätzlich zu den 22 aufgezeigten Programmen wurden außerdem noch die im Betriebssystem mitgelieferten Jugendschutzfilter von Microsoft (Windows Live Family Safety, ab Windows Vista) und Apple (Mac OS X Parental Controls) untersucht. Beide Programme schnitten bei der Bewertung im Mittelfeld ab.[66] Eine weitere Besonderheit im Testfeld war der Dienst OpenDNS des in San Francisco ansässigen, gleichnamigen Unternehmens.[67] Im Gegensatz zu den anderen Werkzeugen handelt es sich hierbei nicht um ein Programm, das auf einem Endgerät installiert wird, sondern um einen Domain Name System (DNS) Dienst, der zum Beispiel für einen gesamten Haushalt direkt in einem Router, also einem Netzwerkgerät, konfiguriert werden kann. Alle Internetzugriffe werden so über den OpenDNS Dienst umgeleitet und nur solche Seitenaufrufe weitergeleitet, die durch die Konfiguration der Eltern oder entsprechend einer Filterliste freigegeben worden sind. Hierdurch erhalten alle Geräte (PCs mit beliebigen Betriebssystemen, mobile Endgeräte, Spielekonsolen, usw.) automatisch den gleichen Schutz. Nicht berücksichtigt sind allerdings Verbindungen, die nicht über einen zentralen Router geführt werden – typischerweise Mobilfunkverbindungen in Smartphones. Die Studie attestiert dem Dienst zwar eine Filter-Effektivität im unteren Drittel des Testfelds [Cybion et al., 2012], dennoch stellt die Grundidee des Betriebs des Filters an einem zentraleren Punkt – weg vom Endgerät hin zur Quelle der Inhalte – eine Besonderheit dar, die in Kapitel 6 noch einmal näher betrachtet wird.

Im Wesentlichen basieren alle Programme auf vorgegebenen Filterlisten. In weniger Fällen bieten die Programme die Möglichkeit, eine Whitelist zu verwenden bzw. selber einzurichten, der überwiegende Teil blockiert Webseiten von den mitgelieferten Negativlisten. Das Programm *Owl 2012*, dass sowohl in der Studie bei der Filterleistung als auch in der Gesamtbewertung in beiden Altersstufen am Besten abgeschnitten hat, verfügt unter wenigen darüber hinaus über einen *On-The-Fly* Filter, der Webseiten vor der Darstellung beim Benutzer analysiert und beim Verdacht auf entwicklungsbeeinträchtigende Inhalte blockiert. Bei diesem Programm gibt es zwei Modi: Eine *Active Content Recognition (ACR)* zur Analyse von Webseiten und einen sogenannten *Active Chat Inspector (ACI)*, der für den Schutz vor Cyber-Mobbing in Online-Chats dienen soll. Sowohl der ACI als auch der ACR arbeiten mit statistischen Modellen, die textuelle (Auswahl von Wörtern, Wortklassen) und kontextuelle (Links ausgehend von und auf die Webseite u.ä.) Informationen ausnutzen [Puresight Technologies, 2012]. Details zu diesen und ähnlichen technischen Verfahren werden ab dem Kapitel 4 vertieft. In der SIP-Bench Studie wurden außerdem fünf Applikationen für mobile Geräte getestet. Eine Übersicht bietet Tabelle 1.3.

Alle aufgeführten Applikationen für mobile Endgeräte basieren auf Filterlisten und setzen für iOS-basierte Geräte voraus, dass der Standard-Browser des mobilen Betriebssysteme deaktiviert und dafür der alternative, geschützte Browser des Anbieters genutzt wird. Das Abschalten des Standard-Browsers kann über die in iOS integrierte Jugendschutzfunktionen erreicht werden. Auf Windows- und Android-basierten Geräten kann der Schutz unabhängig vom genutzten Browser eingesetzt werden. Die SIP-Bench Studie attestiert allen mobilen Versionen eine schlechtere Filterleistung gegenüber den PC-Versionen der Anwendungen. Speziell die sehr hohen Underblocking-Werte werden herausgestellt [Cybion et al., 2012]. Allerdings betrachtet die SIP-Bench Studie nur einen Teil der zum Zeitpunkt dieser Studie bereits verfügbaren mobilen Ableger der Jugendschutzfilter. Es

[66] Windows Live Family Safety: Platz 9/24 (Altersgruppe $\leq$ 12 Jahre), Platz 12/24 (Altersgruppe >12 Jahre); Mac OS X Parental Controls: Platz 17/24 (Altersgruppe $\leq$ 12 Jahre), Platz 14/24 (Altersgruppe >12 Jahre) [Cybion et al., 2012]

[67] siehe http://www.opendns.com/, Stand 5.10.2012

Tabelle 1.2
Jugendschutzfilter im Testfeld der SIP-BENCH II 4th Cycle Studie [Cybion et al., 2012].

Hersteller	Name der Software	Eff. (≤12)[a]	Gesamt (≤12)[b]	Eff. (>12)[c]	Gesamt (>12)[d]	Win[e]	Mac[f]	$\bar{G}$[g]
PureSight	Owl	2.30	2.67	2.10	2.74	Ja		2.71
Norton	Online Family	2.20	2.49	1.90	2.44	Ja	Ja	2.47
Kaspersky	Pure	2.00	2.30	1.50	2.18	Ja		2.24
Blue Coat	K9 Web Protection	1.80	2.13	1.60	2.17	Ja	Ja	2.15
ContentWatch	Net Nanny	1.50	1.95	1.50	2.17	Ja	Ja[h]	2.06
Trend Micro	Online Guardian	1.60	1.97	1.70	2.12	Ja		2.05
Deutsche Telekom	Kinderschutz Software	2.20	2.10	1.90	1.85	Ja		1.98
Profile Technology	Parental Filter 2	1.20	1.79	1.40	2.13	Ja		1.96
F-Secure	Internet Security 2012	1.60	1.88	1.70	2.01	Ja		1.95
McAfee	Safe Eyes	1.10	1.75	1.20	2.10	Ja	Ja	1.93
McAfee	Family Protection	1.10	1.71	1.20	2.05	Ja		1.88
SoftForYou	Cybersieve	1.00	1.64	1.00	1.94	Ja		1.79
AVG	Family Safety	1.40	1.74	1.30	1.71	Ja		1.73
Enologic	Net Filter	1.50	1.63	1.50	1.60	Ja		1.62
Mobicip	Mobicip	0.90	1.48	0.80	1.69	Ja		1.59
Xooloo	Xooloo Contrôle Parental	1.40	1.61	1.30	1.54	Ja		1.58
Cybersitter	Cybersitter	1.10	1.38	1.20	1.42	Ja		1.40
CyberPatrol	CyberPatrol Parental Controls	0.60	1.25	0.70	1.48	Ja		1.37
Associazione Davide Onlus	Davide	0.80	1.21	1.10	1.51	Ja	Ja	1.36
Westcoastclouc	Net-Intelligence	1.10	1.28	1.20	1.30	Ja		1.29
WhiteNet	WhiteNet Home	1.30	1.29	1.10	1.09	Ja		1.19
Optenet	Webfilter PC	0.80	1.07	1.10	1.25	Ja		1.16

[a] Effektivität auf einer Skala von 0-4 für Inhalte »unter 13 Jahre«
[b] Gesamtnote auf einer Skala von 0-4 für Inhalte »unter 13 Jahre«
[c] Effektivität auf einer Skala von 0-4 für Inhalte »ab 13 Jahre«
[d] Gesamtnote auf einer Skala von 0-4 für Inhalte »ab 13 Jahre«
[e] Verfügbar für aktuelle Windows-Systeme (XP/Vista/7)
[f] Verfügbar für aktuelle Apple Mac-Systeme
[g] Mittelwert [b d]
[h] Gemäß Herstellerseite für Mac verfügbar, in der Studie nicht berücksichtigt

Tabelle 1.3
Jugendschutzfilter für Mobilgeräte im Testfeld der SIP-BENCH II 4th Cycle Studie [Cybion et al., 2012].

Hersteller	Name der Software	iOS[a]	Android[b]	Windows[c]	Weitere[d]
Blue Coat	K9 Web Protection Browser	Ja, ab 4.2	Ja		
AVG	Family Safety Mobile	Ja, ab 4.3	Ja	Ja, ab 7.5	
F-Secure	Mobile Security		Ja	Ja, nur 6.x	Ja[e]
Mobicip	Safe Browser	Ja, ab 4.3	Ja, ab 2.3		
Norton	Online Family Mobile		Ja, ab 2.2		

[a] Verfügbarkeit für iOS-basierte Geräte (iPhone, iPad, iPod)
[b] Verfügbarkeit für Android-basierte Geräte
[c] Verfügbarkeit für Windows Mobile-basierte Geräte
[d] Verfügbarkeit für sonstige Mobilgeräte
[e] u.a. Nokia Symbian, siehe http://mobile.f-secure.de/downloads/trial/index.html, Stand 24.10.2012

zeichnet sich anhand der jüngeren Produkteinführungen ab, dass die Hersteller von Jugendschutzfiltern stärker in den mobilen Markt eindringen und entsprechende Lösungen für einzelne oder mehrere Betriebssysteme parallel anbieten. Beispielsweise hatte der Hersteller Blue Coat die Android Version der K9 Web Protection im Juni 2012 angekündigt [Blue Coat Systems, 2012]. AVG hat für seine Family Safety Produktlinie erst im Sommer 2012 eine Windows Phone Version veröffentlicht [AVG Technologies, 2012]. Die Mobile Security Produkte der Firma F-Secure wurden Ende Mai 2012 angekündigt [F-Secure GmbH, 2012]. Net Nanny hat Anfang 2012 einen Android-Filter veröffentlicht und ein weiteres iOS-Filterprogramm für 2012 angekündigt [ContentWatch Inc., 2011]. Günstig für diese Entwicklung ist auch, dass die wesentlichen Marktanteile für mobile Smartphone-Betriebssysteme sich immer stärker auf drei Anbieter konzentrieren. Das sind Google mit Android, Apple mit iOS und Microsoft mit Windows Mobile, das auf Nokia-Smartphones seit 2011 anstelle des eigenen Symbian-Systems vertrieben wird [Microsoft, 2011]. Der gemeinsame Marktanteil beträgt im Moment weltweit etwa 86,7 % [Statista, 2012b]. In Deutschland ist dieser Anteil durch die hohe Reichweite von Symbian-Geräten mit circa 69 % noch nicht so groß, allerdings hat sich der Anteil von Nokias Betriebssystem allein zwischen 2011 und 2012 von 42 % auf etwa 24 % nahezu halbiert [Joerges, 2012]. Der weltweite Marktanteil von Symbian-Geräten sinkt seit mehreren Jahren kontinuierlich [Research, 2011]. Außerdem ist das BlackBerry Betriebssystem zu nennen, dass jedoch durch den starken Fokus des Herstellers RIM auf den Businesssektor für den Jugendschutz wohl weniger relevant sein dürfte.

1.4.3 Weitere Jugendschutzfilter

Neben den KJM-anerkannten und SIP-Bench-bewerteten Programmen gibt es auf dem Markt durchaus noch weitere Jugendschutzfilter. Einige ausgewählte werden in diesem Abschnitt vorgestellt.

Das Webangebot der Kindersuchmaschine fragFINN[68] verweist auf eine Reihe von Jugendschutzfiltern, welche die fragFINN-Positivliste unterstützen. Dazu zählt zunächst der Cybits Surf-Sitter,[69] der unter anderem auch Grundlage des JusProg Jugendschutzprogramms ist. Der Surf-Sitter und die Premium-Version des JusProg Jugendschutzprogramms sind, abgesehen von der Meldefunktion für verdächtige Seiten (vgl. Abschnitt 1.4.1.1), zum Zeitpunkt der Studie funktional identisch. Weiterhin aufgeführt ist ein kostenloses Derivat des Surf-Sitters, die sogenannten *Cybits AG Kinderschutzsoftware*, das auf die Freigabe der fragFINN-Whitelist und einige wenige weitere Features wie Surfzeit-Begrenzungen reduziert ist. Cybits vertreibt über den Surf-Sitter hinaus auch eine Version für

[68] siehe http://www.fragfinn.de, Stand 1.10.2012
[69] siehe http://surf-sitter.de/Produkte/PC, Stand 1.10.2012

Windows-Mobile Smartphones, allerdings für die veralteten Betriebssystem Versionen 6.1 bis 6.3, sowie eigene Router, um mehrere Geräte in einem Haushalt über ein Netzwerkgerät zu schützen.

Ebenfalls aufgeführt ist die *Salfeld Kindersicherung 2012*[70] der Salfeld Computer GmbH. Diese ist insofern relevant, da sie sowohl die fragFINN-Liste als auch das BPjM-Modul und damit zwei wesentliche Bedingungen für eine KJM-Anerkennung enthält. Allerdings unterstützt die Software nicht die age-de.xml Kennzeichnung und ermöglicht auch nicht den geforderten altersdifferenzierten Zugang. Darüber hinaus ist sie funktional mit den SIP-Bench-bewerteten Jugendschutzfiltern vergleichbar.

Ebenfalls kommerzielle Security-Suites mit eingebauten Jugendschutzfiltern, die nicht unter SIP-Bench Bewertung stehen, sind der *BitDefender* ab der Variante *Internet Security* der BitDefender GmbH,[71] das Programm *Panda* ab Variante *Internet Security* der Panda Security S.L.,[72] das Programm *Family Safety* der amerikanischen BSecure Technologies, Inc.[73] sowie die *SecuritySuite* der Norman Data Defense Systems GmbH.[74] Die genannten Programme basieren ebenfalls auf vordefinierten und durch den Nutzer erweiterbare bzw. einschränkbare Filterlisten und vergleichbarem Funktionsumfang wie beispielsweise Zeitbegrenzungen.

Die in Berlin ansässige *TIME for kids Informationstechnologien GmbH* vertreibt in Deutschland eine Antivirus- und Filterlösung und richtet sich dabei konkret an Schulen.[75] Darüber hinaus bietet TIME for kids auch eine Kombination aus Hardware in Form von Routern und der Filtersoftware an. Die Software basiert auf der Cobion-Filterliste von IBM.[76] Die Jugendschutzlösung ist bereits deutschlandweit an verschiedenen Schulen in Betrieb.

Einen gänzlich anderen Ansatz verfolgt das amerikanische Startup *care4teen*.[77] Das seit 2011 aktive Unternehmen vertreibt eine kommerzielle Software zum Überwachen und Blockieren des Netzverkehrs auf Windows-PC und Android Systemen. Prinzipiell handelt es sich hierbei ebenfalls um eine listengestützte Software (White- und Blacklists), die um verschiedene Überwachungsmöglichkeiten für Eltern erweitert wird. Die Besonderheit liegt in der Zusammenstellung der Filterlisten, die auf einem Abonnement-basierten Eltern-Netzwerk aufsetzen, in dem kollaborativ Webseiten bewertet und diskutiert werden können. Zum Zeitpunkt dieser Studie waren etwa zehntausend Eltern an der Auswertung von rund drei Millionen Seiten beteiligt. Das ist selbstverständlich ein deutlich geringerer Umfang als in anderen Jugendschutzfiltern (vgl. Abschnitt 1.4.1.2, IBM-Filterliste mit rund 100 Millionen Seiten), allerdings ist die Bewertung auf viele Eltern verteilt und daher mit wachsender Abonnentenzahl potentiell geeignet, um sehr viele Webseiten manuell zu bewerten. Ebenso das Projekt *Web Of Trust*,[78] dass Nutzerbewertungen für Webseiten in den Kategorien Vertrauenswürdigkeit, Händlerzuverlässigkeit, Datenschutz und Jugendschutz sammelt. Allerdings stellt das Projekt keinen Jugendschutzfilter im Sinne dieser Studie bereit, sondern nur die Möglichkeit, Warnhinweise anzuzeigen, falls eine Seite als problematisch für Kinder und Jugendliche bewertet worden ist. Solche Social Tagging oder Crowdsourcing Ansätze sind ein weit verbreitetes Phänomen im Internet (besonders bei der Verschlagwortung und Kategorisierung von Web-Inhalten). Die Prinzipien, Vorteile und Risiken im Rahmen des technischen Jugendmedienschutzes werden in Kapitel 6 im Einzelnen besprochen.

[70] siehe http://salfeld.de/software/kindersicherung/, Stand 1.10.2012
[71] siehe http://www.bitdefender.de/solutions/, Stand 1.10.2012
[72] siehe http://www.pandasecurity.com/security-promotion/germany/panda-internet-security/, Stand 1.10.2012
[73] siehe http://www.bsecure.com/, Stand 1.10.2012
[74] siehe http://www.norman.com/, Stand 1.10.2012
[75] siehe http://www.time-for-kids.de, Stand 1.10.2012
[76] siehe http://www.iss.net/, Stand 2.10.2012
[77] siehe http://www.care4teen.com/, Stand 1.10.2012
[78] siehe http://www.mywot.com/, Stand 1.11.2012

Abschließend sei noch erwähnt, dass eine Reihe von Netzwerkgeräte-Herstellern mittlerweile Router und WLAN-Router mit eingebauten Jugendschutzfiltern vertreiben. Darunter fallen auch solche mit Unterstützung für das BPjM-Modul, zum Beispiel verschiedene Versionen der AVM FRITZ!Box[79] und einige Router-Serien der DrayTek GmbH[80].

1.5 Kinder- und Jugendmedienschutz aus Sicht der Eltern

Im Bereich des technischen Jugendmedienschutzes kommen bereits Verfahren und organisatorische Abläufe zum Einsatz, die ohne aktive Beteiligung von Eltern funktionieren. So werden beispielsweise die durch die BPjM indizierten Internetangebote von Suchmaschinenherstellern aus der Ergebnismenge einer Suchanfrage entfernt und jugendschutzrelevante Internetangebote etwa durch das Engagement von jugendschutz.net abgeschaltet. Allein in 2011 wurden von jugendschutz.net durch Hinweise und eigene Überprüfungen etwa 10 000 unzulässige Angebote identifiziert. Die eingeleiteten Maßnahmen führen in rund 80 % der Fälle zur umgehenden Löschung oder Änderung der Angebote im Sinne des Jugendmedienschutzes [jugendschutz.net, 2011].

Über diese zentral organisierten Instrumente hinaus können technische Verfahren zum Jugendmedienschutz in Schutzprogrammen zum Einsatz kommen, die auf den verwendeten Endgeräten oder der Heim-Infrastruktur laufen. Die ZDF Medienforschung hat in Kooperation mit dem Hans-Bredow-Institut (HBI) 2011 eine repräsentative Elternbefragung durchgeführt und ermittelt, welchen Schutzbedarf Eltern wahrnehmen und inwieweit sie die Ziele des Jugendmedienschutzes teilen, konkrete Regelungen kennen und entsprechend im Alltag umsetzen ([Hasebrink et al., 2011], [Hasebrink et al., 2012]). Die Studie war medienübergreifend angelegt und betrachtet die Sorgen der Eltern und die in der Erziehungspraxis angewandten Maßnahmen für die Einzelmedien Fernsehen/Film, Internet und Computer-/Konsolenspiele.

Der Studie zu Folge nutzen nahezu alle Kinder und Jugendliche das Internet, bei den 12- bis 13-Jährigen sind es bereits 94 % und der Anteil nimmt mit jeder Altersstufe zu. Rund drei Viertel der Eltern äußern Sorgen hinsichtlich der Internetnutzung ihrer Kindern, zu den am häufigsten genannten potenziell beeinträchtigenden Inhalten gehören Gewalt, vulgäre Sprache und verstörende oder beängstigende Inhalte. Am häufigsten sorgen sich Eltern der 12- bis 13-jährigen Kinder (87 %), dabei beruhen die Sorgen bei der Hälfte der Eltern auf eigenen Erfahrungen. Hinsichtlich der Nutzung anderer Medien (Computerspiele, Fernsehen, Handy) zeigen sich ebenfalls sehr viele Eltern besorgt, das Internet steht jedoch deutlich im Vordergrund der Besorgnis. Über 90 % der Eltern sind der Meinung, dass Kinder und Jugendliche im Alter ihres Kindes vor möglichen negativen Einflüssen der Medien geschützt werden und dass potentiell beeinträchtigende Inhalte nicht frei verfügbar sein sollten. Fast alle Eltern sehen es als ihre eigene Aufgabe an, ihre Kinder zu schützen. Ein überwiegender Teil der Eltern sieht ebenso Programm- und Inhalteanbieter in der Pflicht, zum Wohle der Kinder und Jugendlichen einen verantwortlichen Umgang mit den Medien zu praktizieren. Darüber hinaus erwartet auch ein großer Teil der Elternschaft Unterstützung im Jugendmedienschutz und die Förderung der Medienkompetenz Heranwachsender durch die Politik und Ministerien/Behörden sowie Kindergärten und Schulen.

Befragt nach der Kenntnis über bestehende Schutzmaßnahmen und deren Nutzung in der täglichen Erziehungspraxis geben über 90 % der Eltern an, dass sie mit den Kennzeichnungen der Altersfreigaben von FSK und USK vertraut seien und diese als Orientierungshilfe für die Mediennutzung ihrer Kinder verwendeten. Altersdifferenzierende Sendezeitregelungen für das Fernsehen sind drei Vierteln der Bevölkerung bekannt, auch diese werden als hilfreich angesehen. Mit technischen Vor-

[79] siehe http://www.avm.de/de/Produkte/FRITZBox/index.php, Stand 1.10.2012
[80] siehe http://www.draytek.de/jugendschutz.html, Stand 1.10.2012

einstellungen für den Jugendschutz an Computern oder Fernsehgeräten sind nur die Hälfte der Eltern vertraut. Jugendschutzsoftware ist weniger als der Hälfte der Befragten bekannt.[81] 25 % der Befragten geben an, Jugendschutzsoftware für den Internetzugang einzusetzen, dabei können Eltern häufig nicht genau benennen, wie die eingesetzte Lösung funktioniert (z. B. Surfzeitbegrenzer, Voreinstellungen im Betriebssystem, Jugendschutzkomponenten in Internet-Security-Suiten[82]). Etwa die Hälfte der Nutzer empfindet den Schutz durch die genutzten Programme als wirksam.

Eltern messen der Medienkompetenz ihrer Kinder eine hohe Bedeutung bei (75 % der Befragten): »Am besten sind Kinder und Jugendliche geschützt, wenn sie ausreichend über Risiken informiert sind und wissen, wie sie sich vorsehen können.« [Hasebrink et al., 2012] Zur Förderung der Medienkompetenz ihrer Kinder wenden Eltern verschiedene erzieherische Maßnahmen an. Am Beispiel des Fernsehens sind das aktive Mediation, indem Kindern Inhalte von Sendungen erläutert werden, die gemeinsame Nutzung des Mediums Fernsehen und eine passive Mediation durch das Aufstellen von Geboten und Verboten hinsichtlich der Fernsehnutzung.

Die Studie des ZDF kommt zum Schluss, dass Eltern durchaus problematische Aspekte in der Mediennutzung ihrer Kinder sehen, insbesondere im Hinblick auf das Internet. Für das Medium Film/Fernsehen stehen den Eltern einige hilfreiche Schutzmaßnahmen zur Verfügung, die sie in der Erziehung ihrer Kinder auch praktisch verwenden beziehungsweise akzeptieren.[83] Die Nutzung des Fernsehens wird von den Eltern mehr eingeschränkt als die Internetnutzung. Die für das Internet zur Verfügung stehenden – im Falle von Filterprogrammen und Jugendschutzsoftware freiwilligen – Schutzmaßnahmen sind weniger bekannt und deren Funktionsweise wird als teilweise unzuverlässig wahrgenommen. Auch ist das Inhalte-Angebot im Internet unkontrollierter im Vergleich zum Fernsehen. Ein überwiegender Teil der Eltern, deren Kinder regelmäßig das Internet nutzen, glaubt nicht, dass die Internetgemeinschaft selbständig eine wirksame Kontrolle zum Schutz von Kindern und Jugendlichen ausübt. Eltern halten die bestehenden Schutzmaßnahmen für zu lasch und fordern Verschärfungen.

Zu einem ähnlichen Ergebnis kommt auch die Studie *Jugendschutzsoftware im Elternhaus: Kenntnisse, Erwartung und Nutzung* [HBI, 2011]. Im Auftrag des BMFSFJ hat das HBI 2011 aktuelle Umfragen und Studien ausgewertet. Das Ergebnis ist eine übersichtliche Bestandsaufnahme dessen, was Eltern in Deutschland in Bezug auf Jugendschutzsoftware wissen, in wieweit sie diese einsetzen und welche Erwartungen sie haben. Neben der bereits erwähnten Studie des ZDF [Hasebrink et al., 2012] wurden v.a. die Ergebnisse des Forschungsprojekts *EU Kids online* [EU Kids Online, 2010b], die von der forsa 2011 durchgeführte Studie *Jugendschutzprogramme aus Sicht der Eltern* [forsa, 2011] und die vom Medienpädagogischen Forschungsverbund Südwest (MPFS) durchgeführten *KIM-/JIM-Studien* der vergangenen Jahre[84] für die Bestandsaufnahme herangezogen. Der Studie zufolge haben viele Eltern bereits von Jugendschutzsoftware gehört, aber nur wenige setzen sie auch tatsächlich ein. Fast allen Eltern ist es wichtig, dass es Jugendschutzsoftware gibt. Eltern sehen sie als wichtiges Hilfsmittel in ihrem Erziehungskonzept an, fühlen sich aber mit der Auswahl und der Anwendung überfordert. Die Hälfte der Elternschaft wünscht sich mehr Informationen über Jugendschutzsoftware. Diese werden sowohl von den Internetanbietern als auch von öffentlichen Einrichtungen (Schulen, Kindergärten, Behörden und Beratungsstellen) und von den Medien eingefordert.

[81] Anmerkung: Die ZDF Studie wurde vor der ersten Anerkennung von Jugendschutzprogrammen durch die KJM durchgeführt.

[82] Internet-Security-Suiten vereinen verschiedene Schutzmechanismen (Virenschutz, Firewall, Anti-Spam, Kinder-/Jugendschutz) in einem Softwarepaket.

[83] Gemeint sind hier die geltende altersdifferenzierte Sendezeitbeschränkung beziehungsweise die Altersverifikationssysteme für manche (digitalen) Pay-TV- und Video-on-Demand Angebote.

[84] Quellen: KIM-Studie 2006 ([MPFS, 2007], [Feierabend und Klingler, 2007]), KIM-Studie 2008 ([MPFS, 2009], [Feierabend und Klingler, 2009]), KIM-Studie 2010 [MPFS, 2011b], JIM-Studie 2011 [MPFS, 2011a].

Zusammenfassung der Elternsicht Eltern möchten ihre Kinder vor Gefahren schützen. Hierzu bedienen sie sich auch gerne technischer Mittel, soweit sie in das eigene Erziehungskonzept passen und von den Eltern benutzbar sind. Schwierigkeiten entstehen für Eltern vor allem aus der Unkenntnis über Gefahren, mit denen ihre Kinder im Internet konfrontiert werden könnten, und der scheinbar verwirrenden Marktsituation rund um den technischen Jugendmedienschutz. Damit verbunden ist auch eine Information über die Wirkungsweise mit ihren Möglichkeiten und ihren Grenzen [Dreyer und Hajok, 2012] und die Abgrenzung von Kinder- und Jugendschutzsoftware gegenüber anderen Sicherheits- oder Überwachungsprogrammen. Die bislang vorliegenden Erkenntnisse aus der Forschung erheben nicht explizit Anforderungen von Eltern an Jugendschutzsoftware, jedoch lassen sich aus den Marktstudien über vorhandene Lösungen Differenzierungsmerkmale ableiten, die Eltern und Erziehungsbeauftragten als Orientierung für die Auswahl einer Lösung dienen können. Dabei können die folgenden Bereiche unterschieden werden (vgl. [Cybion et al., 2012]):

1. Funktionsumfang – zum Beispiel verschiedene Nutzerprofile, nutzerspezifische Anpassbarkeit, Zeitlimits, Nutzungsrestriktionen (Programme, Protokolle, Nutzung von Whitelists, Berücksichtigung von Blacklists, …),
2. Sicherheit – die Software darf nicht von einem Kind oder Jugendlichen umgangen werden können,
3. Effektivität – Zuverlässige Sperrung von Inhalten hinsichtlich von Themen, Altersgruppen, Sprache, User Generated Content (UGC) in Web 2.0-Angeboten (z. B. Film- und Fotoportale), Internet-Kommunikation (z. B. Chats),
4. Usability (Gebrauchstauglichkeit) – Installation, Konfiguration, Deinstallation.

1.6 Jugendmedienschutz aus Sicht von Kindern und Jugendlichen

Verbote und Gebote im Umgang mit den Medien werden von Kindern und Jugendlichen als Teil der Erziehung durch ihre Eltern akzeptiert. Unverständnis herrscht aber teilweise über die Zuordnung von Inhalten und Angeboten in die jeweilige Altersgruppe, da das subjektive Empfinden, welche Angebote für sie selbst ungeeignet erscheinen, durchaus von den Einschätzungen der professionellen Prüfer und denen der Eltern abweicht. Unbenommen ist auch, dass *Verbotenes* eine gewisse Neugier bei Kindern und Jugendlichen weckt, erfahren zu wollen, was es denn tatsächlich ist, wovon sie ferngehalten werden sollen.

Die Ergebnisse einer europaweiten empirischen Untersuchung zur Onlinenutzung durch Kinder und Jugendliche[85] im Rahmen des Projektes *EU Kids Online*[86] zeigen jedoch auf, dass ältere Kinder und Jugendliche selbst ein Schutzbedürfnis entwickeln, insbesondere wenn sie bei der Nutzung des Internets Erfahrungen gemacht haben, die sie unangenehm berührt oder verletzt haben. Rund ein Viertel der 11- bis 16-Jährigen gibt an, Erfahrungen mit Inhalten wie Hassforen, Pro-Anorexia, Selbstverletzung, Drogen oder Selbstmord gehabt zu haben. Sexuelle oder pornografische Inhalte, mit denen Kinder und Jugendliche neben dem Internet auch über Fernsehen, Film und Videos in Kontakt kommen, empfinden ältere Kinder als weniger belastend, für jüngere Kinder sind diese Bilder aber deutlich beunruhigender. Im Rahmen der Studie wurden auch die Eltern über die Kenntnis der Mediennutzung ihrer Kinder befragt. Über die Hälfte der Eltern von Kindern, die Kontakt mit unangenehmen oder verletzenden Inhalten hatten, wussten davon nichts. Aber die meisten betroffenen Kinder (etwa 80 %) haben sich mit dem Erlebten auseinander gesetzt und Unterstützung gesucht. Am häufigsten wurde sich mit Freunden ausgetauscht, teilweise auch mit den Eltern. Die Hälfte der betroffenen Kinder hat technische Maßnahmen angewandt, um sich zukünftig zu schützen (z. B. Löschen von Nachrichten, Blockieren von Absendern, Ändern persönlicher Einstellungen oder Kon-

[85] Zusammenfassung der ersten Befunde [EU Kids Online, 2010a], ausführlicher Bericht [EU Kids Online, 2010b].
[86] siehe http://www.eukidsonline.de, Stand 5.9.2012

taktdaten, Anpassung von Internet-Filtern). Ein Viertel der Betroffenen hat das Internet für eine Weile lang gar nicht mehr genutzt.

Im Vergleich zum europäischen Umfeld legt die Studie dar, dass deutsche Kinder und Jugendliche das Internet weniger häufig und weniger vielfältig nutzen. »Dies geht einher mit einer geringeren Wahrscheinlichkeit, im Internet unangenehme Erfahrungen zu machen. In dieser Hinsicht gehört Deutschland damit einer Gruppe von Ländern mit niedriger Internetnutzung und geringerem Risiko an.« [EU Kids Online, 2010a]

Die Erkenntnisse der Studie belegen die Aussage der in der Studie des Zweites Deutsches Fernsehen (ZDF) [Hasebrink et al., 2012] befragten Elternschaft aus Deutschland, dass medienkompetente Kinder und Jugendliche auch mit negativen Erfahrungen umzugehen wissen, sich selbst vor erneutem Kontakt mit belastenden Inhalten schützen wollen und sich im Rahmen ihrer Möglichkeiten durchaus zu helfen wissen. Auch die öffentliche Hand teilt die Ansicht, dass bei Jugendlichen durch das Erlernen eines bewussten Umgangs mit dem Internet der Selbstschutz gefördert wird. So werden beispielsweise im von der LfM geförderten Projekt *Medienscout NRW*[87] Jugendliche zu Medienscouts ausgebildet mit dem Ziel, ihr Wissen um den sicheren Medienumgang an Mitschüler zu vermitteln und sie bei Fragen zu unterstützen. Die Ergebnisse der Pilotphase sind durchaus vielversprechend:

»Von Januar 2011 bis April 2012 wurden im Rahmen des Pilotprojekts bereits 40 junge Mediennutzer im Schulkontext erfolgreich zu sog. ›Medienscouts‹ ausgebildet. […] Durch die Qualifizierung einer vergleichsweise kleinen Schülergruppe konnten weitreichende Effekte erzielt werden, da die ausgebildeten Medienscouts anschließend selbst als Referenten agieren und ihre Mitschüler qualifizieren sowie diesen als Ansprechpartner bei medienbezogenen Fragen und Problemen dienen. Heranwachsende sollen jüngeren Schülerinnen und Schülern beratend bei ihrer Mediennutzung zur Seite stehen und Fragen, die sich für junge Nutzer rund um die Themen Social Web, Internet & Co. ergeben können, beantworten.« [LfM, 2012]

Zusammenfassung der Sicht von Kindern und Jugendlichen Kinder und Jugendliche haben das Grundrecht auf informationelle Selbstbestimmung und möchten in ihren Freiheiten möglichst wenig beschränkt werden. Sie sind aber grundsätzlich bereit, erzieherisch aufgestellten Regeln Folge zu leisten. Opfer von verstörenden Inhalten im Internet entwickeln ein eigenes Schutzbedürfnis, bei ihnen ist die Bereitschaft zur Akzeptanz von Jugendschutzsoftware besonders ausgeprägt.

1.7 Zusammenfassung

Der aktuelle technische Jugendmedienschutz in Deutschland stellt sich als ein breites Spektrum rechtlicher Rahmenbedingungen, Organisationen und Möglichkeiten der praktischen Durchführung dar. Abschnitt 1.1 hat aufgezeigt, dass es eine umfassende rechtliche Grundlage für den Jugendschutz allgemein und den Jugendmedienschutz im Speziellen gibt. Die Durchsetzung des Jugendschutzes im Sinne der Verfassung und der entsprechenden Gesetze ist klare Aufgabe des Staates. Der Jugendmedienschutz basiert auf Staatsverträgen und Richtlinien zwischen den Ländern beziehungsweise den Landesmedienanstalten. Obwohl die jeweiligen Rahmenwerke von Bund und Ländern aufeinander aufbauen, sind die zahlreichen Aktivitäten im operativen Bereich nicht vergleichbar hierarchisch strukturiert.

Dieser Pluralismus spiegelt sich in der Umsetzung der Rahmenbedingungen wieder. Abschnitt 1.2 hat gezeigt, dass auf Bundes- und Länderebene zahlreiche Einrichtungen und Initiativen existieren, die sich in ihren Kompetenzen und Zielen zum Teil durchaus überschneiden. Diese Initiativen reichen

[87] siehe http://www.medienscouts-nrw.de, Stand 5.9.2012

von der bundesweiten Förderung der Medienkompetenz von Kindern und Jugendlichen über die freiwilligen Selbstkontrollen der Medienbranche bis hin zum Engagement von Unternehmen der freien Wirtschaft. Darüber hinaus gibt es auch europäische Initiativen, die jedoch nicht im Fokus dieser Studie stehen.

Für den technischen Jugendmedienschutz im Internet steht Eltern eine große Auswahl an Werkzeugen zur Verfügung. Zum Teil sind die entsprechenden Filterprogramme bereits unter Begutachtung öffentlicher Stellen, welche die Umsetzung des Jugendmedienschutzes im Internet verfolgen. Die KJM als zentrales Organ der Landesmedienanstalten hat dazu zwei Jugendschutzprogramme gemäß dem JMStV unter Auflagen anerkannt. Anbieter von entwicklungsbeeinträchtigenden Inhalten können sich durch die Programmierung ihrer Angebote gemäß den Vorgaben der KJM beziehungsweise des JMStV zumindest in Deutschland gesetzlich absichern. Diese Privilegierung der Anbieter gilt momentan nur für Inhalte bis einschließlich der Altersstufe »ab 16 Jahre«. Weiterhin gibt es eine davon unabhängige Auswahl an Jugendschutzfiltern für verschiedene Endgeräte wie PCs, Smartphones und Netzwerkgeräte, die in ihrem Funktionsumfang und in ihrer Filterleistung unterschiedliche Schwerpunkte setzen und auch unterschiedliche Qualitätsgrade erreichen. Bis auf wenige Programme, die Echtzeit-Analysefunktionen bieten, basieren diese Programme im Wesentlichen auf Filterlisten, die beim Abruf von Angeboten im Internet durch die Programme abgeglichen werden und gegebenenfalls den Zugriff auf diese Angebote einschränken. Diese, zum Teil altersdifferenzierten, Filterlisten sind in der Regel spezifisch für die einzelnen Hersteller. Mit den Listen des BPjM-Moduls und der fragFINN-Positivliste gibt es bisher zwei Beispiele für redaktionell gepflegte Filterlisten, die von allen Herstellern im Rahmen von Kooperationsprojekten verwendet werden können.

KAPITEL 2

ANFORDERUNGEN AN TECHNISCHE VERFAHREN ZUR DETEKTION JUGENDSCHUTZRELEVANTER WEB-INHALTE

Dieses Kapitel zeigt bestehende Anforderungen an technische Verfahren zur Detektion jugendschutzrelevanter Web-Inhalte auf. Diese dienen im weiteren Verlauf der Studie zur Prüfung der Eignung von bereits vorhandenen technischen Verfahren für den Einsatz im technischen Jugendmedienschutz.

Die in Abbildung 2.1 illustrierte Abgrenzung von Begriffen im Kontext des technischen Jugendmedienschutzes dient der besseren Orientierung. Entlang dieser Struktur ist das aktuelle Kapitel aufgebaut. *Jugendschutzsoftware* bezeichnet technische Mittel, mit denen Kinder und Jugendliche vor dem Zugriff auf beeinträchtigende oder jugendgefährdende Inhalte geschützt werden sollen (Näheres hierzu vgl. Abschnitt 2.3). Als *technische Jugendschutzstrukturen* werden alle technisch umgesetzten und technologisch nutzbaren Maßnahmen zur Förderung des Jugendmedienschutzes im Internet bezeichnet (Näheres hierzu vgl. Abschnitt 2.4).

Klassifikationsverfahren bezeichnen im Sinne der vorliegenden Studie von Computern eigenständig ausführbare Programme, die in der Lage sind, einen zu untersuchenden Web-Inhalt (zum Beispiel eine Webseite) vorgegebenen Klassen von Web-Inhalten zuzuordnen. *Technische Verfahren zur Detektion jugendschutzrelevanter Web-Inhalte* sind dementsprechend Klassifikationsverfahren, die einen zu untersuchenden Web-Inhalt entweder der Klasse der jugendschutzrelevanten Web-Inhalte oder der Klasse der nicht jugendschutzrelevanten Web-Inhalte zuordnen können.

Im Folgenden werden die Anforderungen an Klassifikationsverfahren zur Detektion jugendschutzrelevanter Web-Inhalte systematisiert betrachtet. Dabei wird grundsätzlich zwischen *funktionalen Anforderungen* (vgl. Abschnitt 2.1) und *nicht-funktionalen Anforderungen* (vgl. Abschnitt 2.2)[1] unterschieden. Ergänzt wird die Betrachtung durch Exkurse in die Technologiebereiche *Anforderungen an Jugendschutzsoftware* (vgl. Abschnitt 2.3) und *Anforderungen an technische Jugendschutzstrukturen* (vgl. Abschnitt 2.4), die über den eigentlichen Fokus der vorliegenden Studie herausreichen. Die Exkurse dienen der Identifikation impliziter Anforderungen an Klassifikationsverfahren, die sich zum Beispiel aus Fragen der Integration von Verfahren in existierende Programme oder technische Strukturen ableiten.

Quellen für Anforderungen zum technischen Jugendmedienschutz Abgeleitet wurden die im Folgenden dargestellten Anforderungen aus verschiedenen Quellen. Hierzu wurden die in Kapitel 1 zitierten Dokumente der verschiedenen Interessensgruppen im Kontext des Jugendmedienschutzes, die aus ihren eigenen Betrachtungswinkeln heraus teilweise auch unterschiedliche Bedürfnisse

[1] »Funktionale Anforderungen geben an, was das Softwaresystem oder einzelne seiner Komponenten tun sollen (und was nicht). Typischerweise werden diese Anforderungen definiert aus der Funktionsperspektive (Systeminput und -output, Fehlersituationen), aus der Datenperspektive (Datenstrukturen und Integritätsbedingungen) und aus der Verhaltensperspektive (Systemzustände, Zustandsübergänge, Ereignisse).« [Pohl, 2010]
Eine *nichtfunktionale Anforderung* legt fest, welche *Eigenschaften* das Produkt haben soll, zum Beispiel in welcher Zeitspanne eine bestimmte Systemreaktion erfolgen soll.

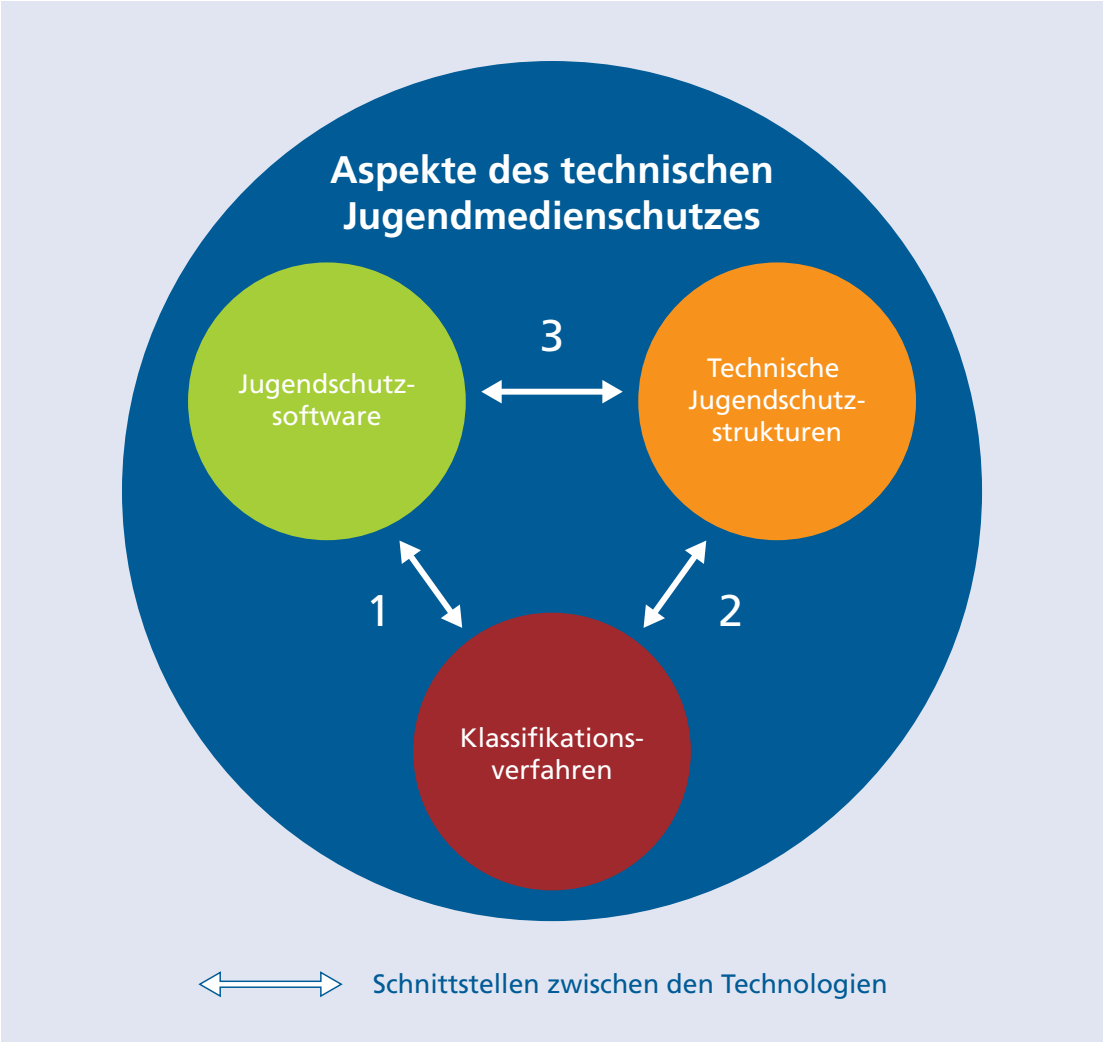


Abbildung 2.1
Jugendschutztechnologien und Schnittstellen

hinsichtlich des technischen Jugendmedienschutzes haben, konsultiert. Desweiteren wurde im Rahmen der Studie eine kooperative und informelle Befragung mit den Herstellern der beiden durch die KJM anerkannten Jugendschutzprogramme durchgeführt. Dabei handelt es sich einerseits um das *JusProg Jugendschutzprogramm* des JusProg e. V. sowie die *Kinderschutz Software* der Deutschen Telekom (vgl. Abschnitt 1.4.1). Ziel der Befragung war eine Konkretisierung der funktionalen und nicht-funktionalen Anforderungen an Klassifizierungsverfahren für den Jugendmedienschutz, die bereits in Jugendschutzprogrammen integriert sind oder in Zukunft integriert werden sollen.

2.1 Funktionale Anforderungen an Klassifikationsverfahren

Der Begriff *funktionale Anforderung* entstammt der Informatik und dient der systematischen Klassifikation von Anforderungen an ein System oder einen Dienst. »Eine funktionale Anforderung definiert eine vom System beziehungsweise von einer Systemkomponente bereitzustellende Funktion oder einen bereitzustellenden Service. Als Benutzeranforderung kann eine funktionale Anforderung sehr allgemein beschrieben sein. Als Bestandteil einer Spezifikation beschreibt eine funktionale Anforderung detailliert Eingaben und Ausgaben sowie bekannte Ausnahmen.« [itemis, 2012]

Die Untersuchung der Quellen hat ergeben, dass zwei wesentliche Anforderungen an technische Verfahren im Sinne der vorliegenden Studie existieren:

1. Klassifikationsverfahren müssen jugendschutzrelevante Web-Inhalte detektieren können.
2. Klassifikationsverfahren müssen jugendschutzrelevante Web-Inhalte altersdifferenziert klassifizieren können.

Auf Grund der Vielfältigkeit möglicher jugendschutzrelevanter Web-Inhalte ergibt die nähere Betrachtung dieser beiden grundsätzlichen Anforderungen eine Vielzahl von ableitbaren Detailanforderungen. Die folgenden Unterabschnitte betrachten die beiden Anforderungen im Detail.

2.1.1 Arten und Kategorien von Web-Inhalten

Die Anforderung *Klassifikationsverfahren müssen jugendschutzrelevante Web-Inhalte detektieren können* ist sicherlich die wesentliche und grundsätzliche Voraussetzung für die maschinelle Unterstützung der Klassifikation von Web-Inhalten. In der heutigen Praxis des Jugendmedienschutzes (vgl. Abschnitt 1.4) wird diese Art der Bewertung von Web-Inhalten nur teilweise maschinell vorgenommen. Häufig, insbesondere wenn die existierenden technischen Verfahren keine eindeutigen oder zuverlässigen Ergebnisse liefern, findet eine manuelle Bewertung durch die redaktionelle Arbeit von Jugendmedienschützern statt. Der Zweck automatischer Verfahren zur Klassifikation von Web-Inhalten ist die Ergänzung und Entlastung der redaktionellen Arbeit durch eine weitgehende Automatisierung, so dass möglichst viele jugendschutzrelevante Inhalte detektiert werden können. Jedes technische Verfahren verfügt über Grenzen der Zuverlässigkeit, so dass eine vollständige automatische Klassifikation sämtlicher Web-Inhalte auf absehbare Zeit nicht möglich sein wird. Insofern wird eine (ergänzende) redaktionelle Bewertung auch zukünftig immer notwendig sein.

Der Begriff *jugendschutzrelevante Web-Inhalte* umfasst im Sinne der vorliegenden Studie die gesamte Bandbreite von Web-Inhalten, wie sie im § 4 JMStV (unzulässige Angebote) und § 5 JMStV (entwicklungsbeeinträchtigende Angebote) definiert ist. Unzulässige Inhalte lassen sich im Sinne des JMStV den in Abschnitt 1.1.4.1 dargestellten (thematischen) Kategorien zuordnen.

Für eine derartige Kategorisierung ist allerdings zu berücksichtigen, dass im internationalen Vergleich unterschiedliche Gesetzgebungen gelten. Für den technischen Jugendmedienschutz bedeutet das letztendlich, dass Filterprogramme auch vor solchen Angeboten schützen müssen, die zwar nach deutschem Recht unzulässig aber im jeweiligen Betreiberland durchaus legitim und demnach im Internet frei verfügbar sein können.

2.1.1.1 Absolut unzulässige Inhalte

Absolut unzulässige Inhalte im Sinne des § 4 Absatz 1 JMStV sollten aufgrund der Tatsache, dass sie einem Verbreitungsverbot unterliegen, gar nicht online zu finden sein. Die Tatsache, dass derlei Inhalte von deutschen Anbietern im WWW dennoch zu finden sind, betrifft daher nicht nur den Jugendmedienschutz, sondern auch und vor allem wegen der strafrechtlichen Relevanz die entsprechenden Strafverfolgungsbehörden. Daher sollte es nicht vorrangige Aufgabe von Verfahren für den Jugendmedienschutz sein, entsprechende Inhalte zu identifizieren.

In der täglichen Praxis stoßen die Jugendmedienschützer im WWW allerdings häufig auf entsprechende Angebote. jugendschutz.net berichtet, dass im Jahr 2011 jedes fünfte überprüfte Angebot unzulässige Inhalte enthielt (etwa 10 000 Fälle), etwa 27 % davon betrafen Verstöße in Deutschland [jugendschutz.net, 2011]. Zum Schutze der Heranwachsenden wird es somit doch Aufgabe des Jugendmedienschutzes, entsprechende Angebote zu identifizieren und Kinder und Jugendliche vor ihnen zu schützen.

2.1.1.2 Relativ unzulässige Inhalte

Die Verbreitung relativ unzulässiger Inhalte (»einfache Pornografie«) über das WWW ist im Sinne des § 4 Absatz 2 JMStV zulässig, »wenn vonseiten des Anbieters sichergestellt ist, dass sie nur Erwachsenen zugänglich gemacht werden (geschlossene Benutzergruppe)«. Daraus resultieren zwei mögliche Anforderungen an technische Verfahren:

1. Relativ unzulässige Inhalte dürfen, so lange keine erfolgreiche Altersverifikation stattgefunden hat, im WWW nicht auffindbar sein. Da die Altersverifikation eine Personenidentifikation beinhaltet, muss also sicher gestellt werden, dass automatische Analyseverfahren nicht auf entsprechende Inhalte zugreifen können. Falls ein Verfahren dennoch entsprechende Inhalte auffinden kann, liegt ein entsprechender Verstoß vor.
2. »Die Authentifizierung dient der Sicherstellung, dass nur die jeweils identifizierte und altersgeprüfte Person Zugang zu geschlossenen Benutzergruppen erhält […].« [KJM, 2012f] Das Authentifizierungsverfahren für die Nutzung einer geschlossenen Benutzergruppe muss also sicher sein und darf auch nicht von technischen Verfahren überwunden werden können. Zur Überprüfung dieser Sicherheit können entsprechende Verfahren eingesetzt werden.[2]

2.1.1.3 Entwicklungsbeeinträchtigende Inhalte

Im Sinne des § 5 JMStV ist die Verbreitung von Inhalten, die geeignet sind, die Entwicklung von Kindern oder Jugendlichen bis zu einer bestimmten Altersstufe zu beeinträchtigen, gestattet, wenn der

[2] Diese Verfahren dienen jedoch nicht der Detektion oder Klassifikation von Web-Inhalten und werden daher in der vorliegenden Studie nicht weiter betrachtet.

Anbieter dafür Sorge trägt, dass Kinder und Jugendliche der betroffenen Altersgruppe sie »üblicherweise nicht wahrnehmen« (vgl. § 5 [JMStV, 2003]). Hieraus ergeben sich eine Reihe von Anforderungen an die Beurteilung von Web-Inhalten, die durch technische Verfahren nachgebildet werden könnten:

1. Technische Verfahren müssen entwicklungsbeeinträchtigende Inhalte detektieren. Dies kann nur an Hand eindeutiger und quantisierbarer Kriterien erfolgen. Für die im Rahmen der vorliegenden Studie betrachteten Inhalte-Kategorien erfolgt die Analyse und ausführliche Betrachtung entsprechender Kriterien in Kapitel 3 (Gewalt vgl. Abschnitt 3.2.2, Extremismus vgl. Abschnitt 3.3.2, Selbstgefährdung vgl. Abschnitt 3.4.2).
2. Die Eignung zur Entwicklungsbeeinträchtigung muss immer im Kontext einer bestimmten Altersstufe betrachtet werden. Daher müssen Inhalte im Sinne des § 5 JMStV immer altersdifferenziert betrachtet werden. Nähere Ausführungen hierzu finden sich ebenfalls in Kapitel 3.
3. Die Verbreitung entsprechender Inhalte ist zulässig, wenn der Anbieter der Sorgfaltspflicht genügt. Automatische technische Verfahren müssen demnach erkennen, ob der geforderten Sorgfaltsanforderung entsprochen ist, zum Beispiel durch
   (a) die Programmierung des Webangebots für ein anerkanntes Jugendschutzprogramm,[3]
   (b) eine geeignete Sendezeitbeschränkung,
   (c) die Trennung von expliziten Kinderangeboten, sofern eine Entwicklungsbeeinträchtigung nur für Kinder[4] anzunehmen ist (vgl. § 5 Absatz 5 JMStV) oder
   (d) ein anderes technisches oder sonstiges Mittel, das die Wahrnehmung des Angebots durch Kinder oder Jugendliche der betroffenen Altersstufe unmöglich macht oder wesentlich erschwert (§ 5 Absatz 3 JMStV).
4. Die Verbreitung entwicklungsbeeinträchtigender Inhalte in Nachrichtensendungen und vergleichbaren Angeboten in Telemedien ist unter bestimmten Voraussetzungen explizit zulässig (vgl. § 5 Abs 6 JMStV). Technische Verfahren müssen also differenzieren können, ob ein als entwicklungsbeeinträchtigend klassifizierter Inhalt im Kontext von Nachrichten angeboten wird.[5]

Die nähere Betrachtung der im JMStV definierten Kriterien zeigt somit auf, dass an die vollautomatische und rein technische Beurteilung der Zulässigkeit einer Verbreitung von Inhalten komplexe Anforderungen zu stellen sind. Der Fokus der vorliegenden Studie liegt nun auf der Detektion jugendschutzrelevanter Inhalte, welche im Folgenden näher untersucht wird. Abschnitt 2.1.2 betrachtet die Anforderungen an die altersdifferenzierte Klassifikation tiefergehend und vergleicht die in Deutschland Anwendung findenden Altersklassifikationen im internationalen Umfeld. Die Prüfung der Zulässigkeit der Verbreitung eines als jugendschutzrelevant detektierten Angebots ist eine damit verbundene aber unabhängig davon zu betrachtende Aufgabe und wird in der vorliegenden Studie nicht weiter untersucht.

2.1.1.4 Jugendschutz in der Werbung

Der JMStV sieht ausführliche Regelungen für den Jugendschutz hinsichtlich Werbung und Teleshopping vor (vgl. § 6 JMStV), die auch Anwendung im WWW finden. Die Überprüfung der Einhaltung der vorgesehenen Regeln ist ebenfalls Aufgabe des Jugendmedienschutzes. Es liegt jedoch nicht im

[3] Voraussetzung hierfür ist, dass zumindest ein Jugendschutzprogramm nach § 11 JMStV anerkannt ist.

[4] I. S. d. § 3 Abs 1 JMStV ist Kind, wer noch nicht 14 Jahre alt ist.

[5] In diesem Kontext sei erwähnt, dass – insbesondere anlässlich der versuchten Novellierung des JMStV im Jahr 2010 – eine rege Diskussion entstanden ist, in wieweit Internet-Blogs als »bei Telemedien vergleichbare Angebote zu Nachrichtensendungen oder Sendungen zum politischen Zeitgeschehen im Rundfunk« anzusehen sind [Assion, 2012].
Auch einzelne Hersteller von Computerspielen sehen diese als »legitime Ergänzung zur Berichterstattung im Fernsehen und in den Print-Medien« an, wie 2011 anlässlich dem Erscheinen des Kriegsspiels *The Death of Osama bin Laden* von KUMA, LLC. verlautbart wurde. In dem Spiel kann der Spieler die Geheimoperation *Geronimo* vom 1.5.2011 nachspielen [Gamestar, 2011, Kuma Games, 2011].

Fokus der vorliegenden Studie, technische Verfahren zur Überprüfung der Einhaltung dieser Regelungen zu untersuchen. Daher wird dieser Bereich von der vorliegenden Studie nur in soweit betrachtet, wie jugendschutzrelevante Inhalte Bestandteil von Werbung im WWW sind.

2.1.2 Altersdifferenzierte Klassifikation

Die altersdifferenzierte Klassifikation von Inhalten ist von zentraler Bedeutung für die Beurteilung von Web-Inhalten hinsichtlich ihrer Eignung zur Entwicklungsbeeinträchtigung von Kindern und Jugendlichen. Die gesetzliche Grundlage hierfür ist mit §§ 5 und 11 JMStV gelegt.

Die altersdifferenzierte Betrachtung von Inhalten und Angeboten ist eine sehr komplexe Aufgabe, bei der zahlreiche Kriterien berücksichtigt werden müssen, wie das folgende Beispiel exemplarisch für Gewaltdarstellungen beschreibt: »Die Beurteilung von Gewaltdarstellungen ist beispielsweise abhängig vom Gesamtkontext und von der Art der Einbettung der Darstellungen in das Gesamtangebot. Hierbei sind einzelne Gewaltdarstellungen in ihrer Ausgestaltung und Intensität zu berücksichtigen. Die zu prüfenden Faktoren lauten insbesondere: nachvollziehbarer und verständlicher Handlungsverlauf, Einsatz von adäquaten und nicht selbstzweckhaften Gewalthandlungen, Einbettung in den Gesamtkontext, Grundstimmung des Angebots, Formen der Gewalt (physisch, psychisch, verbal, strukturell), Relevanz (Bezug zur Lebenswelt von Kindern und Jugendlichen), Intensität (Maß an Brutalität, Ausgespieltheit, Detailfreude), Inszenierung (drastische und blutige Bilder, hohe Emotionalität, Hervorhebungen mittels Musik), Anteil der brutalen und exzessiven Gewaltszenen, Darstellung der Folgen der Gewalt, Einbeziehung in reale oder lebensnahe Kontexte, beispielsweise in Nachrichtenbeiträgen, Spannung (Ruhepausen, Happy End), Subjekte der Gewalt, Objekte der Gewalt, angebotsinterne Sanktionierung von Gewalt, Art der Inszenierung, dramaturgische Verläufe sowie formal-ästhetische Gestaltungsmittel.« [Hopf und Braml, 2010]

Der Beitrag zeigt die Bandbreite der bei der altersdifferenzierten Klassifizierung zu berücksichtigenden Faktoren auf. Für viele der Faktoren bestehen Ansätze für technische Verfahren, die belastende, aber auch entlastende Momente identifizieren können. Ob die zur Verfügung stehenden technischen Verfahren ausreichen, um automatisch eine Gesamtbewertung vornehmen zu können, muss im Einzelfall an jedem einzelnen Gefährdungsszenario betrachtet werden. Unabhängig von der Frage, ob eine Bewertung vollautomatisch vorgenommen werden kann oder ob nur einzelne be- oder entlastende Indizien automatisch erkannt werden können, ist für den Einsatz technischer Verfahren die Festlegung objektiver und quantisierbarer Kriterien für jede einzelne Altersstufe eine wesentliche Voraussetzung.

Kapitel 3 betrachtet tiefergehend bestehende Kriterienkataloge zur altersdifferenzierten Klassifikation von Inhalten, die bislang vor allem für die redaktionelle Bewertung von Inhalten herangezogen werden und stellt diese den für die vorliegende Studie zur Verfügung gestellten Testdatenbeständen gegenüber.

Die Wahl der Altersgruppen für die altersdifferenzierte Klassifizierung stellt ebenfalls Anforderungen an technische Verfahren. Je feingranularer Altersstufen gewählt werden, umso mehr Kriterien sind für die Klassifizierung festzulegen. Ein sehr grobes Raster an Altersstufen läuft hingegen den Informationsbedürfnissen der Älteren jeder Altersstufe zuwider, denn »die Beurteilung der Beeinträchtigung hat an den schwächeren und noch nicht so entwickelten Mitgliedern der Altersgruppe zu erfolgen« [JuSchRiL, 2005]. Im internationalen Vergleich finden sehr unterschiedliche Altersstufen Anwendung, zum Beispiel definieren Finnland und Österreich sechs Altersstufen, Belgien hingegen nur zwei (vgl. Tabelle 2.1). Im Folgenden werden die in verschiedenen Jugendschutzsystemen gewählten Altersstufen in Deutschland und im internationalen Vergleich näher betrachtet.



Abbildung 2.2
Altersfreigabekennzeichnung nach FSK

2.1.2.1 Altersstufen verschiedener Jugendschutzsysteme in Deutschland

In Deutschland werden Altersempfehlungen und Altersbeschränkungen von verschiedenen Einrichtungen und Stellen vorgenommen. Diese werden nachfolgend vorgestellt. Von den Einrichtungen der freiwilligen Selbstkontrolle in Deutschland wird medienübergreifend angelehnt an das JuSchG ein einheitliches Klassifikationsschema für Altersstufen angewandt. Generell werden Angebote nach folgendem Schema bewertet:

- Freigegeben ab 0 Jahren
- Freigegeben ab 6 Jahren
- Freigegeben ab 12 Jahren
- Freigegeben ab 16 Jahren
- Freigegeben ab 18 Jahren

Die FSK und die USK verwenden für die Auszeichnung der Altersstufen durch Etiketten ein einheitliches Farbsystem (vgl. Abbildung 2.2 und Abbildung 2.3).

Anzumerken ist in diesem Zusammenhang, dass gemäß § 3 JMStV eine abgrenzende Definition der Begriffe *Kind* und *Jugendlicher* vorgenommen wird. Kind ist demnach, »wer noch nicht 14 Jahre« alt ist und Jugendlicher, »wer 14 Jahre, aber noch nicht 18 Jahre alt ist« [JMStV, 2003]. Entgegen dieser Definition wenden die anerkannten Einrichtungen der Selbstkontrolle ein Klassifikationsschema an, dass eine Alterseinteilung »ab 12« beziehungsweise »ab 16« vorsieht. Eine Unterteilung in Altersgruppen »bis 13« und »ab 14« findet de facto nicht statt.

Die *Kriterien der KJM für die Anerkennung von Jugendschutzprogrammen im Bereich des World Wide Web* [KJM, 2010] fordern Konfigurationsmöglichkeiten für mindestens die Altersstufen »bis unter 12 Jahre«, »12 bis unter 16 Jahre« und »16 Jahre bis unter 18 Jahre« (vgl. Abschnitt 1.4.1).

Eine Besonderheit findet sich in der bundeseinheitlichen Regelung für Filmvorführungen (Kino). So dürfen unter 14-jährige Jugendliche Filme der Freigabe FSK 12 konsumieren, so lange sie *vor 20 Uhr* enden. Jugendliche im Alter von 14 oder älter dürfen dieselben Filme konsumieren, auch wenn sie *nach 20 Uhr* aber *vor 22 Uhr* enden (§ 11 Absatz 3 JuSchG [JuSchG, 2008]).

Freiwillige Selbstkontrolle der Filmwirtschaft (FSK) Die Hauptaufgabe der FSK besteht in der Prüfung von Filmen, DVDs und sonstiger Medienträger, die in Deutschland zur öffentlichen Vorführung vorgesehen sind, im Sinne des JuSchG. Mit der Abteilung FSK.online ist sie seit September 2011 eine anerkannte Einrichtung der Freiwilligen Selbstkontrolle für Webangebote im Sinne des JMStV [FSK, 2012d], wie in Abschnitt 1.2.2.3 bereits dargelegt wurde. Die Altersstufen der FSK richten sich nach § 14 JuSchG und sind die oben genannten.

Unterhaltungssoftware Selbstkontrolle (USK) Die USK ist in Deutschland die verantwortliche Stelle für die Altersfreigabe von Computerspielen nach dem JuSchG (vgl. Abschnitt 1.2.2.3). In diesem Sinne überprüft die USK Spiele, welche auf physischen Datenträgern vorgelegt werden. Seit September 2011 ist die USK neben dem Trägermedienbereich auch für die freiwillige Selbstkontrolle



Abbildung 2.3
Altersfreigabekennzeichnung nach USK

im Bereich der Online-Medien nach dem JMStV anerkannt. Die Altersstufen der USK richten sich nach § 14 JuSchG und sind die oben genannten.

Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e. V. (FSM) Die FSM war die erste für den Bereich der Online-Medien anerkannte Einrichtung der Freiwilligen Selbstkontrolle, wie bereits in Abschnitt 1.2.2.3 vorgestellt wurde. Die FSM bietet ein Altersklassifizierungssystem an, auf dem Anbieter mit Hilfe eines Fragebogens ihre Onlineinhalte gemäß den gesetzlichen Bestimmungen bewerten und einer Altersfreigabe zuordnen können. Die von der FSM vorgesehenen Altersstufen sind die oben genannten [FSM, 2011b]. Ein wesentlicher Unterschied zu der Begutachtung durch FSK und USK besteht in der unterschiedlichen rechtlichen Verbindlichkeit der Alterskennzeichnung. Die von FSK und USK im Rahmen eines – juristisch als Verwaltungsakt zu betrachtenden – Prüfvorgangs erteilten Alterskennzeichen sind rechtlich bindend, während die – eher als Gutachten zu betrachtenden – Altersklassifikationen der FSM nur für den Bereich der Medienaufsicht einen verbindlichen Charakter hat.

Freiwillige Selbstkontrolle Fernsehen (FSF) Die FSF ist eine anerkannte Selbstkontrolleinrichtung privater Fernsehanbieter in Deutschland, die Altersfreigaben für die Bereiche Fernsehen und WWW erteilt (vgl. hierzu Abschnitt 1.2.2.3). Die Altersstufen orientieren sich dabei an den oben genannten [FSF, 2012a]. Sie können im Sinne des § 5 JMStV entsprechenden Sendezeitbegrenzungen (vgl. Tabelle 1.1) zugeordnet werden. Davon machen einige Anbieter von Online-Filmportalen, insbesondere die öffentlich-rechtlichen Rundfunkanstalten Deutschlands, Gebrauch (vgl. Abschnitt 2.4.3.1).

Bundesprüfstelle für jugendgefährdende Medien (BPjM) Die BPjM ist zuständig für die Prüfung und Aufnahme jugendgefährdender Medien in die Liste jugendgefährdender Medien.[6] Sie dient dem medialen Jugendschutz. Nach § 18 Abs 1 JuSchG bedeutet jugendgefährdend, dass »die Entwicklung von Kindern oder Jugendlichen oder ihre Erziehung zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit« [JuSchG, 2008] in Gefahr ist. Beispielhaft werden Medien genannt, die »unsittlich sind, verrohend wirken, oder zu Gewalttätigkeit, Verbrechen oder Rassenhass anreizen« [ebd.]. Nach § 15 Absatz 2 JuSchG unterliegen bestimmte Medieninhalte wegen ihrer offensichtlichen Jugendgefährdung schon kraft Gesetzes einer Verbreitungsbeschränkung, ohne dass es einer Aufnahme in die Liste der jugendgefährdenden Medien bedürfte. Dazu zählen zum Beispiel die nach dem StGB verbotenen Inhalte wie Volksverhetzung, Anleitung zu Straftaten, Gewaltverherrlichung und -verharmlosung, Aufstachelung zum Rassenhass, Gewalt- und Kinderpornographie, Medien, die den Krieg verherrlichen oder Kinder oder Jugendliche in unnatürlicher, geschlechtsbetonter Körperhaltung darstellen. Die Indizierung jugendgefährdender Medien durch die BPjM erfolgt ohne weitere Altersdifferenzierung.

2.1.2.2 Altersklassifikation im internationalen Umfeld

Im internationalen Umfeld werden teilweise andere Altersstufen für die Kategorisierung von Angeboten verwendet. Auch finden je nach gesellschaftlichen Moral- und Wertvorstellungen unter-

[6] Dies wird gemeinhin als *Indizierung* bezeichnet.



Abbildung 2.4
Altersfreigabekennzeichnung nach PEGI



Abbildung 2.5
Inhaltekennzeichnung nach PEGI

schiedliche Bewertungsmaßstäbe Anwendung. Ergänzend zu den Altersauszeichnungen findet bei verschiedenen Systemen auch eine optische Kennzeichnung der Art des potentiell entwicklungsbeeinträchtigenden Inhalts mittels Icons statt. Somit kann von den Eltern auf einen Blick eine weitergehende Einschätzung des Angebots vorgenommen werden.

Nachfolgend werden einige internationale beziehungsweise nicht-deutsche Systeme zur Altersklassifikation betrachtet und in ihren spezifischen Merkmalen gegenüber gestellt.

Pan European Game Information (PEGI) Pan European Game Information (PEGI) ist das erste europaweite Alterseinstufungssystem für Computerspiele und wird von der *Interaktiven Softwareföderation Europas (ISFE)* verwaltet. Für die praktische Umsetzung des PEGI-Systems ist das niederländische *Institut für die Klassifizierung audiovisueller Medien (NICAM)* zuständig. PEGI wurde als einheitliches europäisches System entwickelt, um die Uneinheitlichkeit zwischen den Software-Einstufungssystemen verschiedener Länder zu beseitigen und den europaweiten Verkauf von Spielen zu erleichtern. Deutschland übernahm das Pan-European Game Information System (PEGI-System) nicht. Stattdessen benutzt man die rechtlich maßgebliche Alterskennzeichnung der USK. Dennoch ist die PEGI-Klassifizierung (siehe Abbildung 2.4) oft auf deutschsprachigen Produkten zu finden. Das PEGI-System besteht aus zwei Teilen: einer Altersempfehlung und einer Inhaltsbeurteilung. Auf der Spieleverpackung werden beide durch diverse Symbole angegeben [PEGI, 2012]. Die Altersempfehlung nach PEGI gibt an, ob es für Spieler einer bestimmten Altersklasse akzeptabel ist. Die Altersstufen sind:

- ab 3 Jahren (ab 4 Jahren in Portugal)
- ab 7 Jahren (ab 6 Jahren in Portugal)
- ab 12 Jahren
- ab 16 Jahren
- ab 18 Jahren

Die Beurteilung der Inhalte erfolgt in acht inhaltlichen Rubriken, die zum Teil eine Altersdifferenzierung vorsehen (siehe auch Abbildung 2.5):

1. Gewalt: Das Spiel enthält Gewaltdarstellungen oder verherrlicht/verharmlost Gewalt,
2. Sex: Spiel zeigt Nacktheit und/oder sexuelle Handlungen oder spielt auf sexuelle Handlungen an,
3. Drogen: Spiel bezieht sich auf Drogenkonsum oder zeigt diesen,
4. Angst: Spiel bereitet kleinen Kindern Angst oder ist gruselig,
5. Schimpfwörter: Spiel verwendet Schimpfwörter,
6. Diskriminierung: Spiel zeigt Diskriminierung oder Spielinhalt fördert Diskriminierung,



Abbildung 2.6
Altersfreigabekennzeichnung nach BBFC

7. Glücksspiel: Spiel fordert zum Glücksspiel auf oder gibt Anleitung dazu und
8. Online: Spiel kann online gespielt werden.

British Board of Film Classification (BBFC) Das British Board of Film Classification (BBFC) ist eine nicht-staatliche britische Organisation, die Filme, aber auch Computer- und Videospiele, bewertet und Altersfreigaben erteilt. Alle Filme auf Trägermedien (VHS, DVD, usw.) müssen vom BBFC geprüft werden, sollen sie in Großbritannien vertrieben werden. Ist ein solcher Film nicht vom BBFC freigegeben, so darf er nicht vertrieben werden. Ergänzend bietet BBFC.online die freiwillige Bewertung von Filmen, die per Internet zum Download oder Streaming angeboten werden. Zurzeit benutzt die BBFC folgende Freigaben (classifications, siehe auch Abbildung 2.6) [BBFC, 2009]:

- U: »Universal« – Für alle geeignet.
- PG: »Parental Guidance« – Für jede Altersgruppe erlaubt. Eltern seien aber darauf hingewiesen, dass manche Szenen für Kinder unter 7 Jahren nicht geeignet sind.
- 12A: »12 Accompanied/Advisory« – Nur für Personen über 12 Jahren. Mit Begleitung einer erwachsenen Person darf auch ein Kind unter dem 12. Lebensjahr sich den Film anschauen. Diese Freigabe gibt es nur bei Kinofilmen.
- 12: »12« – Freigegeben ab 12 Jahren. Anders als bei der 12A Freigabe muss diese Person über 12 Jahren sein, um den Film im Kino anschauen zu dürfen.
- 15: »15« – Freigegeben ab 15 Jahren.
- 18: »18« – Freigegeben ab 18 Jahren.
- R18: »Restricted 18« – Nur für Personen über 18 Jahre freigegeben und darf nur in Sexshops verkauft werden.

Entertainment Software Rating Board (ESRB) Das Entertainment Software Rating Board (ESRB) bewertet Computerspiele nach deren Eignung für Kinder und Jugendliche in den USA und Kanada, ähnlich der USK in Deutschland, jedoch mit dem Unterschied, dass es sich hierbei um eine reine Selbstregulationseinrichtung der Computerspieleindustrie ohne Rechtsverbindlichkeit handelt. Die Bewertung eines Spiels erfolgt auf freiwilliger Basis des Herstellers. Die ESRB verwendet die folgenden, jedoch nicht verbindlichen Altersempfehlungen [ESRB, 2012]:

- eC = early Childhood; ab 3 Jahren geeignet
- E = Everyone; ab 6 Jahren geeignet
- E10+ = Everyone 10 and older; ab 10 Jahren geeignet
- T = Teen; ab 13 Jahren geeignet
- M = Mature; ab 17 Jahren geeignet
- AO = Adults Only; ab 18 Jahren geeignet
- RP = Rating Pending; bedeutet, dass noch keine Altersempfehlung beschlossen wurde

Ergänzend zu den Altersempfehlungen verwendet das ESRB ein umfangreiches Vokabular an *Content Deskriptoren*, die den Käufern als eine weitere Orientierungshilfe für die bewerteten Spiele dienen sollen [ESRB, 2012]:

1. Alcohol Reference: Reference to and/or images of alcoholic beverages



Abbildung 2.7
Altersfreigabekennzeichnung nach ESRB

2. Animated Blood: Discolored and/or unrealistic depictions of blood
3. Blood: Depictions of blood
4. Blood and Gore: Depictions of blood or the mutilation of body parts
5. Cartoon Violence: Violent actions involving cartoon-like situations and characters. May include violence where a character is unharmed after the action has been inflicted
6. Comic Mischief: Depictions or dialogue involving slapstick or suggestive humor
7. Crude Humor: Depictions or dialogue involving vulgar antics, including »bathroom« humor
8. Drug Reference: Reference to and/or images of illegal drugs
9. Fantasy Violence: Violent actions of a fantasy nature, involving human or non-human characters in situations easily distinguishable from real life
10. Intense Violence: Graphic and realistic-looking depictions of physical conflict. May involve extreme and/or realistic blood, gore, weapons, and depictions of human injury and death
11. Language: Mild to moderate use of profanity
12. Lyrics: Mild references to profanity, sexuality, violence, alcohol, or drug use in music
13. Mature Humor: Depictions or dialogue involving »adult« humor, including sexual references
14. Nudity: Graphic or prolonged depictions of nudity
15. Partial Nudity: Brief and/or mild depictions of nudity
16. Real Gambling: Player can gamble, including betting or wagering real cash or currency
17. Sexual Content: Non-explicit depictions of sexual behavior, possibly including partial nudity
18. Sexual Themes: References to sex or sexuality
19. Sexual Violence: Depictions of rape or other violent sexual acts
20. Simulated Gambling: Player can gamble without betting or wagering real cash or currency
21. Strong Language: Explicit and/or frequent use of profanity
22. Strong Lyrics: Explicit and/or frequent references to profanity, sex, violence, alcohol, or drug use in music
23. Strong Sexual Content: Explicit and/or frequent depictions of sexual behavior, possibly including nudity
24. Suggestive Themes: Mild provocative references or materials
25. Tobacco Reference: Reference to and/or images of tobacco products
26. Use of Drugs: The consumption or use of illegal drugs
27. Use of Alcohol: The consumption of alcoholic beverages
28. Use of Tobacco: The consumption of tobacco products
29. Violence: Scenes involving aggressive conflict. May contain bloodless dismemberment
30. Violent References: References to violent acts

Motion Picture Association of America (MPAA) Die Motion Picture Association of America (MPAA) ist die Selbstorganisation der amerikanischen Filmproduzenten und -verleiher. Zu den Funktionen der Organisation zählt heute die Entscheidung, für welche Ziel- und Altersgruppen sich individuelle Filme eignen (Rating). Die Altersfreigaben/Empfehlungen (siehe auch Abbildung 2.8) der MPAA sind:[7]

[7] siehe http://www.mpaa.org/ratings/what-each-rating-means, Stand 20.11.2012

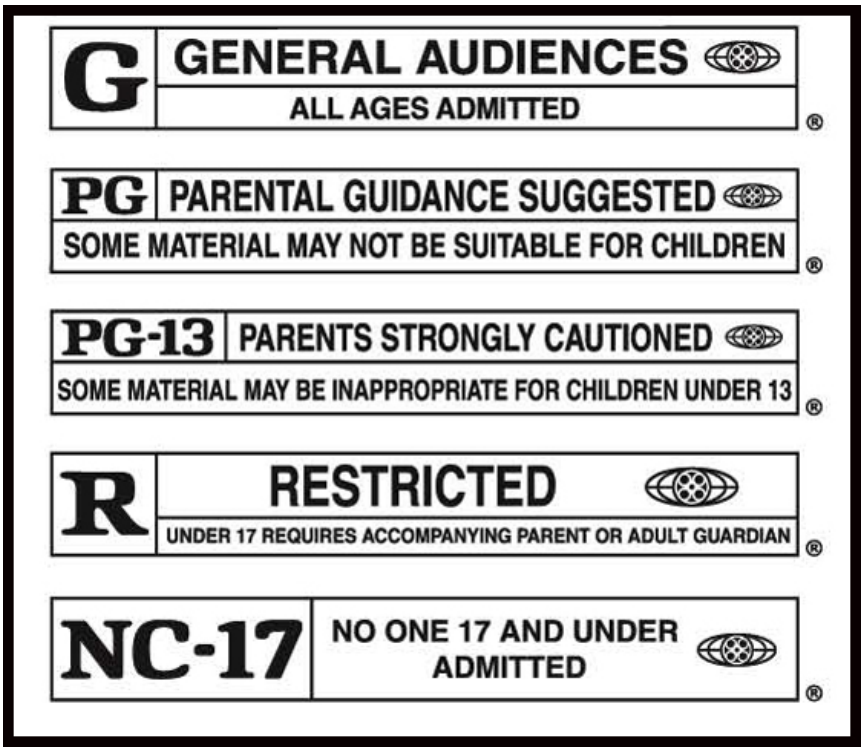


Abbildung 2.8
Altersfreigabekennzeichnung der MPAA

- G (general audiences): Entspricht dem deutschen »Freigegeben ohne Altersbeschränkung«. Der Film ist für Personen jeglichen Alters freigegeben.
- PG (parental guidance suggested): Diese Einstufung empfiehlt eine vorherige Begutachtung oder die Begleitung durch einen Elternteil bzw. einen Erwachsenen, jedoch ist es nur eine Empfehlung, was bedeutet, dass Kinder unter 13 Jahren dennoch frei den Film im Kino besuchen können.
- PG-13 (parents strongly cautioned): Diese Einstufung ist eine deutliche Verschärfung der PG-Freigabe. Es wird empfohlen, den Film erst Jugendlichen ab 13 Jahren zugänglich zu machen.
- R (restricted): Diese Einstufung erlaubt es Jugendlichen unter 17 Jahren nur in Begleitung eines Elternteils oder Erwachsenen, den Film zu besuchen.
- NC-17 (no one 17 and under admitted): Der Film ist ohne Ausnahme nicht zugänglich für Jugendliche, die 17 Jahre oder jünger sind. Also ist der Film nur für Personen ab 18 Jahren zugänglich.

Nicht von der MPAA geprüfte Filme oder Filmfassungen werden als »Unrated« bezeichnet. Die MPAA-Ratings unterscheiden sich insofern von den deutschen FSK-Freigaben, als mit Ausnahme der NC-17-Filme alle Filme grundsätzlich von Kindern und Jugendlichen jeglichen Alters gesehen werden dürfen. Es wird höchstens die Begleitung durch Erwachsene vorgeschrieben (R) oder empfohlen (PG, PG-13).

Computer Entertainment Rating Organization (CERO) Die Computer Entertainment Rating Organization (CERO) ist die japanische Organisation zur Bewertung von Computer- und Videospielen. Japanische Hersteller senden Videos und anderes Material über die zu bewertenden Spiele an die CERO, die anhand dessen die Spiele einstuft. Die Einstufung nach CERO ist zwar rein freiwillig, de facto hat sich die CERO-Einstufung jedoch zum Standard auf dem japanischen Markt entwickelt. Der Vertrieb von Produkten ohne CERO-Klassifikation wäre nur schwer möglich. CERO verwendet folgende Altersempfehlungen [IGN, 2012] (siehe auch Abbildung 2.9):



Abbildung 2.9
Altersfreigabekennzeichnung nach CERO



Abbildung 2.10
Inhaltekennzeichnung nach CERO

- CERO A: Ohne Alterseinschränkung
- CERO B: geeignet ab 12 Jahren
- CERO C: geeignet ab 15 Jahren
- CERO D: geeignet ab 17 Jahren
- CERO Z: geeignet ab 18 Jahren

Ähnlich dem europäischen PEGI-System verwendet auch CERO eine Klassifikation zur Kennzeichnung von Inhalten (siehe auch Abbildung 2.10):

- Romantik (Love)
- Gewalt (Violence)
- Alkohol-/Tabak-Genuss (Drinking/Smoking)
- Horror (Horror)
- Sexualität (Sexual content)
- Verbrechen (Crime)
- Drogen (Controlled substances (drugs))
- Glücksspiel (Gambling)
- Anstößige Sprache und Sonstiges (Language and others)

Office of Film and Literature Classification (OFLC) Die Office of Film and Literature Classification (OFLC) ist eine Behörde in Australien, die für die Altersfreigaben zuständig ist. Die OFLC klassifiziert Filme und Videospiele. Um in Australien vertrieben werden zu können, muss ein Medium gekennzeichnet sein (Kennzeichnung siehe Abbildung 2.11), ansonsten ist es gleichgestellt mit verbotenen Medien. Die OFLC verwendet zwei unterschiedliche Arten von Altersfreigaben [OFLC, 2012]. Die folgenden Freigaben haben lediglich beratenden Charakter:

- G (General) – bedeutet, dass das Medium für alle Altersgruppen freigegeben ist.
- PG (Parental guidence recommended) – bedeutet, dass bei jüngeren Kindern Begleitung durch die Eltern angebracht ist.



Abbildung 2.11
Altersfreigabekennzeichnung nach OFLC

- M (Recommended for mature audiences) – bedeutet, dass das Medium nur für ältere Personen empfohlen wird.

Die folgenden Freigaben sind verbindlich:

- MA15+ (Not suitable for people under 15. Under 15s must be accompanied by a parent or adult guardian) – bedeutet, dass Personen unter 15 Jahren von einem Erwachsenen begleitet sein müssen.
- R18+ (Restricted to 18 and over) – Ab 18 Jahren freigegeben.
- X18+ (Restricted to 18 and over) – Ab 18 Jahren freigegeben (betrifft nur Pornografie).
- Eine Freigabe kann auch verweigert werden, was bedeutet, dass das entsprechende Medium verboten ist. (Refused Classification)

Die Freigaben bei Videospielen entsprechen den Freigaben für Filmen, jedoch sind »R18+« und »X18+« nicht möglich.

Übersicht von Altersfreigaben im internationalen Vergleich Die vorgenannten nationalen und internationalen Systeme zur Alterseinstufung und -kennzeichnung von Inhalten zumeist für Filme oder Computerspiele weisen eine große Bandbreite an Granularität und Entscheidungskriterien auf. Kulturelle Unterschiede verschiedener Länder oder Regionen spiegeln sich in der Einschätzung der Eignung von Angeboten für Minderjährige wider. Im internationalen Vergleich erscheinen die in Deutschland Anwendung findenden Regelungen zur Kennzeichnung sehr übersichtlich, jedoch fehlt in dieser jegliche weitergehende Information über den Grund der Alterseinstufung beziehungsweise Hinweise auf den potentiell als kritisch zu betrachtenden Inhalt.

Tabelle 2.1 vergleicht in der Übersicht die staatlich vorgesehenen Altersstufen verschiedener Länder. Deutlich wird, dass in den meisten Ländern die Altersgrenzen 12, 16 und 18 Jahre für Altersfreigaben von Medien Anwendung finden, dass aber auch eine breite Streuung der als geeignet angesehenen Altersstufen sowie der erforderlichen Granularität (zwei bis sechs Altersstufen) vorherrscht. Bemerkenswert ist auch, dass die USA als ein für Deutschland und weite Teile der Welt bedeutsamer Inhalteproduzent deutlich von den deutschen und europäischen Standards abweicht. Demnach können die im Ursprungsland vorgenommenen Altersfreigaben allenfalls als Indiz zur eigenen nationalen Bewertung herangezogen werden.

2.1.2.3 Zusammenfassung: Anforderungen an technische Verfahren hinsichtlich der Altersdifferenzierung bei der Klassifikation von Medien

In Deutschland ist die Altersdifferenzierung auf Basis inhaltlicher Klassifikation eine wesentliche Anforderung an die Bewertung von Webangeboten. Daraus ist abzuleiten, dass Klassifizierungsverfahren, welche diese Unterscheidung unterstützen oder begünstigen, anderen Verfahren vorzuziehen sind.

Für die maschinelle Unterstützung der altersdifferenzierten Klassifikation ist es erforderlich, dass technische Verfahren die Eignung von Inhalten für die vorgegebenen Altersstufen beurteilen. National und international sind zahlreiche Organisationen für die Prüfung und Erteilung von Altersfreigaben zuständig. Jede Organisation verfügt über einen eigenen Katalog von Kriterien, der es den Prüfern ermöglicht, eine geeignete Altersfreigabe zu erteilen.[8] Die Kataloge und die für die Klassifikation vorgesehenen Altersstufen spiegeln dabei unterschiedliche gesellschaftliche Wert- und Moralvorstellungen wider und sind international nur begrenzt vergleichbar.

[8] Beispiele für solche Kataloge werden in Bezug auf die in der vorliegenden Studie betrachteten Gefährdungsszenarien in Kapitel 3 vorgestellt und näher betrachtet.

Tabelle 2.1
Altersfreigaben im internationalen Vergleich [FSK, 2011a].

Land	Stufen	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Deutschland	5	x						x						x				x		x
Belgien	2	x																x		
Bulgarien	3	x												x				x		
Dänemark	4	x							x				x				x			
Estland	5	x						x						x		x		x		
Finnland	6				x				x				x		x		x			x
Frankreich	4	x												x				x		x
Griechenland	3	x													x				x	
Großbritannien	4	x												x			x			x
Irland	5	x												x			x	x		x
Italien	3	x														x				x
Lettland	5	x							x					x				x		x
Malta	5	x												x		x		x		x
Niederlande	5	x						x			x			x				x		
Österreich	6	x						x				x		x		x		x		
Polen/Poland	4	x							x								x			x
Portugal	5				x			x						x				x		x
Rumänien	4	x												x			x			x
Schweden	4	x							x				x				x			
Spanien	5	x							x					x				x		x
Tschechien	4	x												x			x			x
Ungarn	4	x												x				x		x
Australien	3	x															x			x
Japan	4	x												x			x			x
Neuseeland	5	x													x		x	x		x
Singapur	3	x																x		x
USA	4	x													x				x	x

Für die in den nachfolgenden Kapiteln der vorliegenden Studie beschriebene Untersuchung technischer Verfahren wird in Anlehnung an die in Deutschland etablierten Jugendschutzsysteme eine Altersdifferenzierung entlang der folgenden Altersstufen betrachtet:

- unter 12: Kinder unter 12 Jahren
- 12–15: Jugendliche, die mindestens 12 aber noch nicht 16 Jahre alt sind
- 16–17 (unter 18): Jugendliche, die mindestens 16 aber noch nicht 18 Jahre alt sind

Soweit möglich und sinnvoll werden auch Erkenntnisse zur altersdifferenzierten Klassifikation von Inhalten aus dem internationalen Umfeld berücksichtigt. Spiegeln die Kataloge und die für die Klassifikation vorgesehenen Altersstufen zwar unterschiedliche gesellschaftliche Wert- und Moralvorstellungen wider und sind international nur begrenzt vergleichbar, so können sie dennoch wertvolle Anhaltspunkte für technisch überprüfbare Merkmale von Medien (Kriterien) liefern.

2.1.3 Besondere Aspekte bei der Klassifikation von Web-Inhalten

Bei der Bewertung von Webangeboten sind einige besondere Aspekte zu betrachten, die das Medium Internet auszeichnen und zu denen es keine Entsprechungen bei der Betrachtung von »herkömmlichen« Medien wie Trägermedien oder Rundfunk gibt. Dies sind insbesondere die Dynamik des Internets, die Möglichkeit der Aggregation unterschiedlicher Web-Inhalte in einem Webange-

bot, die Grenzenlosigkeit und die Beteiligung der Konsumenten an der Produktion von Web-Inhalten (UGC).

Erkennung von einzelnen Webseitenelementen Die Klassifikation einer einzelnen Webseite setzt sich in der Regel aus der Bewertung von einzelnen Elementen (z. B. eingebettete Videos) und des Gesamtkontextes zusammen. Beispielsweise führen Aufnahmen aus der Zeit des NS-Regimes im Kontext eines Angebots für historische Aufklärung gegenüber derselben Aufnahme in einem geschichtsrevisionistischen Kontext zu stark abweichenden Gesamtbewertungen. Klassifikationsverfahren sollten die Möglichkeit bieten, einzelne Webseitenelemente zu erkennen und getrennt zu bewerten. Die Gesamtbewertung einer Webseite sollte aus Sicht der Hersteller von Jugendschutzprogrammen automatisch, zum Beispiel auf Basis einer Aggregation von Einzelbewertungen oder durch Kontextanalyse getroffen werden. Eine manuelle Gesamtbewertung oder die Überprüfung der automatisch erfolgten Klassifikation kann von den Herstellern von Jugendschutzprogrammen nur in begrenztem Umfang geleistet werden.

Kontextanalyse Ähnlich zur Erkennung von Webseitenelementen können prinzipiell auch weitere Anhaltspunkte für die Gesamtbewertung einer Webseite herangezogen werden. Beispielsweise liefern Links auf und hin zu einer analysierten Seite oft Aufschluss über den inhaltlichen Kontext. Bevorzugt werden von den Herstellern Verfahren, die bereits solche oder vergleichbare Ansätze unterstützen.

Betriebsmodi für Bewertung einzelner Webseiten und ganzer Webangebote Die grundlegende Anforderung an die Klassifikationsverfahren ist eine Bewertung eines Webangebots unter einem vollständig qualifizierten Domainnamen (fully qualified domain name (FQDN)[9]). Dabei sind prinzipiell mehrere Betriebsmodi denkbar. Webangebote können zum Beispiel auf Basis der höchsten Altersdifferenzierungsstufe einer darunter geführten Webseite oder auf Basis einer gemittelten Bewertung aller verfügbaren Webseiten ausgewertet werden. Optimalerweise besteht die Möglichkeit der Einzelklassifikation von Webseiten, zum Beispiel für das Blockieren einzelner entwicklungsbeeinträchtigender YouTube-Videos, ohne YouTube gänzlich zu blockieren.

Berücksichtigung von eingebetteten Einzelinhalten Ähnlich der genannten Anforderung an Betriebsmodi müssen die Verfahren dazu geeignet sein, eine differenzierte Bewertung für eine Webseite in Abhängigkeit davon zu treffen, inwieweit eingebettete Inhalte von anderen Webangeboten (beispielsweise YouTube-Videos) Einfluss auf die Altersklasse haben und, im Umkehrschluss, welche Implikationen für die Quelle der eingebetteten Inhalte folgen.

Priorisierung von Einzelklassifizierungen Gerade nach der Integration von verschiedenen Klassifikationsverfahren offenbart sich die Frage, wie die Einzelbewertungen geeignet kombiniert werden können. Vergleichbar mit den Rangfunktionen moderner Suchmaschinen müssen die Verfahren quantitative Bewertungen ausstellen, die eine Kombination ermöglichen. Darüber hinaus muss entschieden werden, welche Gesamtentscheidung bei unterschiedlichen Klassifikationen erfolgen soll – zum Beispiel Verfahren A klassifiziert ein Webangebot als »ab 18 Jahre«, Verfahren B dasselbe Angebot als »ab 16 Jahre«. Dies stellt eine besondere Anforderung an das Gesamtklassifikationssystem dar.

Zuordnung zu thematischen Klassen Aufgrund der unerschöpflichen Menge von zu bewertenden Webangeboten ist es wünschenswert, dass Klassifikatoren Webangebote möglichst zuverlässig in vordefinierte thematische Klassen einordnen. Aus Erfahrung verschiedener Hersteller ergibt sich hier eine durchaus dreistellige Anzahl von Themenklassen. Auf Basis der Themenklassen können Webangebote akkumuliert in Altersgruppen eingeordnet werden. Diese Anforderung birgt aller-

[9]Der Begriff FQDN beschreibt die umgangssprachlich verwendeten Bezeichnungen *Domain* bzw. *Subdomain*.

dings die Gefahr von Falschbewertungen und nicht begründetes Anzeigen oder die nicht begründete Sperre eines Webangebots beim Nutzer.

Berücksichtigung von nutzergenerierten Inhalten im Web 2.0 Insbesondere in den Bereichen Web 2.0 und UGC ergeben sich für Jugendschutzprogramme besondere Herausforderungen bei der Klassifikation, aber auch Chancen bei der Ausnutzung bestehender Klassifikations- und Jugendschutzmechanismen. Beispiele für solche Angebote sind Videoportale wie YouTube oder Fotoportale wie flickr. Zum einen entsteht durch die zentrale Rolle des Nutzers als Inhalteproduzent und die weitestgehende Abwesenheit redaktioneller Kontrolle das Problem, dass sich bei einem Anbieter aus Sicht des Jugendschutzes vollkommen irrelevante Inhalte neben, im Extremfall, strafrechtlich relevanten Inhalten befinden. Zum anderen verfügen viele dieser Plattformen bereits über eigene Alarm- oder Registrierungssysteme, deren Ergebnisse ggf. technisch ermittelt und für die Altersklassifizierung berücksichtigt werden können.

Berücksichtigung dynamischer Inhalte Seit der Einführung des Web 2.0 gewinnen sogenannte *dynamische Webangebote* zunehmend an Relevanz. Im Vergleich zu statischen Webseiten, die genauso ausgeliefert werden, wie sie auf einem Webserver vorliegen, werden Inhalte von gängigen Web 2.0-Portalen abhängig von einer Vielzahl dynamischer Parameter in ihrer inhaltlichen Zusammensetzung beeinflusst. Ein klassisches Beispiel in Deutschland sind die vielen YouTube-Videos, die aufgrund von urheberrechtlichen Belangen für deutsche Nutzer gesperrt sind. Die entsprechenden Seiten werden dynamisch in Abhängigkeit von der Herkunft des Nutzers dargestellt. Darüber hinaus gibt es viele weitere Beispiele, wie etwa die nutzerabhängige Darstellung von Inhalten bei registrierten und eingeloggten Nutzern oder die Darstellung in Abhängigkeit vom genutzten Endgerät. Hier besteht daher eine Anforderung an ein Integrationskonzept, diese dynamischen Aspekte zu berücksichtigen und ggf. in die Bewertung einfließen zu lassen.

Inhalte auf Austauschplattformen Weitere Austauschkanäle für jugendgefährdende Inhalte neben den Angeboten und Diensten im WWW sind Peer-to-Peer Netzwerke[10] und andere Austauschplattformen wie zum Beispiel Rapidshare oder das mittlerweile geschlossene Megaupload. Da solche Plattformen auch für den Austausch von urheberrechtlich geschützten Materialien genutzt werden, besteht nicht nur aus Sicht des Jugendmedienschutzes ein Interesse an praktisch einsetzbaren Verfahren, die ungesetzliche Aktivitäten erkennen beziehungsweise verhindern.

Klassifizierung fremdsprachiger Inhalte Ein wesentliches Kernproblem der existierenden Jugendschutzprogramme ist der Umgang mit fremdsprachigen Inhalten und den damit einhergehenden fremden Schriftzeichen und Ähnlichem. Klassifikationsverfahren, die sprachunabhängig funktionieren, sind gemäß dieser Anforderung sprachabhängigen Verfahren vorzuziehen.

Hoher Automatisierungsgrad Eine naheliegende Kernanforderung ist der Automatisierungsgrad der Klassifikationsverfahren. Ziel der Hersteller ist es, Verfahren zu integrieren, die es ihnen ermöglichen, mehr Webangebote mit gleichbleibender Qualität und bei gleichbleibendem Personalaufwand zu bewerten. Teilautomatisierte Prozesse sind dabei nicht ausgeschlossen.

Transparenz der Klassifikation Automatische Klassifikationen müssen möglichst transparent für den Anwender sein. Das heißt, dass beim Blockieren einer Webseite detaillierte Informationen vorliegen müssen, aus welchem Grund diese Seite blockiert wird, insbesondere wenn es sich um eine Kombination verschiedener Teilbewertungen handelt. Darüber hinaus wäre es denkbar, dass zum

[10] Als Peer-to-Peer-Netzwerke (P2P) werden PC-Netze bezeichnet, die allen Rechnern im Netz eine gleichberechtigte Zusammenarbeit bieten. Das bedeutet, dass jedes System im Netz anderen Systemen Funktionen und Dienstleistungen anbieten und andererseits von anderen Systemen angebotene Funktionen, Ressourcen, Dienstleistungen und Dateien nutzen kann [ITwissen.info, 2012b].

Beispiel der Hersteller eines Jugendschutzprogramms diese Information für die Unterrichtung der Eltern zum Grund einer Blockierung nutzen könnte.

2.1.4 Zusammenfassung: funktionale Anforderungen an Klassifikationsverfahren

Die nähere Betrachtung der beiden funktionalen Grundanforderungen *Detektion von jugendschutzrelevanten Inhalten* und *altersdifferenzierte Klassifikation von jugendschutzrelevanten Inhalten* zeigt, dass je Inhaltekategorie und je betrachteter Altersgruppe quantisierbare Kriterien definiert sein müssen, damit automatische technische Verfahren überhaupt erst zum Einsatz kommen können. Das betrachtete Themenfeld ist dabei so breit, dass kein Verfahren sämtliche Unteranforderungen abdecken kann. Vielmehr muss für bestehende und zukünftige technische Verfahren festgelegt werden, welchen Detailanforderungen diese genügen sollen, damit ihre Wirksamkeit überprüft werden kann. Dabei ist den besonderen Eigenschaften der Angebote über das Medium Internet Rechnung zu tragen. So kann es gelingen, aus einer sinnvollen Kombination verschiedener Verfahren einen effektiven Schutzmechanismus zu entwickeln, der je nach aktueller Bedrohungslage (z. B. thematisch oder hinsichtlich verschiedener Medientypen) flexibel angepasst werden kann.

In Kapitel 3 werden bestehende Kriterienkataloge hinsichtlich ihrer Eignung für automatische technische Verfahren näher betrachtet und mit der der Studie zur Verfügung gestellten Testdatenmenge abgeglichen. In Abschnitt 4.2 wird überprüft, welche Kriterien für die technischen Verfahren tatsächlich nutzbar (quantisierbar) sind.

2.2 Nicht-funktionale Anforderungen an Klassifikationsverfahren

Im Gegensatz zu funktionalen Anforderungen, die definieren, *was* ein System oder Dienst leisten soll, und die je nach beschriebenem System oder Dienst sehr spezifisch und unterschiedlich ausfallen können, beschreiben nicht-funktionale Anforderungen (auch: nichtfunktionale Anforderungen), *wie gut* ein System oder Dienst seine Funktionen erfüllen soll. Sie werden auch als Attribute von Funktionen bezeichnet. Um konkrete nicht-funktionale Anforderungen zu spezifizieren, müssen die zunächst abstrakt formulierten Attribute (z. B. »das System muss ausreichend schnell sein«) in detaillierte (messbare) Qualitätsattribute verfeinert werden. Beispiel hierfür ist, wie schnell (genaue Zeitangabe in Sekunden) ein System auf Benutzereingaben reagieren soll. Die Messbarkeit der einzelnen Qualitätsattribute ermöglicht im Zuge von Systemtests (im Rahmen der Qualitätssicherung) eine einfache und schnelle Überprüfung der zugehörigen Anforderungen. Ziel nicht-funktionaler Anforderungen ist es, klare Qualitätsvorgaben an die systemunterstützten Geschäftsprozesse und Benutzeraufgaben zu erheben und schließlich umsetzen zu können [Fraunhofer-Gesellschaft, 2007].

Für nicht-funktionale Anforderungen existieren typische Gliederungen, beispielsweise nach Volere[11] oder der ISO/IEC 250xx-Normenreihe, die derzeit vom deutschen Institut für Normung zur Überführung in eine DIN-Norm vorbereitet wird.[12]

In diesem Abschnitt werden *nicht-funktionale Anforderungen an technische Verfahren zur Klassifikation von Web-Inhalten* unter verschiedenen Aspekten des praktischen Einsatzes für solche Verfahren betrachtet und entlang der folgender Struktur systematisiert, soweit diese hinsichtlich der betrachteten Verfahren relevant sind:

[11] siehe http://www.volere.co.uk/, Stand 30.11.2012

[12] siehe http://www.nia.din.de/projekte/DIN+ISO%2FIEC+25000/de/136443377.html, Stand 21.11.2012. Bis 2006 bestand die deutsche Norm DIN 66272 zur Festlegung von Qualitätsmerkmalen für Software.

1. Zuverlässigkeit
2. Korrektheit
3. Benutzbarkeit und Gebrauchstauglichkeit
4. Leistung und Effizienz
5. Betrieb und Umgebungsbedingungen
6. Wartbarkeit
7. Portierbarkeit
8. Flexibilität
9. Skalierbarkeit
10. Sicherheit
11. Randbedingungen

2.2.1 Zuverlässigkeit

Die Zuverlässigkeit eines Systems ist eine Eigenschaft, die angibt, wie verlässlich eine dem System zugewiesene Funktion in einem Zeitintervall erfüllt wird. Sie unterliegt einem stochastischen Prozess und kann qualitativ oder auch quantitativ beschrieben werden, sie ist nicht unmittelbar messbar. Hiervon sind die sogenannten deterministischen Eigenschaften (Merkmale) eines Produkts zu unterscheiden, welche direkt messbar sind.[13] Das Merkmal Zuverlässigkeit ist allen technischen Systemen inhärent, das heißt, dass kein technisches System frei von der Möglichkeit ist, auszufallen. Die Zuverlässigkeit von Software wird definiert als die Wahrscheinlichkeit einer fehlerfreien Software-Anwendung über eine spezifizierte Zeitdauer und unter spezifizierten Umgebungsbedingungen. Typische Merkmale für die Definition der Zuverlässigkeit von Software sind Systemreife, Wiederherstellbarkeit und Fehlertoleranz.

An die zu betrachtenden technischen Verfahren sind hohe Anforderungen an die Zuverlässigkeit zu stellen, da sie nur dann im Rahmen von Jugendschutzprogrammen oder Jugendschutzstrukturen, die ihrerseits qualitativen und quantitativen Mindestanforderungen genügen müssen (vgl. z. B. die Anforderungen zur Anerkennung von Jugendschutzprogrammen [KJM, 2011a]), eingesetzt werden können.

Dennoch können Verfahren, die sich seitens der Forschung noch in der Entwicklung oder Erprobung befinden, eine sinnvolle und wertvolle Ergänzung für den technischen Jugendmedienschutz darstellen und sollten nicht wegen mangelnder Systemreife oder Fehlertoleranz ignoriert werden. Beim Einsatz solcher Verfahren muss jedoch durch geeignete Mechanismen potentiellen Mängeln an der Zuverlässigkeit Rechnung getragen werden, so dass eine unzuverlässige Komponente nicht das gesamte Schutzsystem beeinträchtigen kann.

Fehlertoleranz gegenüber Anwendungsfehlern Bei der Fehlertoleranz gegenüber Anwendungsfehlern handelt es sich um eine grundsätzliche Anforderung an Softwaresysteme. Die Klassifikatoren müssen eine Toleranz gegenüber Fehlern der Anwender, in dem Fall diejenigen der Hersteller von Jugendschutzprogrammen, aufweisen und solche Fehler entweder kompensieren, geeignete Hinweise oder Alternativen anbieten oder korrigierte Neuklassifikationen ermöglichen.

[13] Allgemeine technische Eigenschaften können zum Beispiel sein: Gewicht, Abmessungen, Festigkeit, Farbe, elektrische und Wärmeleitfähigkeit. Beispiele für messbare Eigenschaften von Softwareprodukten können die Anzahl der Quellcodezeilen, Klassen oder Methoden sein.

2.2.2 Korrektheit

Unter dem Begriff Korrektheit wird die Eigenschaft eines Systems oder Dienstes verstanden, gemäß einem gegebenen Auftrag und den vorhandenen Rahmenbedingungen ein fehlerfreies Ergebnis zu erzeugen. Die Korrektheit beschreibt also die Wirksamkeit oder Effektivität technischer Verfahren. Automatische Verfahren sind aber in der Regel fehlerbehaftet. Insofern liegt hier der Fokus darauf, die einzusetzenden Verfahren auf die Höhe der Klassifikationsqualität in Abhängigkeit vom Aufwand für Integration, Wartung und Betrieb trainierbarer und justierbarer Verfahren hin zu untersuchen.

Im Rahmen des technischen Jugendmedienschutzes hängt die Korrektheit der Ergebnisse einer Klassifikation vor allem von der Verfügbarkeit maschinell verwendbarer Klassifikationskriterien ab. Kapitel 3 zeigt auf, dass die Definition eindeutiger Kriterien für die maschinelle Verarbeitung kaum möglich und der Einsatz vergleichender und lernender Verfahren (siehe auch Kapitel 4) sinnvoll ist. Diese Verfahren liefern verschiedene Indizien, die zusammengenommen eine Klassifikation ermöglichen. Die identifizierten Indizien treffen dabei mit einer gewissen Wahrscheinlichkeit auf den zu bewertenden Inhalt zu, so dass eine hundertprozentige Korrektheit im Sinne einer »ja oder nein« Entscheidung nicht gewährleistet werden kann. Aus diesem Grund soll die Korrektheit der Verfahren als Maß für die Wahrscheinlichkeit einer richtigen automatischen Klassifikation verstanden werden. Dieses Maß kann zum Beispiel an Hand einer Testdatenmenge bestimmt werden, wie es in den Kriterien für die Anerkennung von Jugendschutzprogrammen durch die KJM definiert ist: »Die KJM orientiert sich dabei – angelehnt an das sog. Paretoprinzip – an der 80:20-Regel. Demnach ist die geforderte Zuverlässigkeit derzeit gegeben, wenn das Programm mindestens vier von fünf Angeboten richtig behandelt« [KJM, 2011a]. Das Maß berechnet sich dabei aus dem Verhältnis der Fehler von Overblocking zu Underblocking. Des Weiteren werden die für Kindersuchmaschinen wie zum Beispiel *fragFINN* und *Blinde Kuh* redaktionell ausgewählten Webangebote (Whitelists) als Testdatenmenge verwendet, deren Inhalte im Kontext der Jugendschutzprogramme nicht als problematisch eingestuft werden sollen (false positives).

2.2.3 Benutzbarkeit und Gebrauchstauglichkeit

Die Benutzbarkeit eines (Software-)Systems oder Dienstes beschreibt, welcher Aufwand der Einsatz der Software von den Benutzern fordert und wie er von diesen beurteilt wird. Sie beschreibt das Verhältnis von Effizienz und Effektivität.

Gebrauchstauglichkeit[14] bezeichnet das Ausmaß, in dem ein (Software-)System oder Dienst durch bestimmte Benutzer in einem bestimmten Anwendungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen.[15]

Die zu betrachtenden technischen Verfahren werden nur mittelbar durch Benutzer bedient werden, da sie für den Einsatz im technischen Jugendmedienschutz in geeignete Programme oder Systemumgebungen integriert werden müssen. Anforderungen an Benutzbarkeit und Gebrauchstauglichkeit sind vor allem an die Integration (vgl. Kapitel 6) zu stellen.

[14] engl.: Usability, vgl. EN ISO 9241 Teil 11

[15] Die Begriffe Gebrauchstauglichkeit und Benutzerfreundlichkeit sind eng miteinander verwandt. Die Gebrauchstauglichkeit bildet dabei die Mindestanforderung ab, dass ein System zur Zufriedenstellung der Nutzer arbeitet (Gebrauchstauglichkeit = Benutzbarkeit + Zufriedenstellung), wohingegen die Benutzerfreundlichkeit auch die emotionalen Aspekte des gesamten Nutzungserlebnisses, der sogenannten User Experience berücksichtigt.

Des Weiteren müssen technische Verfahren hinsichtlich ihrer Funktionsweise und ihrer Konfigurationsmöglichkeiten ausführlich und für die anvisierte Zielgruppe leicht verständlich dokumentiert sein. Dies stellt jedoch im engeren Sinne keine Anforderung an das technische Verfahren selbst dar.

2.2.4 Leistung und Effizienz

Leistung (im Sinne der IT: Rechenleistung) ist ein Maß für informationstechnische Systeme. Es gibt diverse Maßeinheiten, in denen die Rechenleistung angegeben wird. Viele Maßeinheiten für Prozessoren beziehen sich auf die Anzahl an Rechenoperationen in einer bestimmten Zeit. Allgemein werden auch komplexere Operationen – sogenannte Benchmarks – betrachtet, die unter vorgegebenen Rahmenbedingungen eine vergleichende Bewertung der Rechenleistung von Rechnern ermöglichen [ITwissen.info, 2012d].

Typische Leistungskriterien oder Kenngrößen für die Bewertung der Leistung technischer Verfahren zur Klassifikation von Web-Inhalten können zum Beispiel die Anzahl der untersuchten Webseiten pro Zeiteinheit oder die Geschwindigkeit der Analyse von Inhalten mit einer für den menschlichen Konsum vorgesehenen Abspiellänge (Audio, Video) sein. Hieraus leitet sich die Effizienz eines Verfahrens ab, die die Antwortzeiten, also die Zeitspanne zwischen dem Aufruf eines Verfahrens und dessen Antwort im Sinne einer Klassifikation, im Verhältnis zu den eingesetzten Ressourcen (Rechenleistung) und somit die Wirtschaftlichkeit betrachtet.

Die Effizienz der Analyseverfahren ist ein ausschlaggebender Punkt für deren Eignung im späteren Betrieb (vgl. Abschnitt 2.2.5). Da das Angebot im WWW praktisch unerschöpflich ist und stetig anwächst, müssen Verfahren ausgewählt werden, die ein Angebot in angemessener Zeit analysieren und klassifizieren können. Besonders interessant sind außerdem Verfahren, die auf dem End- oder Netzwerkgerät des Nutzers eine Echtzeitanalyse[16] der soeben aufgerufenen Seiten ermöglichen können.

2.2.5 Betrieb und Umgebungsbedingungen

Anforderungen an Betrieb und Umgebungsbedingungen für technische Verfahren zur Klassifikation von Web-Inhalten stellen den umfangreichsten Anteil der nicht-funktionalen Anforderungen dar, denn die Verfahren müssen in Jugendschutzprogramme (vgl. Abschnitt 2.3) oder Schutzsysteme (vgl. Abschnitt 2.4) integriert werden, damit sie zum technischen Jugendmedienschutz beitragen. Alleinstehend können Verfahren lediglich evaluiert und erprobt werden, für den Betrieb bedarf es der geeigneten Integration. Mit den Aspekten der Integration der in der vorliegenden Studie untersuchten Verfahren (vgl. Kapitel 5) beschäftigt sich das Kapitel 6.

Betriebliche Anforderungen stehen in engem Zusammenhang mit Anforderungen an die Wartbarkeit (vgl. Abschnitt 2.2.6), Portierbarkeit (vgl. Abschnitt 2.2.7) und Sicherheit (vgl. Abschnitt 2.2.10).

Nachfolgend werden die wesentlichen betrieblichen Anforderungen aus verschiedenen praktischen Betrachtungswinkeln heraus beleuchtet.

[16] Mit Echtzeitfähigkeit ist hier gemeint, dass der Einsatz von Verfahren zum Beispiel zum Zeitpunkt des Aufrufs einer Webseite eine Verzögerung (Latenz) in der User Experience des Benutzers verursacht, die einen bestimmten Schwellenwert nicht übersteigen darf. Die Wahl des Schwellenwerts (akzeptable Latenzzeit) hängt von verschiedenen Randbedingungen ab, zum Beispiel Nutzungsszenario, allgemeine (serverseitige) Geschwindigkeit des Angebots, Medientyp u. a. m.

2.2.5.1 Plattformen

Eine wichtige Fragestellung im Kontext von Jugendschutzsoftware ist die Plattform, auf der diese zum Einsatz kommen sollen. Abschnitt 1.4 hat bereits aufgezeigt, dass verschiedene Ansätze existieren, wie Jugendschutzmechanismen realisiert werden können. Technische Verfahren müssen demnach für den Einsatz auf der durch Jugendschutzsoftware oder -systeme adressierten Plattform geeignet sein.

PC Betriebssysteme Die zum heutigen Tage anerkannten Jugendschutzprogramme sind für den Einsatz auf einem typischen heimischen PC mit Microsoft Windows Betriebssystem vorgesehen. Damit wird der Verbreitung des Betriebssystems Rechnung getragen, denn auf über 90 % der PCs und Notebooks kommt Microsoft Windows zum Einsatz.[17] Die Programme integrieren sich in das Betriebssystem und bieten zum Teil komfortable Konfigurationsmöglichkeiten zum Beispiel über die Integration der Schutzfunktionalität in das Benutzermanagement des Betriebssystems. Auch für die PC-Betriebssysteme Linux und Mac OS existieren am Markt Schutzsysteme (vgl. [Cybion et al., 2012]), jedoch werden diese Betriebssysteme bislang nicht von in Deutschland anerkannten Jugendschutzprogrammen abgedeckt.

Sofern technische Verfahren innerhalb von Jugendschutzprogrammen verwendet werden, die auf dem heimischen PC zum Einsatz kommen sollen, ist die Kompatibilität mit den gängigen oben genannten Betriebssystemen eine wesentliche Anforderung an die technischen Verfahren. Sind technische Verfahren für den Einsatz auf Serverumgebungen vorgesehen, sind entsprechende Anforderungen an die dort verwendeten Betriebssysteme zu formulieren.

Smart Devices Als Smart Devices werden informationstechnisch aufgerüstete Alltagsgegenstände bezeichnet, die einen Mehrwert durch sensorgestützte Informationsverarbeitung und Kommunikation erhalten und im Sinne der vorliegenden Studie über einen Internetzugang verfügen [Lackes und Siepermann, 2012b]. Eine Kategorisierung von Smart Devices kann zum Beispiel nach ihrem ursprünglichen Nutzungszweck oder an Hand des Nutzungskontextes (z. B. stationär oder mobil) vorgenommen werden.

Typische Beispiele für Smart Devices sind Mobiltelefone, Tablet PCs, Spielekonsolen, Fernsehgeräte oder Blu-ray Player, mittels derer der Anwender einen beschränkten oder unbeschränkten Zugang zum Internet und dem WWW erhalten kann. Ein beschränkter Zugang zeichnet sich dadurch aus, dass zum Beispiel nur Inhalte aus einem Portal des Geräteherstellers konsumiert werden können oder der Online-Zugang über spezielle Apps erfolgt, die ihrerseits eine begrenzte Menge an Inhalten zugreifbar machen. Ein unbeschränkter Zugang zum WWW ist gewährt, wenn über einen Web-Browser sämtliche Web-Inhalte abrufbar sind.[18] Die technologische Entwicklung zeigt auf, dass zukünftig immer mehr Endgeräte mit einem Internetzugang ausgestattet und im Lebensraum von Kindern und Jugendlichen anzutreffen sein werden, zum Beispiel Internetradio-Empfänger, Navigations- und Entertainmentsysteme in Fahrzeugen, Foto- und Filmkameras[19] oder gar die auf den großen Unterhaltungselekronikmessen der vergangenen Jahre vorgestellten Kühlschränke mit Internetzugang.

Im Jahr 2012 nutzen bereits 46 % der 14- bis 19-Jährigen in Deutschland den Zugang zum Internet über ein mobiles Gerät, wie die ARD/ZDF-Onlinestudien belegen [Projecter, 2012]. Daher gewinnt der Jugendmedienschutz für die mobile Nutzung des Internets zunehmend an Bedeutung und sollte dem

[17] Dabei entfallen 52 % auf Windows 7, 27 % auf Windows XP, 7 % auf Windows Vista. Zum Vergleich: Die Verbreitung von Mac OS X beträgt etwa 7 % (Daten: Stand September 2012 [Statista, 2012a]).

[18] Ein unbeschränkter Zugang ist im Sinne der vorliegenden Studie auch dann gewährt, wenn spezielle Inhaltstypen (z. B. Flash oder spezielle Video- oder Audiocodecs) nicht wiedergegeben werden können.

[19] So können beispielsweise Aufnahmen direkt im Internet veröffentlicht werden oder soziale Netzwerke genutzt werden, Beispiel: Samsung ST80 [Netzwelt, 2012].

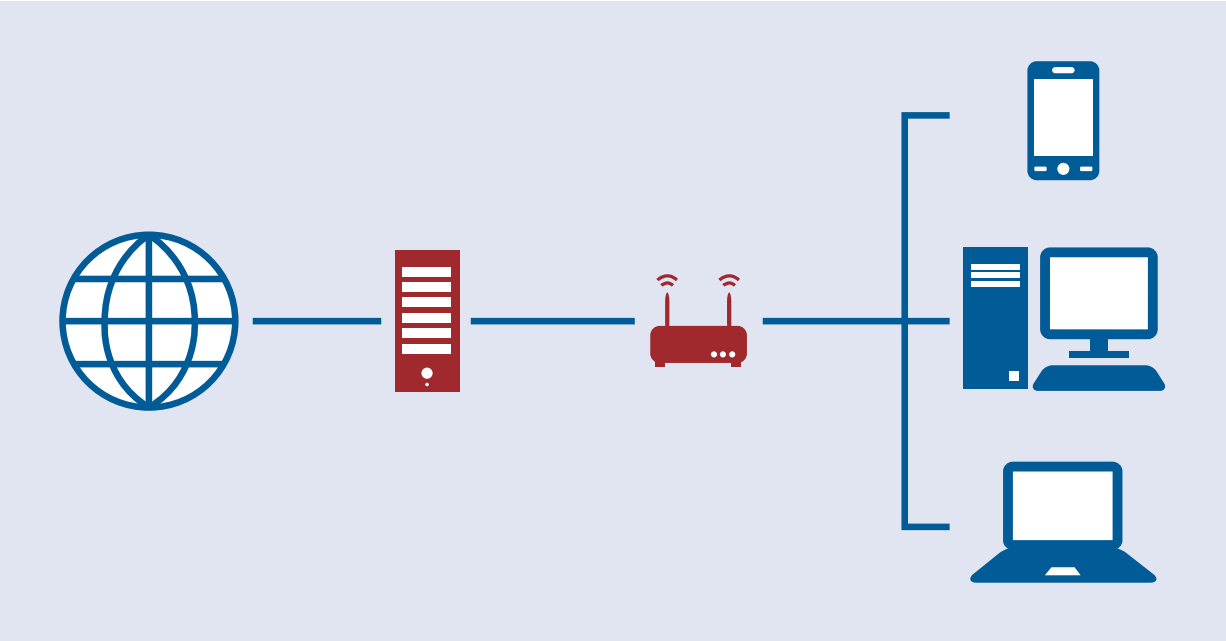


Abbildung 2.12
Jugendschutzmechanismen in Infrastrukturkomponenten

Schutz des heimischen PCs gleichgestellt sein. Verfahren für den technischen Jugendmedienschutz sollten demnach wenn möglich ebenfalls für den Schutz des mobilen Internetzugangs geeignet sein. Die daraus resultierenden Anforderungen an die Übertragbarkeit von technischen Verfahren in einen anderen Anwendungskontext werden in Abschnitt 2.2.7 näher betrachtet.

Infrastrukturkomponenten Jugendschutzmechanismen können auch außerhalb des vom Nutzer für den Internetzugang verwendeten Endgerätes stattfinden, also im Bereich des Internetzugangs (vgl. Abbildung 2.12). Hier bieten sich verschiedene Infrastrukturkomponenten für den Einsatz von Schutzsoftware an, zum Beispiel ein Router, der den Übergabepunkt im Heimnetzwerk realisiert[20] oder ein Zugangs-Proxy, der von einem unabhängigen Dienstleister[21] oder im Rechenzentrum des ISP betrieben werden kann.

Abhängig von der gewählten Lösung stehen den Jugendschutztechnologien unterschiedliche Hardware-Ressourcen zur Verfügung. Ein Router verfügt über vergleichsweise geringe Ressourcen, während eine Proxy-Lösung im Rechenzentrum den Anforderungen von Softwarekomponenten entsprechend angepasst werden kann. Sollen Verfahren auf Geräten mit sehr begrenzten Ressourcen zum Einsatz kommen, müssen sie insbesondere wegen der begrenzten Hardware-Ressourcen entsprechenden Performance-Anforderungen genügen. Die im vorhergehenden Abschnitt an Smart Devices formulierten Anforderungen gelten entsprechend (vgl. auch Abschnitt 2.2.7). Bestehende Verfahren lassen sich demnach auch danach kategorisieren, ob sie für den Einsatz auf Geräten mit stark eingeschränkten Ressourcen zum Einsatz kommen können oder nur in vergleichsweise Performanzmächtiger Infrastruktur Verwendung finden können.

[20] Ein Beispiel für ein bereits existierende derartiges Produkt ist der »Jugendschutzrouter« von Surf-Sitter (siehe http://surf-sitter.de/Produkte/PlugAndPlay, Stand 15.10.2012).

[21] Ein Beispiel für eine derartige Lösung ist das Projekt »Jugendschutzfilter für Schulen« (siehe http://www.jugendschutzfilter.net/, Stand 15.10.2012).

2.2.5.2 Einsatzszenario für das Verfahren

Wie bereits betrachtet, stellen die Hardware-Komponenten, die für die Ausführung von Schutzsoftware Anwendung finden sollen, wesentliche Anforderungen an die Performanz der eingesetzten technischen Verfahren. Grundsätzlich sollte dabei unterschieden werden, ob Verfahren während des Zugriffs auf einen Web-Inhalt zum Einsatz kommen sollen oder ob Verfahren zeitlich unabhängig von Nutzerzugriffen auf Web-Inhalte eingesetzt werden. In ersterem Fall sind dabei Anforderungen an die Echtzeitfähigkeit der Verfahren zu richten, die sich zusätzlich danach orientieren, ob die Verfahren auf dem vom Benutzer verwendeten Endgerät oder serverseitig Anwendung finden.

Echtzeitfähige Verfahren Echtzeitfähige Verfahren können zum Beispiel in Jugendschutzprogrammen auf dem Endgerät, als Bestandteil eines Betriebssystems oder einer Infrastrukturkomponente bei jedem Aufruf eines Web-Inhalts durch den Nutzer zur Anwendung kommen. Ihre Aufgabe ist dann die Prüfung des abgerufenen Inhalts und im Falle einer Detektion problematischer Inhalte die Verhinderung der Anzeige des entsprechenden Inhalts. Aus Nutzersicht ist dabei wesentlich, dass die durch die Überprüfung erzeugte Verzögerung (Latenzzeit) in einem akzeptablen Rahmen liegt.[22]

Asynchrone Verfahren Werden Verfahren zeitlich unabhängig (asynchron) vom Zugriff auf Web-Inhalte durch einen Nutzer eingesetzt, bestehen zunächst keine Anforderungen an die Echtzeitfähigkeit der Verfahren. Diese finden zum Beispiel für die Erstellung von Blacklists Anwendung, bei der Webseiten geprüft und im Verdachtsfalle entsprechend in einer Liste zur altersdifferenzierten Sperrung oder zur manuellen Nachkontrolle durch eine Redaktion vermerkt werden. Die Auswahl der zu überprüfenden Webseiten liegt in der Hand des jeweiligen Listen-Betreibers und kann zum Beispiel aufgrund von Beschwerden durch Nutzer, dem Nutzungsverhalten von Kindern und Jugendlichen[23] oder an Hand anderer Kriterien erfolgen. Dennoch sind je nach Anwendungsfall Anforderungen an die Performanz der eingesetzten Verfahren zu stellen. Als typische Kenngrößen können die Verarbeitungsdauer einer Webseite, die Verarbeitungsgeschwindigkeit von Multimedia-Inhalten[24] oder die Zeitdauer für den Vergleich eines Mediums mit einem Index[25] betrachtet werden.

Internet-basierte Kommunikation Aus der Praxis ergibt sich für den Jugendmedienschutz auch ein Schutzbedürfnis für Internet-basierte Kommunikationsmechanismen wie Chat-Räume oder Dialogfunktionen innerhalb von Online-Spielen. Neben dem Austausch entwicklungsbeeinträchtigender Inhalte (z. B. Selbstgefährdung) ist auch der Schutz vor Belästigungen und Cybermobbing von Bedeutung.[26] Sollen Verfahren für den Schutz von Internet-basierter Kommunikation zum Einsatz kommen, sind ganz erhebliche Anforderungen an die Performanz der Verfahren zu stellen. Die Untersuchung von technischen Verfahren für den Einsatz im Kontext Internet-basierter Kommunikation ist nicht Bestandteil der vorliegenden Studie, dennoch können Erkenntnisse der vorliegenden Studie für diesen Anwendungsfall dienlich sein.

[22] Welche Zeitspannen vom Nutzer als »akzeptabel« empfunden werden, hängt vom Nutzungskontext und dem zu konsumierenden Medientyp ab. Verfügt der Nutzer beispielsweise über einen sehr schnellen Internetzugang, bei dem der Aufruf von Webseiten ohne merkliche Verzögerung erfolgt, liegt die Akzeptanzschwelle deutlich geringer als bei einer mobilen und typischerweise langsameren Internetverbindung. Beim Abruf von Videomaterial ist ein Nutzer »gewöhnt«, dass es einen Moment lang dauert, bis das Video betrachtet werden kann.

[23] Eine derartige Möglichkeit bietet zum Beispiel das anerkannte JusProg Jugendschutzprogramm (vgl. Abschnitt 1.4.1.1).

[24] Als Verarbeitungsgeschwindigkeit wird das Verhältnis der Abspieldauer eines Mediums zur Verarbeitungsdauer durch technische Verfahren bezeichnet.

[25] Dies ist zum Beispiel relevant für den Einsatz von Fingerprinting-Technologien (vgl. Abschnitt 4.5.1).

[26] siehe hierzu auch http://www.chatten-ohne-risiko.net/erwachsene/risiken/, Stand 15.10.2012

2.2.5.3 Integration von Klassifikationsverfahren

Eine naheliegende und zentrale Anforderung an die Verfahren ist eine möglichst einfache und kostengünstige Integration in die Programme beziehungsweise die technische Infrastruktur der Hersteller. In diesen Bereich fallen mehrere Teilanforderungen wie zum Beispiel die Verfügbarkeit und Güte von geeigneten Softwareschnittstellen, die Unterstützung von aktuellen Betriebssystemen und Hardware,[27] die ausgetauschten Datenformate oder geeignete Schutzmechanismen bei einer netzwerkorientierten Integration.

Integration von Klassifikationsverfahren in Jugendschutzprogramme Anforderungen von Jugendschutzprogrammen an technische Verfahren zur Klassifikation von Web-Inhalten (siehe Schnittstelle 1 in Abbildung 2.1) betreffen vor allem die Gesamt-Leistungsfähigkeit der Jugendschutzprogramme, die von der Leistungsfähigkeit und Zuverlässigkeit der eingesetzten Verfahren abhängig ist. In Abschnitt 2.3 werden in einem Exkurs die Anforderungen an Jugendschutzsoftware diskutiert und daraus Anforderungen an Klassifikationsverfahren abgeleitet. Als wesentliche Anforderungen ergeben sich hierbei technische Aspekte der Integrierbarkeit, Performanz, Effektivität und Zuverlässigkeit. Darüber hinaus sind organisatorisch/wirtschaftliche Fragestellungen der Lizenzierbarkeit und Haftung von Bedeutung, die der ausgewählten Systematik folgend in Abschnitt 2.2.11 näher betrachtet werden.

Integration von Klassifikationsverfahren in technische Jugendschutzstrukturen Die Anforderungen an technische Verfahren zur Klassifikation von Web-Inhalten aus Sicht technischer Jugendschutzstrukturen (siehe Schnittstelle 2 in Abbildung 2.1) resultieren aus der Betrachtung der Fragestellung, wie automatische Verfahren beispielsweise in das bestehende Meldewesen von Beschwerdestellen integriert werden oder zur Generierung zentral bereit gestellter Blacklists beitragen können. Die Anforderungen sind dabei häufig denen der Jugendschutzprogramme (vgl. Abschnitt 2.2.5.3) ähnlich.

In einem Exkurs werden in Abschnitt 2.4 die Anforderungen an technische Jugendschutzstrukturen diskutiert und die daraus resultierenden Anforderungen an Klassifikationsverfahren zusammengefasst (vgl. Abschnitt 2.4.5). Ähnlich wie bei den Jugendschutzprogrammen stehen auch Fragen der Lizenzierbarkeit im Fokus der Betrachtung (vgl. Abschnitt 2.2.11).

2.2.6 Wartbarkeit

In der Softwaretechnik bezeichnet der Begriff *Softwarewartung* Tätigkeiten, »die nach der Einführung eines Softwaresystems dessen Verfügbarkeit und Nutzbarkeit sicherstellen oder die Lebensdauer der Software verlängern. Neben der Fehlerbehebung gehören die Funktionserhaltung und die Funktionserweiterung ebenso dazu wie perfektionierende Maßnahmen.« [Lehner, 2012]

Dabei wird zwischen folgenden Arten der Wartung unterschieden:

- korrektive Wartung: die Beseitigung von Fehlern
- perfektionierende Wartung: Verbesserung von Attributen wie etwa der Performanz oder der Wartbarkeit. Darunter fällt insbesondere die Bereinigung des Entwurfs oder der Implementierung durch Reengineering, Refactoring usw.

[27] Hier geht die Bewertung der beiden befragten Hersteller auseinander. Während ein Hersteller eher *handelsübliche* Server fordert, so erweitert der zweite Hersteller die Anforderung auf eine möglichst gänzliche Plattform- und Hardwareunabhängigkeit.

- adaptive Wartung: Anpassung der Software an veränderte technische Bedingungen der Umgebung

In diesem Sinne schließt die Wartbarkeit auch die Analysierbarkeit, Prüfbarkeit, Erweiterbarkeit und Änderbarkeit von Software ein.

Bei der Formulierung von Anforderungen an die Wartbarkeit technischer Verfahren muss grundsätzlich unterschieden werden, ob der Quellcode der Verfahren demjenigen, der die Verfahren einsetzt, zur Verfügung steht oder ob es sich um eine lizenzierte ausführbare Software handelt, bei der der Quellcode nicht offen gelegt wird (proprietäre Software). Ersteres ist vor allem gegeben bei Software, die unter einer Open Source-Lizenz genutzt werden kann oder bei Software, die in einem Auftrag unter der Maßgabe entwickelt wurde, dass der Quellcode dem Auftraggeber übergeben wird. Hierbei ist es für den nachhaltigen und zukunftssicheren Betrieb der Software unabdingbar, dass die oben genannten Kriterien hinsichtlich der Wartbarkeit eingehalten werden. Bei der Verwendung proprietärer Software sind sinngemäße Anforderungen an den Lizenzgeber der Software zu richten und Rahmenbedingungen zu schaffen, dass auch im Falle einer Beendigung der Unternehmertätigkeit des Lizenzgebers die Wartbarkeit der lizenzierten Software sichergestellt ist (z. B. durch den Abschluss eines geeigneten Escrow-Vertrags[28]).

2.2.7 Portierbarkeit

Portierbarkeit oder Portabilität beschreibt die Übertragbarkeit eines Softwareprodukts auf andere Umgebungsbedingungen. Dies betrifft vor allem die für die Ausführung der Software zu Grunde liegende Plattform. Die Portabilität eines Softwareprodukts ist hoch, wenn der Anpassungsaufwand bei Übertragung in eine andere Umgebung (anderer Computer oder anderes Betriebssystem) gering ist. Sie ist ein wichtiges Qualitätsmerkmal insbesondere bei langlebigen Softwareprodukten [Siepermann, 2012].

In Abschnitt 2.2.5 wurden bereits betriebliche Anforderungen an technische Verfahren zur automatischen Klassifikation von Web-Inhalten aufgezeigt. Im Fokus stehen dabei vor allem PC Betriebssysteme, Smart Devices und Infrastrukturkomponenten verschiedenster Art. Entsprechend der betrieblichen Anforderungen sind auch Anforderungen an die Portierbarkeit eines Verfahrens auf andere Plattformen zu stellen, wobei neben den technischen auch wirtschaftliche Randbedingungen wie der Aufwand für die Portierung (Kosten, Dauer) zu berücksichtigen sind. Im Folgenden werden einige möglichen Plattformen näher betrachtet.

2.2.7.1 Mobile Plattformen

Wie bereits aufgezeigt wurde, nimmt die mobile Nutzung des Internets insbesondere bei den jüngeren Nutzern rasant zu. Zur Standardausstattung eines Mobiltelefons (ugs.: Handy) gehört inzwischen der Internetzugang. Im Juli 2012 veröffentlichte der Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. (BITKOM) die Ergebnisse einer repräsentativen Marktumfrage, wonach vier von zehn Mobilfunknutzern ein Smartphone (Handy mit Internetzugang) besitzen und 43 % aller Smartphone-Besitzer täglich das Internet nutzen. Nur 42 % aller Smartphone-Besitzer telefonieren täglich mit dem Handy. »Die mobile Internetnutzung per Smartphone ist ein absolutes

[28] Im Rahmen eines Escrow-Vertrags wird der Quelltext von Software in zum Beispiel notarieller Obhut verwahrt und vereinbart, dass – im Falle bestimmter Ereignisse (vor allem der Insolvenz) – dem Kunden Einblick in den Quelltext zu gewähren ist.

Massenphänomen geworden, das in der Breite der Bevölkerung angekommen ist«, sagte BITKOM-Präsidiumsmitglied René Schuster [BITKOM, 2012].

Bei der mobilen Internetnutzung bestehen grundsätzlich die gleichen jugendschutzrelevanten Gefahren wie bei der Nutzung des Internets mittels PC. Einige Eigenschaften von insbesondere Handys eröffnen darüber hinaus weitere potentiell problematische Aspekte:

- Apps: Die speziell für Smartphones entwickelten Anwendungsprogramme verwenden für den Zugriff auf Inhalte, die am heimischen PC über die standardisierten Technologien des WWW zugreifbar sind, teilweise proprietäre Protokolle. Obwohl dabei die gleichen Inhalte aufgerufen werden, können die für das WWW etablierten Schutzmechanismen nicht mehr greifen.[29]
- Integrierte Kamera: Die in nahezu jedem internetfähigen Handy integrierte Kamera ermöglicht es jederzeit und sehr einfach, Bilder und Videos zu erstellen und über das Internet zu verbreiten. Ein Verstoß gegen § 201a (StGB)[30] ist so mit wenigen Fingerbewegungen geschehen, häufig ohne dass die strafrechtliche Relevanz den Kindern und Jugendlichen bekannt ist.
- Persönliches Gerät: Das Handy wird im Gegenteil zu einem häuslichen oder schulischen Computer als persönlicher Gegenstand betrachtet, auf dem die Kontrolle der nutzbaren Funktionen und die Nachvollziehbarkeit konsumierter Inhalte begrenzt sind. Das Gerät ist jederzeit verfügbar und die Nutzungsdauer beziehungsweise -häufigkeit ist technisch nicht kontrollierbar. Die Internetnutzung entzieht sich so der elterlichen Aufsicht und Kinder können im Prinzip jederzeit und überall surfen.

Für Jugendliche ist die wichtigste Anwendung auf dem Smartphone die Nutzung sozialer Netzwerke. Dies bestätigten 56 % der befragten Smartphone-Nutzer in der JIM-Studie 2011. Insbesondere im Kontext von Mobbing über soziale Netzwerke »spielt das Handy eine Rolle, da hier schnell beleidigende Inhalte wie Bilder, Texte oder Filme weiterverbreitet werden können« [MPFS, 2011a].

Das Erfordernis für die Etablierung wirksamer technischer Schutzmaßnahmen für den Jugendmedienschutz hinsichtlich der Nutzung von Smartphones ist längst erkannt. Mobilfunkbetreiber und einige Anbieter sozialer Netzwerke haben dies durch die Vereinbarungen von Selbstkontrollen (vgl. Abschnitt 1.2.4) bestätigt, konnten bislang jedoch vornehmlich organisatorische und beratende Schutzmechanismen etablieren. Die zum Zeitpunkt der Erstellung der vorliegenden Studie am Markt befindlichen Lösungen für Handys realisieren nur einen minimalen Schutz, der den gesetzlichen Mindestanforderungen nicht genügt. So sind zum Beispiel reine Whitelist-basierte Ansätze vorzufinden oder Programme, die hauptsächlich eine Überwachung der Handynutzung durch die Eltern ermöglichen.

Um dem erkannten Bedarf an effektiven Schutzmechanismen Rechnung zu tragen, ist neben der Entwicklung neuer Verfahren vor allem die Frage der Portierbarkeit bestehender – gegebenenfalls echtzeitfähiger – Verfahren und Technologien zu betrachten und als Anforderung an technische Verfahren festzuhalten.

2.2.7.2 Infrastruktur

Wie bereits dargestellt, können Jugendschutzfunktionen auch in Infrastrukturkomponenten realisiert werden. Dies bietet gegenüber der Absicherung einzelner Endgeräte einige Vorteile, die in Kapitel 6 diskutiert und den damit verbundenen Herausforderungen und eventuellen Nachteilen gegenüber gestellt werden. Zum Beispiel müssen Schutzfunktionen und zu Grunde liegende technische Verfah-

[29] Ein Beispiel hierfür ist die altersdifferenzierte Auszeichnung von Webangeboten mittels des age-de.xml-Standards.
[30] § 201a (StGB): Verletzung des höchstpersönlichen Lebensbereichs durch Bildaufnahmen.

ren nicht auf neue Plattformen, die insbesondere im Bereich der Smart Devices einer rasanten Evolution unterliegen, portiert werden. Auch böte ein derartiger Ansatz eine Möglichkeit zur Konfiguration von Nutzerprofilen, die Geräte-übergreifend angewendet werden können. Einige Verfahren könnten aufgrund von Performanz-Anforderungen (vgl. »Echtzeitfähigkeit« in Abschnitt 2.2.5.2) und den recht begrenzten Hardware-Ressourcen von Smart Devices nur seitens eines Diensteanbieters (z. B. Anbieter sozialer Netzwerke oder von Chatrooms) oder in der Internet-(zugangs)-infrastruktur realisiert werden.

Die genannten Gründe zeigen auf, dass technische Jugendschutzstrukturen seitens der Inhalte- und Diensteanbieter, Mobilfunkanbieter, ISPs und weiterer Infrastrukturbetreiber und -anbieter eine sinnvolle und wirksame Ergänzung zu Jugendschutzprogrammen auf den Endgeräten darstellen können. Daher ist die Portierbarkeit bestehender technischer Verfahren auf Infrastrukturkomponenten eine bedeutende Anforderung.

2.2.8 Flexibilität

Als Flexibilität wird die Eigenschaft eines Systems oder Dienstes bezeichnet, auch unter veränderten Rahmenbedingungen funktionsfähig zu sein. Dies betrifft im Rahmen von Software zum Beispiel die Evolution von Softwareumgebungen (wie z. B. neue Betriebssystemversionen) und im betrachteten Fall vor allem die stete Fortentwicklung der Technologien des WWW. Flexibilität betrachtet im konkreten Fall auch die Möglichkeit zur Portierung eines technischen Verfahrens in eine andere Systemumgebung, zum Beispiel ein anderes Endgerätebetriebssystem oder auf serverseitige Infrastrukturkomponenten, wie in Abschnitt 2.2.7 dargestellt.

In der Softwaretechnik wird eine größtmögliche Flexibilität durch den Einsatz technischer Standards realisiert. Hierzu gehören Plattform-übergreifend verwendbare Programmiersprachen und standardisierte Eingabe- und Ausgabeschnittstellen. Konkrete Anforderungen an die Flexibilität einzelner Verfahren ergeben sich auch aus der Integration derselben in größere Systeme oder Umgebungen, wie auch Kapitel 6 aufzeigt.

Bei der Betrachtung des Nutzungskontextes entstehen weitere Anforderungen an die Flexibilität technischer Verfahren. Ausgehend vom *klassischen* WWW besteht zunächst die Anforderung, dass technische Verfahren eine Webseite und die darauf befindlichen Inhalte klassifizieren sollen. Die Webseite ist dabei ein zu einem gegebenen Zeitpunkt als statisch zu betrachtendes Dokument, das über eine eindeutige Adresse erreichbar ist. So lange der Urheber (z. B. Person oder Redaktion) keine Veränderung an der Webseite vornimmt, ist das Dokument reproduzierbar abrufbar. Mit der fortschreitenden Entwicklung und Anwendung neuer Webtechnologien und der Einführung des Web 2.0 entstehen neue Herausforderungen für die Klassifikation von Inhalten im WWW. Die Kombination verschiedener Internet-Datenquellen auf einer Webseite (sog. Mash-Ups), die Personalisierung von Adressen zum Aufruf von Webseiten[31] und die Aggregation von UGC auf Inhalteplattformen stellen die bestehenden Technologien oftmals vor bislang nicht gelöste Probleme. Beispielsweise können die etablierten Jugendschutzprogramme nicht zwischen einzelnen Objekten in einem UGC-Medienportal wie flickr oder YouTube[32] differenzieren, so dass der Zugriff auf das gesamte Portal in Abhängigkeit von der gewählten Altersstufe[33] entweder komplett freigegeben oder gänzlich blockiert wird.

[31] Zum Beispiel durch die Ergänzung von Session-IDs oder anderer Informationen.

[32] Die beiden genannten Portale stehen stellvertretend für sämtliche Plattformen im Internet, über die Benutzer Informationen oder Medien bereitstellen können. Dazu gehören neben den angesprochenen Video- und Fotoportalen auch Musikportale, Diskussionsforen, Tauschbörsen, Marktplätze und Flohmärkte (wie z. B. eBay), Webangebote mit Rezensionen- und Kommentarfunktionen bis hin zu Online-Chats u. v. m.

[33] Beim JusProg Jugendschutzprogramm ist YouTube ab 12 Jahren freigegeben.

Resümierend lässt sich festhalten, dass die Flexibilität von Verfahren für den Einsatz in jedem heute etablierten und zukünftig möglichen Nutzungsszenario des Internets eine wichtige Anforderung darstellt. Grundsätzlich müssen Verfahren dabei in der Lage sein, einzelne Inhaltselemente wie Textabschnitte, Bilder, Videos und Audios zu klassifizieren, aber auch den Kontext jedes Inhaltselements (z. B. innerhalb einer Webseite oder die Menge der auf eine Webseite zeigenden Hyperlinks) bei der Bewertung von Angeboten zu berücksichtigen. Einzelne Klassifikationsverfahren können dabei zur Bewertung der Angebote Indizien beisteuern, während aggregierende Bewertungsverfahren unter Berücksichtigung verschiedener Indizien eine – im Idealfall – altersdifferenzierte Klassifikation des gesamten Angebots vornehmen. Desweiteren wird es im Web 2.0 zunehmend wichtiger, auch einzelne Inhaltselemente identifizieren und klassifizieren zu können, so dass eine Unterscheidung von Inhalten innerhalb eines Portals im Sinne des Jugendmedienschutzes möglich wird.

Anpassbarkeit der Klassifikationsverfahren In Zusammenhang mit Anforderungen an die technische Klassifikation sind quelloffene oder umfangreich konfigurierbare und erweiterbare Verfahren sogenannten Blackbox-Anwendungen vorzuziehen. Gerade im laufenden Betrieb müssen Klassifikationssysteme an neue Anforderungen (z. B. neue jugendschutzrelevante Themen) angepasst werden können.

2.2.9 Skalierbarkeit

Unter Skalierbarkeit wird im Zusammenhang mit Software die Eigenschaft verstanden, Änderungen des Problemumfangs durch eine Anpassung der zur Verfügung stehenden Ressourcen zu bewältigen. Eine Anforderung an die Skalierbarkeit eines Softwaresystems muss immer in einem konkreten (messbaren) Zusammenhang formuliert werden. Im konkreten vorliegenden Fall bedeutet Skalierbarkeit beispielsweise, dass die Menge der verarbeiteten Informationen pro Zeiteinheit durch die Bereitstellung weiterer Ressourcen erhöht (gleichzeitige, parallele Verarbeitung von Informationen – *horizontale Skalierbarkeit*) oder die Geschwindigkeit, mit der Daten verarbeitet werden, durch geeignete Maßnahmen (z. B. das Hinzufügen von mehr Hauptspeicherplatz oder das Austauschen der central processing unit (CPU) durch ein schnelleres Modell) vergrößert werden können soll (*vertikale Skalierbarkeit*).

2.2.10 Sicherheit der technischen Verfahren vor Manipulation durch Dritte

»Informationssicherheit ist der Präventivschutz für Persönlichkeits- und Unternehmens-Informationen und ist auf kritische Geschäftsprozesse fokussiert. Ein solcher Schutz bezieht sich gleichermaßen auf Personen, Unternehmen, Systeme und Prozesse und wird durch Integrität, Verfügbarkeit, Vertraulichkeit, Verbindlichkeit, Nachweisbarkeit und Authentizität erzielt. Die Informationssicherheit soll den Verlust, die Manipulation, den unberechtigten Zugriff und die Verfälschung von Daten verhindern.« [ITwissen.info, 2012a]

Hinsichtlich der Sicherheit sind verschiedene Anforderungen an technische Klassifikationsverfahren zu richten, damit ein erwartungsgemäßer Betrieb der Verfahren gewährleistet werden kann, der nicht von Dritten kompromittierbar ist. So muss sichergestellt werden, dass die Ergebnisse der Verfahren nicht verfälscht werden können oder die Funktionsweise der Verfahren umgangen werden kann (z. B. durch Abschaltung oder Absturz eines Verfahrens). Greifen Verfahren auf gelagerte Daten zurück (z. B. Indizes der Fingerprinting-Technologien, trainierte Konfigurationen oder Listen bereits klassifizierter Webseiten), muss sichergestellt werden, dass die gelagerten Daten selbst nicht kompromittierbar sind oder Datensätze gänzlich ausgetauscht werden können. Sollten die Funktionsweise von Verfahren oder verfahrensspezifische Daten der Vertraulichkeit unterliegen, muss durch tech-

nische und ggf. organisatorische Schutzmaßnahmen sichergestellt werden, dass die Vertraulichkeit unter allen Umständen gewahrt bleibt.[34]

2.2.11 Randbedingungen

Als Randbedingungen für technische Verfahren zur Klassifikation von Web-Inhalten sind nicht-technische Anforderungen an die Integrierbarkeit der Verfahren in betriebsfähige Produkte festzuhalten.

Attraktive Lizenzierbarkeit Von besonderem Interesse ist dabei die Fragestellung, unter welchen Bedingungen ein Verfahren lizenziert werden kann, welche Kosten damit verbunden sind und wie die Erfüllung der gestellten technischen Anforderungen gewährleistet werden kann. Übernimmt der Lizenznehmer mit seinem Produkt dem Endkunden gegenüber eine Funktionsgarantie, deren Erfüllung das lizenzierte Verfahren garantieren muss, sind auch die Fragen der Haftung im Falle des Versagens der Verfahrens zu klären.

Die Bandbreite bei der Lizenzierung technologischer Verfahren umfasst eine Vielzahl von Möglichkeiten von freien Softwarelizenzen, zum Beispiel die Lesser General Public License (LGPL)[35], bis hin zu kostenintensiven kommerziellen Lizenzen. Da neben der eigentlichen Lizenzierung auch Kosten bei der Integration anfallen, müssen Klassifikationsverfahren daher über geeignete Lizenzierungsmodelle verfügen, die im Idealfall keine Kosten erzeugen beziehungsweise eine Amortisation über die Zeit des Einsatzes ermöglichen. Interessant sind daher insbesondere Freeware- beziehungsweise Open Source Software.

Software- und Dokumentationsqualität Die eingesetzte Klassifikationssoftware muss gängigen Industriestandards im Bereich Softwarequalität und Dokumentation genügen. Eine unzureichend qualitätsgesicherte oder dokumentierte Software verursacht ansonsten unnötige Kosten im gesamten Betriebsprozess.

Aktive Updatekanäle Die eingesetzten Verfahren sollten in aktiver Entwicklung sein. Insbesondere bei nicht quelloffener Software, die nicht durch Dritte übernommen und weitergeführt werden kann, ist diese Anforderung von großer Bedeutung für den langfristigen Einsatz.

2.2.12 Zusammenfassung: nicht-funktionale Anforderungen an Klassifikationsverfahren

Wie dargelegt bestehen zahlreiche nicht-funktionale Anforderungen an technische Verfahren zur Klassifikation von Web-Inhalten. Die gewählte Systematik zur Ordnung der Anforderungen macht deutlich, dass es sich dabei einerseits um »übliche Anforderungen an gute Software« handelt. Andererseits muss zahlreichen Spezifika hinsichtlich der Komplexität des Themenfeldes Jugendmedienschutz und des sich technologisch rasant entwickelnden Einsatzgebietes WWW – beziehungsweise weiter gefasst und heute de facto nicht mehr scharf differenzierbar: Internet – Rechnung getragen werden. Gewiss wird kein Verfahren allen aufgezeigten Anforderungen genügen können. Vielmehr sollen die Anforderungen als Orientierung dienen, mit welchem Ziel und in welchem Kontext einzelne Verfahren zur Anwendung kommen können und wie verschiedene Verfahren sinnvoll zu einem effektiven Schutzmechanismus verbunden werden können. Die aufgezeigten Anforderungen dienen

[34] Dies betrifft zum Beispiel die dem BPjM-Modul zugrunde liegenden Daten der indizierten Webseiten. Diese dürfen aus gesetzlich verankerten Gründen nicht öffentlich zugänglich gemacht werden. Dieser gesetzlichen Vorgabe wurde bei der Implementierung des BPjM-Moduls Rechnung getragen.

[35] siehe http://www.gnu.org/licenses/lgpl.html, Stand 20.11.2012

in den nachfolgenden Kapiteln der vorliegenden Studie als Orientierung zur Bewertung der technischen Verfahren (vgl. Kapitel 5) und der Betrachtung derer Integrationsmöglichkeiten (vgl. Kapitel 6).

2.3 Exkurs: Anforderungen an Jugendschutzsoftware

Jugendschutzsoftware ist ein technische Mittel, mit dem Kinder und Jugendliche vor dem Zugriff auf beeinträchtigende oder jugendgefährdende Inhalte geschützt werden sollen. Einige der in der vorliegenden Studie betrachteten technischen Verfahren (vgl. Kapitel 4) könnten im Rahmen von Jugendschutzsoftware eingesetzt werden, um jugendschutzrelevante Inhalte zu detektieren. Die Eignung eines Verfahrens für dessen Einsatz im Rahmen von Jugendschutzsoftware wird im Folgenden an Hand der Anforderungen an Jugendschutzsoftware betrachtet.

2.3.1 Quellen für Anforderungen

So unterschiedlich die verwendeten Begrifflichkeiten sind, so vielfältig können auch Anforderungen an Jugendschutzsoftware definiert werden. Im Folgenden werden einige Quellen für Anforderungen näher betrachtet, die als etabliert angesehen werden können.

2.3.1.1 Technische Anforderungen zur Anerkennung von Jugendschutzprogrammen in Deutschland

Für die offizielle Anerkennung von Jugendschutzprogrammen im Sinne des JMStV hat die KJM einen Katalog von *Kriterien der KJM für die Anerkennung von Jugendschutzprogrammen im Bereich des World Wide Web* [KJM, 2011a] herausgegeben, der Anbietern von Jugendschutzprogrammen offen legt, welche technischen und organisatorischen Anforderungen gemäß der gesetzlichen Vorgabe überprüft werden. Diese wurden bereits in Abschnitt 1.4.1 ausführlich vorgestellt. Sie umfassen die von der KJM als notwendig erachteten Mindestanforderungen für die offizielle Anerkennung sowie weitergehende Empfehlungen der KJM, stellen dabei aber keine Beschränkung eines möglicherweise sinnvollen Funktionsumfangs von Jugendschutzsoftware dar.

2.3.1.2 Weitergehende Anforderungen an Jugendschutzsoftware

Neben den im vorstehenden Abschnitt genannten Anforderungen der KJM können weitere (optionale) Anforderungen an Jugendschutzsoftware definiert werden, die sinnvolle Ergänzungen zu den regulatorischen Vorgaben sein können. Nachfolgend werden Anforderungen an Jugendschutzsoftware aus verschiedenen Kontexten heraus betrachtet:

Standardisierte Anforderungen in Großbritannien Im Rahmen der seit 2001 tätigen Task Force *Child Protection on the Internet* der britischen Regierung wurden Kriterien zur Beurteilung von Internet-Zugangskontrollsystemen entwickelt. Die Beurteilung der verfügbaren Systeme anhand objektiver Kriterien soll eine Vergleichbarkeit der zur Verfügung stehenden Produkte, Dienstleistungen, Werkzeuge oder anderer Systeme für den kontrollierten Zugang zum Internet ermöglichen und insbesondere Eltern als Orientierung zur Auswahl von Systemen dienen. Die Kriterien wurden in eine technische Spezifikation (*PAS 74:2008* [BSI, 2008]) überführt, die Minimalanforderungen an Internet-Zugangskontrollsysteme definiert, sich primär an die Hersteller der Systeme richtet und die

Erteilung neutraler Konformitätszertifikate ermöglicht. Die PAS 74 definiert Anforderungen in sechs Bereichen:

- Gebrauchstauglichkeit hinsichtlich Installation, Konfiguration und Nutzung
- Effektivität (Leistungsfähigkeit)
- Minimaler Funktionsumfang
- Aktualisierbarkeit
- Verständlichkeit der Nutzerführung
- Kundensupport und Kommunikation

Das standardisierte Konformitätszertifikat verschafft dem Kunden beziehungsweise Nutzer Gewissheit und Vertrauen in die Leistungsfähigkeit eines Systems:

- Unangemessene Inhalte können blockiert werden.
- Unangemessene Kommunikation über Internet-basierte Dienste kann blockiert werden.
- Veränderungen an der Konfiguration oder Umgehen des Schutzsystems wird angemessen verhindert.
- Das System bietet ein angemessenes Schutzniveau vor unangemessenen Inhalten.
- Das System ist individuellen Bedürfnissen entsprechend konfigurierbar.
- Das System wird regelmäßig aktualisiert.
- Das System kann auf Wunsch des Nutzers (Eltern) deaktiviert beziehungsweise deinstalliert werden.

Marktanalyse Jugendschutzsoftware Die europäische Kommission fördert seit dem Jahr 2005 die kontinuierliche Evaluation des Marktes für Systeme mit Kinder- und Jugendschutzfunktionen. Im Rahmen der beiden SIP-BENCH Projekte[36] wurden regelmäßig die auf dem Markt verfügbaren Lösungen evaluiert. Die Untersuchungs- und Bewertungskriterien sind dabei vergleichbar mit den in PAS 74 definierten Anforderungen.

2.3.2 Systematisierung der Anforderungen

Im Folgenden werden die im Rahmen der vorliegenden Studie identifizierten Anforderungen systematisiert wiedergegeben. Die erhobenen Anforderungen an Jugendschutzsoftware gehen dabei über die in Deutschland gesetzlich verankerten Minimalanforderungen an die Anerkennung von Jugendschutzprogrammen hinaus und umfassen ebenfalls sinnvolle und wünschenswerte Funktionalität.

2.3.2.1 Funktionalität

Als grundsätzliche Funktionalität von Jugendschutzsoftware sind den verschiedenen Quellen nach folgende Anforderung gemein:

- Jugendschutzsoftware soll den Zugriff auf ungeeignete Inhalte verhindern.
- Jugendschutzsoftware soll den Zugriff auf geeignete Inhalte nicht blockieren.
- Die Filterfunktionalität von Jugendschutzsoftware soll für Eltern deaktivierbar sein.
- Jugendschutzsoftware soll über geeignete Voreinstellungen verfügen.
- Der gewünschte Schutzgrad im Hinblick auf die Altersstufe soll konfigurierbar sein.

[36] SIP-BENCH, 2005-2008: [European Commission, 2012c]; SIP-BENCH 2, 2009-2012: [European Commission, 2012d]

- Der gewünschte Schutzgrad soll für jeden Nutzer individuell konfigurierbar sein. Dies schließt individuelle, durch die Eltern erstellte Whitelists und Blacklists ein.

Die Anforderungen können durch nachfolgend genannte Funktionen realisiert beziehungsweise durch diese ergänzt werden:

Positivliste (Whitelist) Webseiten speziell für Kinder und Angebote ohne Jugendschutzrelevanz können auf einer Positivliste zusammengefasst werden, die als Grundlage für einen sicheren Surfraum dienen soll. Jugendschutzsoftware kann als Standardeinstellung für Kinder zum Beispiel nur Webseiten zulassen, die auf Positivlisten vermerkt sind. Positivlisten entstehen durch die redaktionelle Beurteilung von Webseiten durch eine Institution oder Nutzergemeinschaft und dienen insbesondere jüngeren Kindern als sicherer Einstieg in die Online-Welt. Für Jugendliche wird dieser Ansatz bislang nicht verfolgt, da die redaktionelle Erstellung und Wartung von Positivlisten sehr aufwändig ist und der somit gesicherte Ausschnitt aus dem gesamten WWW den Bedürfnissen Jugendlicher nicht gerecht werden kann. Die prominentesten Positivlisten in Deutschland führen derzeit vertikale Suchmaschinen wie fragFINN,[37] Blinde Kuh[38] und Helles Köpfchen.[39]

Negativliste (Blacklist) Auf einer Negativliste werden Angebote gesammelt, die jugendschutzrelevant sind und zu denen Kinder und Jugendliche – zumindest bis zu einer bestimmten Altersstufe – keinen Zugang erhalten sollen. Negativlisten können von Jugendschutzsoftwareherstellern, staatlichen Einrichtungen (z. B. BPjM[40]), Communities, Organisationen, Verbänden, Vereinen oder Unternehmen erstellt werden und als Grundlage für die altersdifferenzierte Sperrung von Angeboten durch Jugendschutzsoftware dienen. Negativlisten können redaktionell oder mittels technischer Verfahren automatisch generiert werden.

Elternwhite-/-blacklists Ergänzend zu den Voreinstellungen von Jugendschutzsoftware durch Positiv- und Negativlisten Dritter kann Jugendschutzsoftware die Möglichkeit vorsehen, dass Eltern explizit Angebote für ihre Kinder freischalten oder sperren. Diese individuelle Flexibilität ist wichtig, um die Wirkungsweise von Jugendschutzsoftware dem individuellen Entwicklungsstand eines Kindes anpassen zu können und den Eltern eine weitreichende Kontrolle über die Wirkung der verwendeten Jugendschutzsoftware zu gestatten. In der Praxis spielen vor allem Homepages lokaler Einrichtungen (z. B. Sportvereine, Schule, …) eine große Rolle für die Elternwhitelists, da solche Webangebote in der Regel nicht in den globalen Whitelists enthalten sind.

Echtzeitfilterung Die Filterung von Angeboten zum Zeitpunkt des Zugriffs (*on-the-fly*) ist eine sinnvolle Ergänzung zu den durch Positiv- oder Negativlisten bereits beurteilten Angeboten. Durch derartige Verfahren wird zwangsläufig der Zugriff auf Inhalte zeitlich verzögert. Eine wesentliche Anforderung an die eingesetzten Verfahren ist dabei, dass die entstehende Verzögerung für den Nutzer akzeptabel bleibt. Außerdem darf die Qualität der Echtzeitfilterung nicht deutlich schlechter sein als gegenüber anderen Verfahren, denn insbesondere eine zu große Quote von Overblocking könnte den Nutzer vom Einsatz der Software abschrecken.

»Ruft der Nutzer ein Angebot auf, das weder in der White- noch in der Blacklist enthalten ist, untersucht die Software die gewünschten Inhalte noch vor der Anzeige mit Hilfe von Erkennungsalgorithmen, um möglicherweise unerwünschte Inhalte zu erkennen und diese ggf. entsprechend auszufiltern. Hierbei werden automatische objektbasierte, textsemantische oder Bilderkennungsverfahren genutzt, die zum Beispiel auf bestimmte Stichworte, Ausdrucksweisen, Symbole oder Bildinhalte ansprechen. […] Da eine on-the-fly-Klassifikation den abgerufenen Inhalt direkt analysiert, sind

[37] siehe http://www.fragfinn.de/, Stand 26.10.2012; fragFINN umfasst über 10 100 Domains [fragFINN, 2012c].
[38] siehe http://www.blinde-kuh.de/, Stand 26.10.2012
[39] siehe http://www.helles-koepfchen.de/, Stand 26.10.2012
[40] Die von der BPjM geführten Negativlisten enthalten Internetangebote, die als jugendgefährdend eingestuft wurden.

derartige Verfahren nicht an URLs gebunden, sondern können (theoretisch) jede Inhaltsform untersuchen, darunter auch Chatkommunikation, Inhalte von Social Media Angeboten wie etwa sozialen Netzwerkplattformen sowie Newsgroup- und Instant Messaging-Inhalte.« [HBI, 2011]

Zeitkonten Die Festlegung von Surfzeitbeschränkungen oder bestimmten Zeiträumen, in denen die Kinder online gehen dürfen, sind eine häufig als sinnvoll angesehene Ergänzung zu dem eigentlichen Schutzprinzip von Jugendschutzsoftware. Sie wird in der Regel aber nicht als Funktion des Jugendmedienschutzes angesehen und gehört dementsprechend auch nicht zu den Anforderungen an die Anerkennung eines Jugendschutzprogramms.

Geräte-übergreifende Konfiguration Kinder und Jugendliche nutzen verschiedene Geräte für den Zugang zum Internet. Neben dem heimischen PC kommen vor allem bei Jugendlichen Smartphones zum Einsatz, aber auch Spielekonsolen und zukünftig in zunehmendem Maße auch Smart TVs.[41] Die Möglichkeit zur Konfiguration eines Jugendschutzprofils, dass den Internetzugang auf allen zur Verfügung stehenden Endgeräten gleichermaßen schützt, könnte eine sinnvolle Ergänzung für den umfassenden Jugendschutz darstellen. Dabei sind aber in jedem Fall datenschutzrechtliche Rahmenbedingungen und das Prinzip der *Datensparsamkeit* insbesondere im Hinblick auf die Daten von Kindern und Jugendliche zu berücksichtigen.

Monitoring und Logging Jugendschutzsoftware bietet auch die Möglichkeit zur Kontrolle der besuchten Webseiten. Mittels der Aufzeichnung der Aktivitäten von Kindern und Jugendlichen und der Bereitstellung von Berichten können Eltern analysieren, wie ihre Kinder das Internet nutzen und ggf. die Konfiguration der Jugendschutzsoftware anpassen. Derartige Funktionen sind in einigen verfügbaren Produkten enthalten und werden als wichtige Funktion beworben. Zu beurteilen, inwieweit eine derartige Überwachung in das erzieherische Konzept passt, bleibt den Eltern vorbehalten. Auch ist juristisch zu hinterfragen, ob das Recht auf die Privatsphäre eines Kindes hierbei verletzt wird (vgl. Artikel 16 UN-Kinderrechtskonvention [Vereinte Nationen, 1989]).

Schutz persönlicher Daten »Freizügige Fotos, Infos über Stress mit den Eltern oder über Zoff in der Clique – das sogenannte Selbstoffenbarungsverhalten ist besonders bei Jugendlichen zwischen 15 und 17 Jahren ausgeprägter als bei anderen Altersgruppen – und damit ein Risiko, weil fraglich ist, ob und welche vertraulichen Inhalte überhaupt in sozialen Netzwerken wie Facebook Eingang finden sollten. Zwar hat die Mehrheit der 12- bis 24-jährigen Nutzer restriktive Datenschutzeinstellungen (48 % ›Wenigoffenbarer‹ und 39 % ›Privatsphäre-Manager‹), aber jeder siebte (14 % ›Vieloffenbarer‹) verwendet recht offene Einstellungen, hat einen hohen Anteil an unbekannten Kontakten und zeigt zugleich ein aktives Kommunikationsverhalten im Netz.« [LfM, 2012]

Zu diesen Ergebnissen kommt die Studie *Digitale Privatsphäre: Heranwachsende und Datenschutz auf Sozialen Netzwerkplattformen* [Schenk et al., 2012] der LfM.

Jugendschutzsoftware könnte die – insbesondere unbeabsichtigte – Veröffentlichung persönlicher Nutzerdaten verhindern. Zu diesen Daten sind neben Name, Anschrift, Telefonnummer etc. auch auf dem verwendeten Endgerät gespeicherte Adress- und Telefonbücher, Geolokalisationsdaten (z. B. von mobilen Endgeräten mit GPS-Empfängern) und Daten über per Gesichtserkennung[42] identifizierte Personen zu zählen.

[41] Aktuell bieten zahlreiche Fernsehhersteller (Philips, Sharp, LG, Loewe, Samsung, Sony, Panasonic) Geräte mit Zugang zum Internet an. Neben herstellerspezifischen Apps bieten diese Geräte auch den Zugang zum WWW über einen integrierten Webbrowser.

[42] Eine solche Funktion ist zum Beispiel in einigen Bildverwaltungsprogrammen integriert, z. B. iPhoto von Apple oder Picasa von Google [Garfinkel und Rosenberg, 2009].

Auch das Personen-Tagging hochauflösender Fotografien von öffentlichen Veranstaltungen ist unter dem Schutz persönlicher Daten zu verstehen. So wurde zum Beispiel 2011 im Rahmen eines Projektes des WDR ein Festival in Bonn fotografiert und bis dato wurden über 1500 Teilnehmer identifiziert und mit ihren Facebook-Profilen verknüpft (vgl. Abbildung 2.13) [WDR, 2012]. Derlei *interaktive Gigapixel-Panoramaaufnahmen* erfreuen sich zunehmender Beliebtheit. Weitere Aufnahmen wurden zum Beispiel in Fußballstadien[43] erstellt.

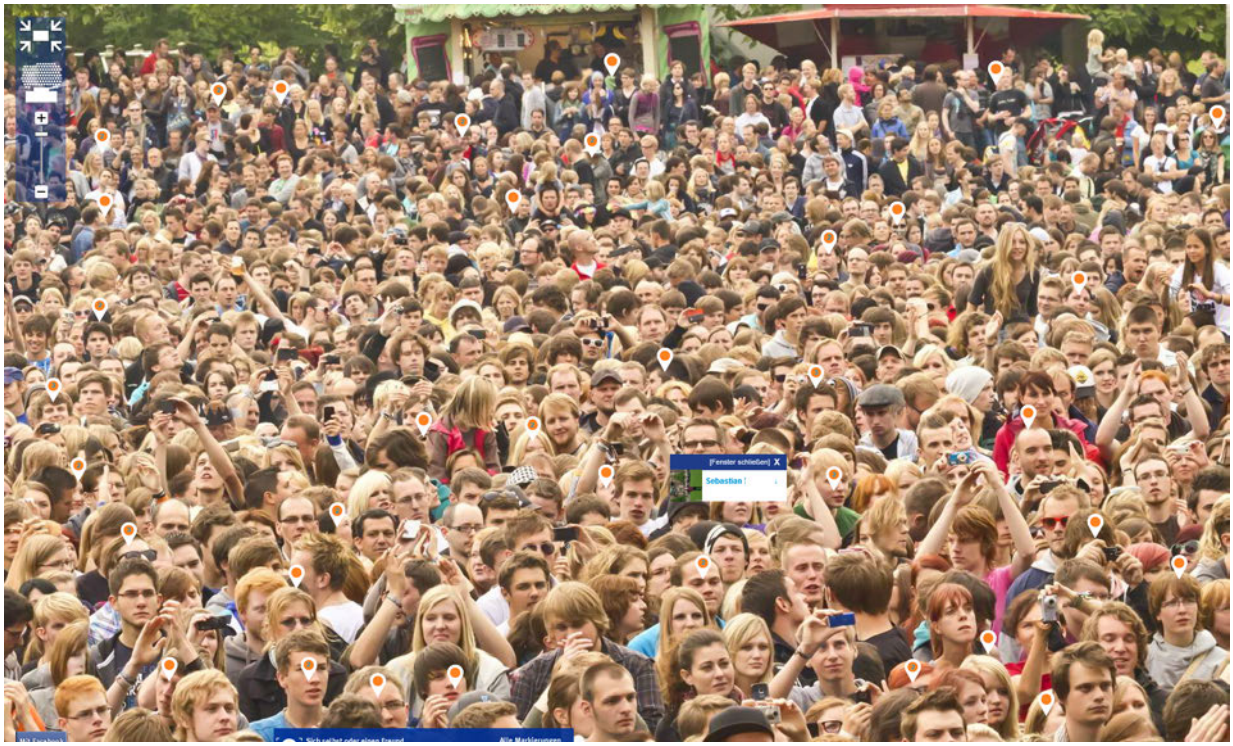


Abbildung 2.13
Personenidentifikation auf öffentlichen Veranstaltungen

Schutz vor eigener Täterschaft Ebenso wie der Schutz vor der – insbesondere unbeabsichtigten – Konfrontation mit jugendschutzrelevanten Inhalten sollen Kinder und Jugendliche auch davor geschützt werden, unangemessene Inhalte zu produzieren und zu verbreiten oder Autor unangemessener Kommunikation zu werden [Cybion et al., 2012]. Dies ist insbesondere relevant, weil Kinder und Jugendliche nicht immer die Grenzen des *Erlaubten* kennen und richtig einschätzen können, was verboten ist (vgl. auch 2.2.7.1).

2.3.2.2 Gebrauchstauglichkeit (Usability)

Ein wesentliches Merkmal für die Akzeptanz von Jugendschutzsoftware durch die Eltern ist die Einfachheit und Verständlichkeit des eingesetzten Produkts und dessen Wirkungsweise. »Um einen großen Teil der Eltern zum Einsatz von Jugendschutzprogrammen zu bewegen, müssen sie benutzerfreundlich sein, dürfen keine unrealistischen Anforderungen an den technischen Sachverstand der Benutzer stellen und keine unangemessen hohen Kosten verursachen.« [KJM, 2011a]

Installation/Deinstallation Der Installationsvorgang soll von kurzer Dauer und ohne besondere technische Kenntnisse durchführbar sein. Jeder Schritt des Installationsvorgangs soll für Eltern verständlich sein. Die Grundeinstellungen für die Installation sollen einen effektiven Schutz gewähr-

[43] siehe zum Beispiel http://wembley360.wembleystadium.com/, Stand 29.10.2012

leisten. Eine weitergehende Konfiguration während der Installation könnte über eine Option für *erfahrene Nutzer* möglich sein.

Eltern sollten in der Lage sein, einmal installierte Jugendschutzsoftware ebenso problemlos wieder zu deinstallieren. Gerade da Jugendschutzsoftware direkt in die Kommunikation des geschützten Endgeräts eingreift, ist eine Entfernung aller Software-Bestandteile und eine vollständige Wiederherstellung des Originalzustands besonders wichtig.

Konfiguration Die Konfiguration der Wirkungsweise von Jugendschutzsoftware soll für die Eltern verständlich und leicht zu erlernen sein. Die Konfiguration soll entlang verständlicher Schutzstufen hinsichtlich eines altersdifferenzierten Internetzugangs und/oder verschiedener Inhaltekategorien erfolgen. Der Anforderung nach möglichst einfacher Konfiguration stehen auch Wünsche von technisch versierteren Nutzern gegenüber, wie zum Beispiel die Möglichkeit der Wahl zwischen verschiedenen vorkonfigurierten Schutzniveaus innerhalb einer Altersstufe, etwa »hoch«, »moderat« und »niedrig« oder die Konfiguration von Jugendschutzfunktionalität auf mobilen Endgeräten über den heimischen PC. Beide Anforderungen, die nach möglichst einfacher und gegen Bedienungsfehler sicherer Konfiguration sowie die Anforderung nach möglichst differenzierten Konfigurationsmöglichkeiten, sind vor dem Hintergrund der angesprochenen Zielgruppe gegeneinander abzuwägen.

Betrieb Im Betrieb darf Jugendschutzsoftware keine störenden Wechselwirkungen mit anderen Programmen aufweisen. Programmbedingte Verzögerungen beim Zugriff auf Internetangebote sollen in einem für den Nutzer akzeptablen Rahmen bleiben. Wird der Zugriff auf Inhalte blockiert, soll eine leicht verständliche Mitteilung auf die Ursache hinweisen und möglichst auf alternative Angebote verweisen.

Administration Eventuelle von Jugendschutzsoftware erzeugte Berichte und Überwachungsmeldungen sollen leicht verständlich und einfach zu handhaben sein. Die manuelle Sperrung oder Freigabe von Angeboten durch Eltern soll auf einfache und schnelle Weise möglich sein.

Aktualisierung Die Aktualisierung von Jugendschutzsoftware und der darin verwendeten Listen soll regelmäßig und in nutzerfreundlicher Weise erfolgen. In den Grundeinstellungen soll eine automatische und nutzerunabhängige Aktualisierung voreingestellt sein.

2.3.2.3 Sicherheit

Der Einsatz von Jugendschutzsoftware muss einem Mindestmaß an Sicherheitsanforderungen genügen. Die gesetzlichen Anforderungen gemäß JMStV entsprechen dabei einem moderaten Sicherheitsniveau, denn Kindern und Jugendlichen soll in diesem Sinne die Wahrnehmung entsprechender Angebote »wesentlich erschwert« (vgl. § 5 Absatz 3 Nummer 1 JMStV) werden. Die Anforderungen der Eltern gehen zum Teil darüber hinaus, denn die Verlässlichkeit auf die Funktion des Programms umfasst auch Aspekte der Sicherheit. So sollen Kinder und Jugendliche den gewährten Schutz nicht einfach umgehen, das Programm abschalten oder deinstallieren können. Dies muss auch Möglichkeiten alternativer Zugriffswege auf Angebote umfassen, wie zum Beispiel die Nutzung von Proxies oder die Verwendung alternativer Webbrowser.

2.3.2.4 Effektivität

Die Effektivität von Jugendschutzsoftware wird häufig an Hand der Quote der nicht blockierten unerwünschten Inhalte (Underblocking) im Verhältnis zur Quote der blockierten unbedenklichen Inhalte

(Overblocking) im Rahmen eines vom Umfang her begrenzten Testszenarios beurteilt. Es ist Aufgabe der Jugendschutzsoftware, an Hand verschiedener Kriterien zu entscheiden, ob ein aufgerufenes Angebot blockiert wird oder nicht. Hierzu stehen dem Programm verschiedene Kriterien zur Verfügung, zum Beispiel:

- die vom Programm verwendeten Positivlisten,
- die vom Programm verwendeten Negativlisten,
- die von Eltern explizit zugelassenen oder gesperrten Angebote,
- die Altersauszeichnung einer Webseite,
- die Ergebnisse von integrierten Echtzeitanalyseverfahren.

Die Effektivität von Jugendschutzsoftware kann demnach am Ergebnis der Aggregation und Priorisierung der verschiedenen Klassifikationen für ein Angebot gemessen werden. Die Effektivität eines Klassifikationsmechanismus hingegen kann im direkten Vergleich mit den altersdifferenzierten Ergebnissen der anderen Mechanismen bestimmt werden.

2.3.2.5 Anforderungen an die Integration von technischen Jugendschutzstrukturen und Jugendschutzsoftware

Als technische Jugendschutzstrukturen werden Mechanismen bezeichnet, die eine altersdifferenzierte Klassifikation von Angeboten durch Dritte vorsehen und für die Nutzung durch Jugendschutzsoftware bereitgestellt werden. Diese können zum Beispiel sein:

- das BPjM-Modul, welches die in den BPjM-Listen erfassten jugendgefährdenden Webseiten umfasst,
- Altersauszeichnungen von Einrichtungen der freiwilligen Selbstkontrolle (z. B. FSK, USK, …),
- Altersauszeichnungen von Webseiten (age-de.xml Auszeichnung),
- Positivlisten vertikaler Suchmaschinen für Kinder und Jugendliche,
- öffentliche Positiv- und Negativlisten von Organisationen, Verbänden oder Interessensgruppen wie zum Beispiel Eltern oder Pädagogen (sog. Crowd-generierte Listen) oder
- inhalte- oder altersspezifische Voreinstellungen von Suchmaschinen und Inhalteportalen (z. B. YouTube, flickr, …).

Jugendschutzsoftware sollte von möglichst vielen bestehenden Strukturen profitieren und die vorgenommenen Altersklassifikationen in geeigneter Weise berücksichtigen.

2.3.3 Zusammenfassung: Anforderungen an technische Verfahren im Kontext von Jugendschutzsoftware

Es bestehen zahlreiche mögliche Anforderungen an Jugendschutzsoftware, die von unterschiedlichen Quellen erhoben wurden (vgl. Abschnitt 2.3.1). Ein Großteil dieser Anforderungen betrifft die Implementierung der Software hinsichtlich Funktionalität und Gebrauchstauglichkeit. Einige Anforderungen an technische Verfahren lassen sich aus den Anforderungen an Jugendschutzsoftware ableiten:

Integrierbarkeit Technische Verfahren für die Detektion jugendschutzrelevanter Inhalte müssen sich in geeigneter Weise in Jugendschutzsoftware integrieren lassen, zum Beispiel durch ein Konzept an offenen Schnittstellen, über die Klassifikationsaufträge an Verfahren übergeben und die Klassifikationsergebnisse für Jugendschutzsoftware nutzbar zurückgemeldet werden. Dies umfasst sowohl

– entsprechend geeignete – echtzeitfähige Verfahren, die Client-seitig in Jugendschutzsoftware integriert werden können, als auch alle Verfahren, die seitens der Anbieter von Jugendschutzsoftware Server-seitig verwendet werden können.

Performanz Für den Verwendungszweck von technischen Verfahren ist die Performanz des Verfahrens von großer Bedeutung. In Echtzeitfiltersystemen können nur Verfahren zum Einsatz kommen, die eine maximale Verarbeitungsdauer[44] nicht übersteigen.

Effektivität und Zuverlässigkeit Technische Verfahren müssen jugendschutzrelevante Inhalte zuverlässig detektieren und dürfen Inhalte ohne Jugendschutzrelevanz nicht fälschlicherweise als problematisch beurteilen. Für die Klassifikation benötigen technische Verfahren detektierbare Kriterien, diese werden in Kapitel 3 näher betrachtet. Die Effektivität und Zuverlässigkeit technischer Verfahren hängt in starkem Maße von den jeweils zu Grunde liegenden Kriterien ab.

2.4 Exkurs: Anforderungen an technische Jugendschutzstrukturen

Als *technische Jugendschutzstrukturen* werden alle technisch umgesetzten und technologisch nutzbaren Maßnahmen zur Förderung des Jugendmedienschutzes im Internet bezeichnet. Geschaffen werden die Strukturen entweder von den Anbietern von Inhalten, zum Teil in Kooperation mit den Einrichtungen der freiwilligen Selbstkontrolle, den Anbietern von Diensten, von staatlicher Seite oder von unabhängigen Dritten wie Initiativen oder Unternehmen.

Heute sind bereits zahlreiche technische Jugendschutzstrukturen im Bereich des Internets etabliert, wie die folgende Aufzählung von Beispielen darlegt:

- Altersfreigaben für Filme und Computerspiele[45]. Diese werden von den Herstellern im Rahmen der freiwilligen Selbstkontrolle und in Kooperation mit den entsprechenden Einrichtungen (z. B. FSK, USK, …) vorgenommen,
- Altersauszeichnungen von Webseiten (age-de.xml-Auszeichnung),
- inhalte- oder altersspezifische Voreinstellungen von Inhalteportalen für UGC (z. B. YouTube, flickr, …),
- altersspezifische Voreinstellungen in Suchmaschinen (*SafeSearch*-Funktionen, vgl. Abschnitt 2.4.1),
- zentral gepflegte Sperrlisten wie das BPjM-Modul,
- Positivlisten vertikaler Suchmaschinen für Kinder und Jugendliche und ausgewiesene Kinderportale,
- WWW-Filterlisten von Initiativen oder Projekten wie zum Beispiel das Projekt *Jugendschutzfilter für Schulen*,[46]
- Kinder- und Jugendschutzfunktionen für einzelne Angebote und Anwendungen im WWW (vgl. Abschnitt 2.4.2).

Nachfolgend werden einige der vorhandenen Jugendschutzstrukturen näher betrachtet.

[44] Diese muss im jeweiligen Anwendungsfall quantitativ definiert werden, da sich eine pauschale Anforderung aufgrund der Unterschiede möglicher Anwendungsfälle nicht festlegen lässt.

[45] Dies sind vor allem Computerspiele, die über Trägermedien vertrieben werden, da eine analoge Regelung für Browsergames nicht existiert.

[46] siehe http://www.jugendschutzfilter.net/, Stand 15.10.2012

2.4.1 SafeSearch-Funktionen

Zu den technischen Jugendschutzstrukturen können auch die *SafeSearch*-Funktionen der großen Suchmaschinen (Google [Google, 2012a], Bing [Microsoft, 2012b], Yahoo [Yahoo!, 2012]) gerechnet werden. Üblicherweise gibt es 3 Schutzstufen (»aus«, »moderat«, »strikt«). Google gibt an, dass sich die SafeSearch-Funktionalität auf die Filterung »sexuell expliziter Inhalte« bezieht [Google, 2012b]. Bing filtert nach eigenen Angaben »nicht jugendfreie Inhalte« aus den Suchergebnissen heraus [Microsoft, 2012a]. Yahoo blockiert nach eigenen Angaben mit SafeSearch »Erwachseneninhalte« [Yahoo!, 2012]. Den Ausführungen der Suchmaschinenanbieter ist nur schwer bis gar nicht zu entnehmen, welche Inhalte tatsächlich mit der SafeSearch-Funktion gefiltert werden und welche technisch umgesetzten Regelwerke – die auf gesellschaftlichen Moralvorstellungen aufbauen – den Filtern zu Grunde liegen. Stichprobenartige Untersuchungen haben ergeben, dass sexuelle oder pornografische Inhalte im strikten SafeSearch Modus im Allgemeinen recht gut gefiltert werden, in anderen Inhaltsgruppen wie Selbstgefährdung oder Gewalt zeigen sich jedoch keine Unterschiede in den Suchergebnislisten für deutschsprachige Webseiten im Vergleich zur abgeschalteten SafeSearch-Funktion.

2.4.2 Kinder- und Jugendschutzfunktionen für einzelne Angebote und Anwendungen im WWW

Stellvertretend für verschiedene Angebote und Dienstleistungen wird nachfolgend ein konkretes Beispiel einer bereits verfügbaren Dienstleistung vorgestellt. Die Wahl dieses Beispiels erfolgt nicht mit werblicher Absicht und soll auch keinen empfehlenden Charakter haben.

Die Firma myON-ID Media GmbH bietet seit Ende 2011 den Online-Service *secure.me* zum Schutz von Nutzern von sozialen Netzwerken und Apps. Ziel des Angebots ist der Schutz der Reputation der Nutzer mittels der Analyse von gewählten Privatsphäreneinstellungen und der Analyse von Kommunikation (Posts) durch oder über eine Person. In der Beschreibung der myON-ID Media GmbH ist zu lesen: »secure.me Privacy Control überwacht Ihre sensiblen persönlichen Daten und zeigt Ihnen, wie Sie diese schützen können. Mit der App-Security innerhalb von secure.me lassen sich Drittanwendungen analysieren, die Zugriff auf Ihr Facebook-Profil haben. secure.me informiert Sie, welche Daten die Apps abrufen und welche Informationen sie auf Ihrem Profil und in Ihrem Freunde-Netzwerk hinterlassen. secure.me Reputation Guard setzt semantische Sprach- und biometrische Gesichtserkennungs-Technologie ein, um Ihre Reputation auf Facebook kontrollieren zu können. Neben dem eigenen Facebook-Profil überwacht secure.me auch Ihr Freunde-Netzwerk.« [secure.me, 2012]

»Das im secure.me enthaltene Facebook-Sicherheitspaket schützt Daten auf Facebook und unterstützt den Jugendschutz. […] Der Jugendschutzfilter überprüft Kommentare und Posts auf jugendgefährdende Inhalte und informiert Sie als Eltern, wenn ein kritischer Beitrag erscheint. So behalten Sie den Überblick über die Inhalte, mit denen ihr Kind umgeben ist und können kritische Themen im geschützten Rahmen Ihrer Familie besprechen.« [WEB.DE, 2012]

Angebote wie *secure.me* sind nicht in erster Linie aus der Motivation des Jugendmedienschutzes heraus entstanden und stellen allein kein angemessenes Jugendschutzprogramm im Sinne des JMStV dar. Sie sind speziell auf einzelne Webangebote ausgerichtet und können in diesem Kontext als sinnvolle Ergänzung zu Jugendschutzsoftware und als Teil der technischen Jugendschutzstrukturen betrachtet werden.

2.4.3 Bestehende Altersklassifikationssysteme im WWW

Für verschiedene Arten von Inhalten im WWW bestehen bereits Altersklassifikationen. Bestehende Altersklassifikationen sollten von den Jugendschutzprogrammen berücksichtigt werden. Auch Verfahren zur altersdifferenzierten Klassifikation von Inhalten können von bestehenden Altersklassifikationen profitieren, in dem zum Beispiel bereits redaktionell bewertete Inhalte als Trainingsdaten für lernende Verfahren verwendet werden.

2.4.3.1 Altersfreigaben in Filmportalen

Im WWW bieten eine Reihe von Portalen Filme und Fernsehsendungen auf Abruf an. Anbieter sind beispielsweise die öffentlich-rechtlichen Rundfunkanstalten Deutschlands, die über ihre Mediatheken[47] die im Fernsehen ausgestrahlte Sendungen eine in der Regel begrenzte Zeit lang zur Verfügung stellen. Auch die privaten Fernsehanbieter aus Deutschland bieten den Abruf ausgestrahlter Sendungen über Videoportale[48] an. Darüber hinaus bestehen weitere kommerzielle Angebote von deutschen Anbietern.[49]

Die Angebote sind teilweise kostenfrei für den Nutzer, teilweise werden Filme gegen eine Leihgebühr oder Pauschalabonnements angeboten. Die oben genannten im Rahmen der vorliegenden Studie betrachteten Anbieter nutzen unterschiedliche Mittel, um den gesetzlichen Jugendmedienschutzbestimmungen nachzukommen. Anbieter, die eine Mitgliedschaft zur Nutzung von Angeboten voraussetzen, setzen beispielsweise auf Altersverifikationssysteme, während für einige frei verfügbare Angebote Sendezeitbeschränkungen angewandt werden. Die Information über bestehende Altersfreigaben erfolgt bei den verschiedenen Anbietern auf unterschiedliche Weise. Entweder werden bestehende FSK-Altersauszeichnungen bei der Beschreibung eines Films oder im Coverbild dargestellt oder der Anbieter gibt an, dass ein Film einer Sendezeitbeschränkung aufgrund einer Alterseinstufung unterliegt.

Die hier betrachteten Angebote stammen von deutschen Anbietern oder von Anbietern, die über deutsche Tochterunternehmen mit Kunden in Deutschland ein Vertragsverhältnis für die kostenpflichtige Nutzung von Online-Videoangeboten anstreben. Daher ist nachvollziehbar, dass diese Anbieter bemüht sind, den in Deutschland geltenden Jugendmedienschutzbestimmungen zu genügen. Ein anderes Bild zeichnet sich bei der Betrachtung von Angeboten, die nicht aus Deutschland stammen und die kein kommerzielles Interesse am deutschen Markt verfolgen, wie beispielsweise das Internet Archive[50]. In dessen frei verfügbarem Angebot sind auch Kinofilme enthalten, die in Deutschland von der FSK mit Altersfreigaben belegt sind.

[47] Videoportale der öffentlich-rechtlichen Rundfunkanstalten in Deutschland sind zum Beispiel *ARD Mediathek* (http://www.ardmediathek.de/, Stand 31.8.2012) und *ZDF Mediathek* (http://www.zdf.de/ZDFmediathek/, Stand 31.8.2012).

[48] Videoportale der privaten Fernsehanbieter in Deutschland sind zum Beispiel das Filmportal *RTL now* (http://rtl-now.rtl.de, Stand 31.8.2012). oder das von der SevenOne Intermedia GmbH übernommene rumänische Portal *Myvideo* (http://www.myvideo.de/, Stand 19.10.2012) .

[49] Videoportale von weiteren Anbietern aus Deutschland sind zum Beispiel *Bild* (http://film.bild.de, Stand 31.8.2012), *Lovefilm* (http://www.lovefilm.de/, Stand 19.10.2012), *Maxdome* (http://www.maxdome.de, Stand 31.8.2012) und die Angebote der Deutschen Telekom *Videoload* (http://www.videoload.de, Stand 31.8.2012) und *Erotic-Lounge* (http://www.erotic-lounge.com, Stand 31.8.2012).

[50] siehe http://archive.org, Stand 31.8.2012.

2.4.3.2 Altersauszeichnung von Onlinespielen

Onlinespiele sind Computerspiele, bei denen die Spieler über das Internet mit dem Spieleanbieter und gegebenenfalls mit anderen Spielern verbunden sind. Eine rechtliche Verpflichtung zur Alterskennzeichnung, wie sie im Sinne des JuSchG für die Verbreitung von Computerspielen auf Trägermedien vorgesehen ist, existiert – bislang – für online-Spiele nicht.[51]

Auch wenn keine gesetzliche Verpflichtung zur Auszeichnung von Onlinespielen besteht, so wurden doch einige Strukturen geschaffen, an denen sich Kinder, Jugendliche und Eltern und letztlich auch Jugendschutzsoftware und technische Verfahren zur Inhalteklassifikation orientieren können. Hier sind die Verpflichtung zur Übernahme der Kennzeichnung von Spielen auf Trägermedien für inhaltsgleiche Online-Angebote gemäß § 12 JMStV, die Aktivitäten der USK.online (vgl. Abschnitt 1.2.2.3) und PEGI *Online*[52] zu benennen. Rechtlich gesehen sind dabei in Deutschland die Kennzeichnungen der USK maßgeblich. Das PEGI-System (vgl. Abschnitt 2.1.2.2) bietet Eltern europaweit detaillierte Empfehlungen zur Eignung von Computerspielinhalten für Kinder und Jugendliche. Durch das PEGI-System wird eine differenzierte Alterseinstufung[53] für Computerspiele realisiert. PEGI Online ist eine Ergänzung des PEGI-Systems, die im Jahr 2007 gestartet wurde. Der Zweck besteht darin, Kinder und Jugendliche in Europa vor ungeeigneten Spielinhalten besser schützen zu können und Eltern das Verständnis für die Risiken und möglichen Gefahren innerhalb dieser Spielumgebung zu erleichtern. Da das PEGI-System zunächst für Spiele auf Trägermedien konzipiert wurde, ist es nicht für alle Arten von Online-Spielen vergleichbar geeignet, so ist zum Beispiel die Anwendung des Systems auf stark dynamische Browserspiele schwierig und für *kleine* Onlinespiele wegen der entstehenden Kosten wirtschaftlich unrentabel.

2.4.4 Beschwerdestellen für Inhalte im WWW

Wie bereits dargelegt (vgl. Abschnitt 1.2) existieren verschiedene Beschwerdestellen für Inhalte im WWW. Jede anerkannte Einrichtung der freiwilligen Selbstkontrolle ist gesetzlich verpflichtet, eine Beschwerdestelle zu betreiben. So verfügen FSK.online und USK.online über eigene Beschwerdestellen. Die FSM betreibt in Zusammenarbeit mit dem eco – Verband der deutschen Internetwirtschaft e. V. *die Internet-Beschwerdestelle für Deutschland*.[54] Zu den Aufgaben von jugendschutz.net, der im JMStV vorgesehenen »gemeinsamen Stelle Jugendschutz aller Länder« (vgl. § 18 Absatz 1 JMStV), die als operativer Arm der Medienaufsicht tätig ist, gehört ebenfalls der Betrieb einer eigenen Beschwerdestelle. Auch die KJM bietet ein entsprechendes Antragsformular für Beschwerden an. Darüber hinaus betreibt auch *naiin – no abuse in internet e. V.*,[55] ein im Jahr 2000 gegründeter Verein von Vertretern der Zivilgesellschaft, Internet-Wirtschaft und Politik, eine Beschwerdestelle zur Verfolgung illegaler Inhalte oder rechtswidriger Aktivitäten im WWW.

Die Beschwerdestellen prüfen eingehende Beschwerden und versuchen Verstöße gegen den Jugendmedienschutz i. d. R. zunächst durch Kontaktaufnahme mit den Inhalte-Anbietern oder den Hostern der Angebote zu lösen. Falls dies nicht erfolgreich ist, wird die zuständige Medienaufsicht eingeschaltet, im Falle von rechtswidrigen Inhalten werden auch direkt die Strafverfolgungsbehörden hinzugezogen.

[51] Ausgenommen sind Spiele, »die ganz oder im Wesentlichen inhaltsgleich sind mit […] Spielen programmierten Datenträgern, die nach § 12 des Jugendschutzgesetzes gekennzeichnet […] sind«, § 12, JMStV.
[52] siehe http://www.pegionline.eu/de/index/id/34, Stand 17.10.2012
[53] Altersgruppen nach PEGI: ab 3, ab 7, ab 12, ab 16, ab 18 Jahren; Quelle: [PEGI, 2011]
[54] siehe http://www.internet-beschwerdestelle.de/, Stand 17.10.2012
[55] siehe http://www.beschwerdestelle.de/, Stand 17.10.2012

Beschwerdestellen sind als Jugendschutzstrukturen im Internet zu betrachten, da dort Datenbanken (oder Listen) überprüfter Verstöße vorliegen. Beschwerdestellen können von technischen Verfahren profitieren, die den personellen Aufwand der manuellen Prüfung von Beschwerden reduzieren. Es ist demnach zu prüfen, ob automatische Verfahren gewinnbringend in den Arbeitsablauf der Beschwerdestellen integriert werden können (z. B. die Wiedererkennung bereits bekannter jugendschutzrelevanter Medien wie Videos oder extremistischer Äußerungen in Musik).

2.4.5 Zusammenfassung: Anforderungen an technische Verfahren im Kontext der technischen Jugendschutzstrukturen

Die beschriebenen Beispiele für technische Jugendschutzstrukturen zeigen eine große Bandbreite von Ansätzen auf, die häufig durch die Art der Angebote geprägt ist. Durch das Zusammenspiel verschiedener Strukturkomponenten kann der Jugendmedienschutz profitieren, da jede Strukturkomponente für sich genommen nur einen kleinen Teil des Jugendmedienschutzes für das WWW abdeckt.

Relevant ist vor allem die Frage der Integration verschiedener Komponenten sowohl untereinander als auch mit Jugendschutzsoftware. Technische Verfahren können die Effektivität einiger Strukturkomponenten fördern. Daher sind auch aus Sicht der Jugendschutzstrukturen bei der weiteren Betrachtung technischer Verfahren auch immer die Anforderungen an die Integrierbarkeit (vgl. Abschnitt 2.3.3) zu berücksichtigen.

2.5 Zusammenfassung

Die Betrachtung der Anforderungen an technische Verfahren zur Detektion von jugendschutzrelevanten Web-Inhalten zeigt einige besondere technologische Herausforderungen auf. Die Anforderungen an die Funktionalität der Verfahren können übersichtlich in nur zwei Anforderungen formuliert werden: jugendschutzrelevante Inhalte im WWW sollen 1. detektiert und 2. altersdifferenziert klassifiziert werden. Die hinter den beiden Anforderungen stehende inhaltliche und mediale Breite stellt die eigentliche Herausforderung dar, denn für jedes einzelne Gefährdungsszenario und jeden Medientyp müssen potentiell sehr unterschiedliche Verfahren eingesetzt werden (vgl. Kapitel 4).

Daneben können die beschriebenen nicht-funktionalen Anforderungen zur Beurteilung technischer Verfahren hinsichtlich ihrer möglichen Verwendung in Jugendschutzprogrammen oder Jugendschutzstrukturen eingesetzt werden. Hierbei ist insbesondere zu prüfen, ob Verfahren echtzeitfähig sind und performant auf Endgeräten des Nutzers oder eher in technischen Jugendschutzstrukturen des Internets zum Einsatz kommen können. Auch die Integrierbarkeit in bestehende Systeme und Strukturen kann entlang der beschriebenen Anforderungen überprüft werden (vgl. Kapitel 6).

KAPITEL 3

CHARAKTERISIERUNG JUGENDSCHUTZRELEVANTER WEB-INHALTE

Ziel dieses Kapitels ist eine möglichst umfassende Charakterisierung von jugendschutzrelevanten Web-Inhalten. Auf dieser Basis werden dann in den folgenden Kapiteln Verfahren ausgewählt und bewertet, die die entsprechenden Charakteristika automatisch detektieren können.

Zunächst werden die bereits existierenden Kriterienkataloge zur redaktionellen Bewertung von jugendschutzrelevanten Webseiten kurz beschrieben. Anschließend werden jeweils für die Themen[1] Gewalt, Extremismus und Selbstgefährdung Definitionen und spezifische Kriterien zur Erkennung und Alterseinstufung angegeben, die sich aus den Kriterienkatalogen ergeben.

Auf Basis der beschriebenen Kriterienkataloge zur redaktionellen Klassifikation von jugendschutzrelevanten Web-Inhalten wurde dem Fraunhofer IAIS von jugendschutz.net eine exemplarische Menge von Webseiten und Videos aus den drei beschriebenen Inhaltsbereichen zur Verfügung gestellt. In den einzelnen Abschnitten zu den jeweiligen Bereichen werden die spezifischen Charakteristika dieser Daten dargestellt. Hier ist von besonderem Interesse, ob sich wiederkehrende Eigenschaften der Inhalte erkennen lassen, die dann durch ein automatisches Verfahren identifiziert werden können.

3.1 Bestehende Kriterienkataloge zur redaktionellen Inhalteklassifikation

Für die redaktionelle Klassifikation von Webseiten liegen heute Kriterienkataloge vor, die jugendschutzrelevante Web-Inhalte charakterisieren. Die Kataloge der verschiedenen Einrichtungen (vgl. Abschnitt 1.2.2) werden in den folgenden Abschnitten beschrieben. Desweiteren wird auf das Buch »Schutzgrade im Jugendmedienschutz« [Liesching, 2011] eingegangen, das ausführliche Informationen zu den Kriterien im Jugendschutz bietet. Einen Überblick gibt die Tabelle 3.1.

3.1.1 Kommission für Jugendmedienschutz der Landesmedienanstalten (KJM)

Die KJM unterscheidet in den *Kriterien für die Aufsicht im Rundfunk und in den Telemedien* [KJM, 2010] unter »Entwicklungsbeeinträchtigung und Entwicklungsgefährdung« die drei Bereiche »Gewalt«, »Sexualität« und »Eigenverantwortung und Gemeinschaftsfähigkeit«. Unter letztere fällt auch strukturelle Gewalt, politischer Extremismus und Selbstgefährdung, die unter »Risikoverhalten: physische und psychische Integrität« aufgeführt wird.

Im Bereich Selbstgefährdung existieren Beurteilungsraster bei jugendschutz.net, die auf der Spruchpraxis der KJM basieren (zum Beispiel im Bereich von Pro-Anorexie-Angeboten [Rauchfuß, 2008]).

[1] Erläuterung zur Themenauswahl siehe Vorwort

Tabelle 3.1
Bestehende Kriterienkataloge zur redaktionellen Inhalteklassifikation.

Quelle	Gewalt	Extremismus	Selbstgefährdung
KJM	Kriterienkatalog mit Unterscheidung von Entwicklungsbeeinträchtigung und Entwicklungsgefährdung		Beurteilungsraster Pro-Ana; Beurteilungsraster Suizid (nicht öffentlich)
FSM	Fragebogen und Prüfgrundsätze mit Kriterien; Unterscheidung von entwicklungsbeeinträchtigend und offensichtlich schwer entwicklungsgefährdend		
Liesching	gesetzliche Grundlage, Rechtsprechung und Einzelfälle; Unterscheidung von 6 Gefährdungsstufen		
PEGI, USK, FSK, FSF	Schwerpunkt bei der Bewertung liegt auf Gewalt; keine detaillierten Kriterien öffentlich		

3.1.2 Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e. V. (FSM)

3.1.2.1 Prüfgrundsätze

In den *Prüfgrundsätzen der FSM* [FSM, 2011d] (vgl. Abschnitt 1.2.2.3) werden – neben den allgemeinen Kriterien zur Bewertung eines Angebots (Domain, Werbung, Pop-ups, Links, Web 2.0 usw.) und den Kriterien zur Bewertung von unzulässigen Angeboten – Kriterien zur Bewertung von *entwicklungsbeeinträchtigenden* und *offensichtlich schwer entwicklungsgefährdenden* (vgl. Abschnitt 1.1.4.2) Angeboten detailliert dargestellt (vgl. Prüfmatrix der FSM: Tabelle 3.2).

Unter die Kriterien für schwer entwicklungsgefährdende Angebote fallen im Bereich Gewalt die »Darstellungen von Gewalttätigkeiten, die zur Nachahmung anreizen, die selbstzweckhaft oder besonders grausam sind oder die verrohend wirken« und die »Befürwortung von Gewalt zur Durchsetzung sexueller Interessen«. Im Bereich Extremismus werden hier »Darstellungen im Sinne demokratiefeindlicher, rassistischer, völkischer oder nationalistischer Ideologien« und das »Beschimpfen von Glaubensbekenntnissen, Religionsgemeinschaften oder Weltanschauungsvereinigungen« aufgeführt. Detailliert sind in den Prüfgrundsätzen die im Sinne der vorliegenden Studie in den Bereich Selbstgefährdung fallenden Kriterien für schwer entwicklungsgefährdende Angebote beschrieben: »Verführung zum Erwerb oder Gebrauch von Suchtmitteln«, »Aufforderung zu Straftaten, die von § 130a StGB nicht erfasst werden« und »Aufruf zum Suizid, zur Selbstverletzung oder zur Selbstgefährdung«.

Die Kriterien für entwicklungsbeeinträchtigende Angebote werden in Gewalt, Sexualität, Extremismus und sozial-ethisch desorientierende Angebote unterteilt. Unter letztere fallen unter anderem die Kriterien zum Thema Selbstgefährdung.

3.1.2.2 Fragebogen zur Altersklassifizierung

Der Fragebogen der FSM zur Altersklassifizierung[2] von Internet-Angeboten bietet dem Anbieter des Angebotes die zumindest derzeit kostenfreie Möglichkeit einer Selbsteinschätzung. Am Ende der Abfrage erhält der Anbieter eine Datei (age-de.xml) zur Installation auf seinem Webserver, aus der Jugendschutzsoftware die Alterskennzeichnung auslesen kann. Der Fragenkatalog umfasst zahlrei-

[2] siehe https://www.altersklassifizierung.de/, Stand 17.10.2012

Tabelle 3.2
Prüfmatrix der FSM [FSM, 2011d]

	Physiologisch-körperliche Entwicklung	Sexuelle Entwicklung	Identitätsbildung/Selbstfindung	Soziale Entwicklung/ Politische Sozialisation	Moralische/religiöse Entwicklung	Sonstige Entwicklungsbereiche	Gesamtergebnis der Prüfung
Angebot: Gefährdungspotential							
- präsentierter Inhalt (z. B. Gewalt, Sexualität, Extremismus)							
- Art bzw. Ziel und Anliegen des Angebots							
- Darstellungsform							
- gestalterische Mittel							
- Jugendaffinität/Nutzerschaft des Angebots							
Nutzer: Gefährdungsneigung							
Durchschnittliche Jugendliche bzw. Risikogruppe							
- Alter							
- Geschlecht							
- Entwicklungsstand							
- Familiäres Anregungsmilieu							
- Persönliche Medienbindung							
- Vorerfahrungen bzgl. des Themas							
- Rezeptionsweisen (Involvement, Identifikation mit Inhalten)							
Resultat der Prüfung i. S. d. JMS							

Tabelle 3.3
Schutzstufen nach Liesching

Stufe	Beschreibung
6	offensichtlich schwer jugendgefährdende Inhalte
5	einfach jugendgefährdende Inhalte
4	entwicklungsbeeinträchtigende Inhalte »ab 18«
3	entwicklungsbeeinträchtigende Inhalte »ab 16«
2	entwicklungsbeeinträchtigende Inhalte »ab 12«
1	entwicklungsbeeinträchtigende Inhalte »ab 0«/»ab 6«

che Fragen zu Sex und Erotik sowie zu Gewalt und bedrohlichen oder ängstigenden Inhalten. Bei den Fragen bezüglich Gewalt wird nicht nur abgefragt, *was* dargestellt wird, sondern zwischen verschiedenen Darstellungsarten (vier Stufen von real bis extrem abstrahiert) differenziert. Es gibt weiterhin einige Fragen zu politischem Extremismus und Selbstgefährdung. Der Fragebogen zeigt jeweils nach dem Ausfüllen einer Seite an, welche Alterseinstufung sich aus den bisher ausgefüllten Fragen ergibt.

3.1.3 Liesching: Schutzgrade im Jugendmedienschutz

Der Rechtsanwalt Dr. Marc Liesching [Liesching, 2011] betrachtet den Jugendmedienschutz aus juristischer Perspektive. Bezüglich der Jugendgefährdung definiert er sechs Gefährdungsstufen (vgl. Tabelle 3.3). Viele der themenspezifischen Begriffe, die in der Literatur häufig verwendet werden, werden in dem Buch definiert. Es werden verschiedene Interpretationen gegenübergestellt, so zum Beispiel die Begriffe »Gefährdungsgeneigtheit« und »durchschnittliche Minderjährige«, die bei der Bewertung von Internet-Angeboten im Hinblick auf die potentiellen Nutzer eine Rolle spielen.

Für jede Gefährdungsstufe führt Liesching die gesetzlichen Grundlagen auf, beleuchtet die Rechtsprechungen und Rechtsliteratur, die Richtlinien, Vorgaben und Spruchpraxis der Jugendschutzinstitutionen und fügt eine systematische Konsolidierung der Auslegungsgrundsätze an. Im Rahmen der Rechtsprechungen werden zahlreiche Einzelfälle aufgeführt, die für die Bewertung von weiteren Internet-Inhalten als Kriterien herangezogen werden könnten.

3.1.4 Kriterienkataloge vergleichbarer Medien

Neben den Kriterienkatalogen zur Bewertung von Internet-Angeboten ist es sinnvoll, einen Blick auf die Vorgehensweisen bei der Bewertung ähnlicher Medien wie Computerspiele, Fernsehen und Kino zu werfen.

Für Filme und andere Trägermedien, die in Deutschland für die öffentliche Vorführung und Verbreitung vorgesehen sind, wird die FSK-Kennzeichnung benutzt. Desweiteren gibt es verschiedene Positivlisten des Kinder- und Jugendfilmzentrums in Deutschland (KJF).[3]

Kennzeichnend für Filme im Gegensatz zu klassischen Webseiten ist die fortlaufende Handlung mit Spannungsauf- und abbau. Es findet keine Interaktion mit dem Nutzer statt. Filme in Form von Werbe-Trailern sind häufig Elemente von Internet-Angeboten.

[3] Dies sind zum Beispiel http://www.top-videonews.de und http://www.kinderfilmwelt.de (Stand 14.11.2012).

Für Computerspiele gibt es die Altersklassifikation nach USK und PEGI. Während die von PEGI angegebene Altersgruppe lediglich eine Empfehlung für Eltern hinsichtlich eines Mindestalters darstellt, ist die USK-Altersstufe in Deutschland rechtlich bindend. Das besondere Kennzeichen von Spielen ist die Interaktion. Computerspiele werden häufig in jugendaffinen Internet-Angeboten mittels Gratisspielen (als Online-Spiel oder als Download) beworben.

3.1.4.1 Unterhaltungssoftware Selbstkontrolle (USK)

Die USK – und für den Bereich des Internets die USK.online – (vgl. Abschnitte 1.2.2.3 und 2.1.2.1) bewertet Computerspiele, die auf PCs, Konsolen, mobilen Endgeräte oder im Internet spielbar sind. Bei der Jugendschutz-Bewertung ist die Darstellung der Gewalt (auch Kriegsdarstellung) ein Schwerpunkt.

Die besonderen Herausforderungen bei der Bewertung von Spielen sind die Interaktivität, das Selbstwirksamkeitserleben (Welche Rolle nimmt der Spieler ein? Welchem Handlungsdruck ist er ausgesetzt? Gibt es gewaltfreie Lösungen? usw.) und der Immersionsgrad (Wie weit taucht der Spieler in eine virtuelle Welt ein?).

3.1.4.2 Pan European Game Information (PEGI)

PEGI hat einen eigenen, kostenpflichtigen Prozess aufgesetzt, den Anbieter zur Altersklassifizierung von Computerspielen nutzen können (vgl. Abschnitt 2.1.2.2). Die PEGI überprüft die Spiele auf Gewalt, Sex, Diskriminierung und andere sensitive Bilder und Töne (»other sensitive image oder sound content« [PEGI, 2011]). Die genauen Kriterien sind nicht öffentlich. Der Anbieter erhält am Ende des Prozesses die Berechtigung, ein entsprechendes »age rating logo« an das Spiel anzubringen.

3.1.4.3 Freiwillige Selbstkontrolle der Filmwirtschaft (FSK)

Die FSK (vgl. Abschnitte 1.2.2.3 und 2.1.2.1) arbeitet auf der Basis der im Grundgesetz geschützten Werte, des Jugendschutzgesetzes. Für den Bereich des Internets ist die FSK.online zuständig und der JMStV rechtlich maßgeblich. Bezüglich der angewandten Kriterien heißt es in § 18 Absatz 2 Punkt 3 der Grundsätze der FSK [FSK, 2010] :

»Die Entwicklung von Kindern und Jugendlichen oder ihre Erziehung zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit können insbesondere Filme oder Trägermedien beeinträchtigen, welche die Nerven überreizen, übermäßige Belastungen hervorrufen, die Phantasie über Gebühr erregen, die charakterliche, sittliche (einschließlich religiöse) oder geistige Erziehung hemmen, stören oder schädigen oder zu falschen und abträglichen Lebenserwartungen verführen.«

In den Erläuterungen zu den Alterseinstufungen der FSK [FSK, 2012a] wird deutlich, dass, je niedriger die Altersstufe ist, desto mehr Relevanz die Frage nach der Identifikation mit den Akteuren der Filme hat. Hier geht es um ängstigende, bedrohende, konflikthaltige und auch gewalttätige Inhalte. Gewalt spielt bei allen Alterseinstufungen eine Rolle, wobei die entwicklungspsychologischen Fähigkeiten, mit den Inhalten umzugehen, ausschlaggebend für die Alterseinstufung ist.

3.1.4.4 Freiwillige Selbstkontrolle Fernsehen (FSF)

Ziel der FSF ist die Umsetzung der Jugendschutzanforderungen für das Fernsehangebot privater Fernsehanbieter in Deutschland. Die Mitgliedschaft in der FSF ist freiwillig. »Geprüft wird das Fernsehprogramm insbesondere hinsichtlich des Gehalts an Gewalthandlungen und sexuellen Darstellungen, von dem die Altersfreigabe und damit die zulässige Sendezeit im Tages-, Abend-, Spätabend- oder Nachtprogramm abhängen.« [FSF, 2012b]

3.2 Charakterisierung von Web-Inhalten im Bereich Gewalt

Im folgenden Abschnitt wird zunächst aufgeführt, was im Allgemeinen und im Kontext des Jugendmedienschutzes unter Gewalt verstanden wird. Anschließend werden die Kriterien zusammengestellt, die in oben genannten Kriterienkatalogen (vgl. Abschnitt 3.1) zur Identifikation von Gewalt und zur Altersklassifikation angegeben werden. Im dritten Teil werden die Daten beschrieben, die dieser Studie als Beispiele für jugendschutzrelevante Webseiten in Bezug auf Gewalt dienen.

3.2.1 Definitionen

Zur Einführung in das Thema sind im Folgenden die verschiedenen Definitionen zu Gewalt beziehungsweise *Gewaltdarstellungen* niedergeschrieben.

Duden

1. »Macht, Befugnis, das Recht und die Mittel, über jemanden, etwas zu bestimmen, zu herrschen«
   (a) »unrechtmäßiges Vorgehen, wodurch jemand zu etwas gezwungen wird«
   (b) »[gegen jemanden, etwas rücksichtslos angewendete] physische Kraft, mit der etwas erreicht wird«
2. »(gehoben) elementare Kraft von zwingender Wirkung«[4]

Im Kontext der Studie sind die Definitionen 1a und 1b relevant.

Wikipedia »Unter den Begriff Gewalt (von althochdeutsch *waltan* – stark sein, beherrschen) fallen Handlungen, Vorgänge und Szenarien, in denen bzw. durch die auf Menschen, Tiere oder Gegenstände beeinflussend, verändernd und/oder schädigend eingewirkt wird. Gemeint ist das Vermögen zur Durchführung einer Handlung, die den inneren bzw. wesentlichen Kern einer Angelegenheit oder Struktur (be)trifft.«[5]

§ 131 Strafgesetzbuch »§ 131 Gewaltdarstellung

(1) Wer Schriften (§ 11 Absatz 3), die grausame oder sonst unmenschliche Gewalttätigkeiten gegen Menschen oder menschenähnliche Wesen in einer Art schildern, die eine Verherrlichung oder Verharmlosung solcher Gewalttätigkeiten ausdrückt oder die das Grausame oder Unmenschliche des Vorgangs in einer die Menschenwürde verletzenden Weise darstellt,
   1. verbreitet,
   2. öffentlich ausstellt, anschlägt, vorführt oder sonst zugänglich macht,

[4] siehe http://www.duden.de/rechtschreibung/Gewalt, Stand 26.9.2012
[5] siehe http://de.wikipedia.org/wiki/Gewalt, Stand 26.9.2012

3. einer Person unter achtzehn Jahren anbietet, überlässt oder zugänglich macht oder
4. herstellt, bezieht, liefert, vorrätig hält, anbietet, ankündigt, anpreist, einzuführen oder auszuführen unternimmt, um sie oder aus ihnen gewonnene Stücke im Sinne der Nummern 1 bis 3 zu verwenden oder einem anderen eine solche Verwendung zu ermöglichen,

wird mit Freiheitsstrafe bis zu einem Jahr oder mit Geldstrafe bestraft.

(2) Ebenso wird bestraft, wer eine Darbietung des in Absatz 1 bezeichneten Inhalts durch Rundfunk, Medien- oder Teledienste verbreitet.

(3) Die Absätze 1 und 2 gelten nicht, wenn die Handlung der Berichterstattung über Vorgänge des Zeitgeschehens oder der Geschichte dient.

(4) Absatz 1 Nummer 3 ist nicht anzuwenden, wenn der zur Sorge für die Person Berechtigte handelt; dies gilt nicht, wenn der Sorgeberechtigte durch das Anbieten, Überlassen oder Zugänglichmachen seine Erziehungspflicht gröblich verletzt.«

Jugendschutzgesetz Nach § 15 JuSchG werden im Kontext von Gewalt als schwer jugendgefährdend Trägermedien eingestuft die,

»[…]

3. Menschen, die sterben oder schweren körperlichen oder seelischen Leiden ausgesetzt sind oder waren, in einer die Menschenwürde verletzenden Weise darstellen und ein tatsächliches Geschehen wiedergeben, ohne dass ein überwiegendes berechtigtes Interesse gerade an dieser Form der Berichterstattung vorliegt,
   (a) besonders realistische, grausame und reißerische Darstellungen selbstzweckhafter Gewalt beinhalten, die das Geschehen beherrschen,

[…]«

Das JuSchG nennt als Kriterien für jugendgefährdende Angebote:

»[…] Dazu zählen vor allem unsittliche, verrohend wirkende, zu Gewalttätigkeit, Verbrechen oder Rassenhass anreizende Medien sowie Medien, in denen

1. Gewalthandlungen wie Mord- und Metzelszenen selbstzweckhaft und detailliert dargestellt werden oder
2. Selbstjustiz als einzig bewährtes Mittel zur Durchsetzung der vermeintlichen Gerechtigkeit nahe gelegt wird.«

Jugendmedienschutz-Staatsvertrag Im JMStV werden die Kategorien »schwer jugendgefährdend« beziehungsweise »jugendgefährdend« nicht eigenständig verwendet. Bezüglich Gewalt werden die folgenden Kriterien aufgestellt, die unter die absolute Verbotsnorm des § 4 Absatz 1 JMStV fallen:

»[…]

5. grausame oder sonst unmenschliche Gewalttätigkeiten gegen Menschen in einer Art schildern, die eine Verherrlichung oder Verharmlosung solcher Gewalttätigkeiten ausdrückt oder die das Grausame oder Unmenschliche des Vorgangs in einer die Menschenwürde verletzenden Weise darstellt; dies gilt auch bei virtuellen Darstellungen,

[…]

8. gegen die Menschenwürde verstoßen, insbesondere durch die Darstellung von Menschen, die sterben oder schweren körperlichen oder seelischen Leiden ausgesetzt sind oder waren, wobei ein tatsächliches Geschehen wiedergegeben wird, ohne dass ein berechtigtes Interesse gerade für diese Form der Darstellung oder Berichterstattung vorliegt; eine Einwilligung ist unbeachtlich,

[…]«

Kommission für Jugendmedienschutz der Landesmedienanstalten (KJM) »Gewalt ist die physische, psychische oder materielle Schädigung von personalen und dinglichen Objekten durch Subjekte, Naturgewalten oder Unfälle.« [KJM, 2010]

Freiwillige Selbstkontrolle Multimedia-Diensteanbieter (FSM) »Gewaltdarstellungen sind mediale Angebote, bei denen eine physische, psychische oder materielle Schädigung von Personen und Tieren sowie dinglichen Objekten dargestellt ist.« [FSM, 2011d]

3.2.2 Kriterien und Altersklassifikation

Der Fragebogen[6] der FSM zur Selbsteinschätzung von Webseiten-Anbietern unterteilt Gewalt in

- Gewalt gegen Menschen oder menschenähnliche Wesen (dem Erscheinungsbild und der Gewaltreaktion nach Menschen weitgehend gleichende Wesen)
  - Vergewaltigung,
  - andere sexuelle Übergriffe,
  - Foltern,
  - Töten,
  - wesentliche Verstümmelung,
  - Blutvergießen,
  - Verletzen.
- Gewalt gegen Tiere (oder dem Erscheinungsbild und der Gewaltreaktion nach Tieren gleichende Wesen)
  - Vergewaltigung,
  - andere sexuelle Übergriffe,
  - Foltern,
  - Töten,
  - wesentliche Verstümmelung,
  - Blutvergießen,
  - Verletzen.
- Gewalt gegen Fantasiefiguren
  - Foltern,
  - Töten,
  - wesentliche Verstümmelung,
  - Blutvergießen,
  - Verletzen.

Neben diesen Kategorien, die der personellen Gewalt (ungleiche Macht zwischen Ausübenden und Betroffenen) (vgl. [FSM, 2011d]) zuzuordnen sind, muss auch die strukturelle Gewalt (ungleiche Herrschafts- und Machtverhältnisse im gesellschaftlichen System) betrachtet werden.

[6] siehe https://www.altersklassifizierung.de/, Stand 17.10.2012

In allen Fällen spielt neben der Tatsache, dass gewaltbezogene Inhalte dargestellt werden, außerdem eine Rolle, wie realistisch die Darstellungen sind. Der oben genannte Fragebogen unterscheidet vier Stufen von »real«, »real wirkend«, »nicht oder nur leicht abstrahiert bzw. verfremdet« bis »extrem abstrahiert bzw. verfremdet«. Laut [KJM, 2010] sind insbesondere Angebote problematisch, »die fiktional sind, aber Realität suggerieren und […] die real sind […], aber Darstellungstechniken fiktionaler Formate […] integrieren […].«

Um festzulegen, welche Kategorie von Gewalt in welcher Ausprägung zu welcher Altersgruppe führt, muss außerdem der Gesamtzusammenhang betrachtet werden. Mit dem Erzählstrang (von Videos, Spielen aber auch Textbeiträgen) wird häufig eine Spannung aufgebaut. Die Höhe der Spannung und das Anhalten sowie das Auftreten entspannender Elemente beziehungsweise des Spannungsabbaus haben beispielsweise eine Auswirkung auf die Alterseinstufung. Ebenso müssen die Erwartungen an das Angebot berücksichtigt werden, der Kontext, die Gesamthandlung, die Folgen der Gewalt und die Möglichkeit der Rezipienten, sich mit den ausführenden Personen/Figuren zu identifizieren.

In den Prüfgrundsätzen der FSM [FSM, 2011d] werden Gewaltdarstellungen, die als schwer entwicklungsfährdend eingestuft werden, wie folgt beschrieben:

»Darstellungen von Gewalttätigkeiten, die zur Nachahmung anreizen, die selbstzweckhaft oder besonders grausam sind oder die verrohend wirken. Dies trifft auch auf Inhalte zu, die unter der Schwelle von § 131 StGB bleiben, da diese ja ohnehin unzulässig sind. Hier sind auch Eigenberichte über selbst vollzogene Gewalthandlungen einzuordnen oder Darstellungen, die die Gewalthandlung als förderlich, lustbringend, amüsant, harmlos oder ohne negative Konsequenzen für den Gewalttäter darstellen. Selbstzweckhaft ist eine Darstellung von Gewalttätigkeit dann, wenn sie ohne eine über die eigentliche Tat hinausgehende Informationsvermittlung auskommt und die Tat nicht nur im Mittelpunkt der Darstellung steht, sondern die Tat das Gesamtbild der Darstellung einnimmt. Um besonders grausame Darstellungen handelt es sich bei der exzessiven Darstellung von Schmerzen, Leid und Ängsten, bei der detaillierten Schilderung der Reaktionen oder des Verhaltens des menschlichen Körpers auf Gewalteinwirkungen (z. B. Beschreibungen, wie die Knochen brechen, das Messer in den Körper eindringt, usw.). Auch die besonders grausame Darstellung der Folgen von Gewalttätigkeiten ist hier zu erwähnen, zum Beispiel bei der Darstellung von geöffneten Körpern nach Gewalteinwirkungen. Darstellungen realen Geschehens sind in diesem Bereich deutlich schärfer zu beurteilen, als Darstellungen, die als deutlich fiktional zu erkennen sind, da reales Geschehen auf die Rezipienten wesentlich höheres Wirkungspotenzial besitzt und die Distanzierungsmöglichkeiten geringer sind. Verrohend sind Darstellungen dann, wenn sie eine zynische oder das Opfer verhöhnende Haltung einnehmen oder wenn sie die Tat derart verharmlosen, dass von einer verrohenden Wirkung ausgegangen werden kann.«

In den Prüfgrundsätzen der FSM [FSM, 2011d, Absatz 9.2.2.1] werden die Kriterien zur Einordnung einer Gewaltdarstellung als entwicklungsbeeinträchtigend ausführlich und zum Teil mit beispielhaften Angaben von Altersgruppen angegeben. Es werden inhaltlich-formale Kriterien angelegt wie beispielsweise der Bezug der Gewaltdarstellung zur »Realität und zur Lebenswelt der Kinder und Jugendlichen«. Weiterhin muss betrachtet werden, wie die Gewalt in die Gesamthandlung eingebettet ist und ob beziehungsweise wie die Folgen der Gewalt dargestellt werden. Nicht zuletzt ist zu prüfen, in wie weit der Rezipient sich mit den ausführenden Personen/Figuren identifizieren kann, wie er aktiv am Geschehen teilnehmen kann usw.

Alterseinstufung Um bei der Bewertung von gewalthaltigen Web-Inhalten eine Alterseinstufung vorzunehmen, muss die Persönlichkeitsentwicklung von Kindern und Jugendlichen berücksichtigt werden. Besonders drastische Gewaltdarstellungen können bei Kindern beispielsweise zu Ängstigungen führen, während bei Jugendlichen die Gefahr einer sozialethischen Desorientierung besteht.

Sowohl der Bezug zur Lebenswelt von Kindern und Jugendlichen als auch die Identifikationsmöglichkeiten, die die mitwirkenden Figuren bieten, sind altersspezifisch zu berücksichtigen.

Kein Kriterienkatalog stellt eine direkte Verbindung von Kriterien zu den vorgeschriebenen Altersstufen (siehe 2.1.2.1) her. In den Prüfgrundsätzen der FSM [FSM, 2011d] wird an verschiedenen Stellen auf die »Altersgruppe unter 14 Jahren« beziehungsweise »über 14 Jahren« verwiesen. In [KJM, 2005] wird bezüglich der rezipientenspezifischen Wirkungsfaktoren die Alterseinteilung »Kinder im Vorschulalter«, »Kinder im Grundschulalter (6-10J)« und »Kinder ab 12 Jahren« vorgenommen.

[Liesching, 2011] nennt als Beispiele für schwer jugendgefährdende, im Bereich Gewalt angesiedelte Angebote:

- »die Verfolgung Dritter in einem vollbesetzten Biergarten mit einer Kettensäge,«
- »regellose Boxkämpfe mit verbundenen Augen (›blind boxing‹),«
- »das Aufdrücken eines Brandeisens auf den Körper eines Gefesselten.«

3.2.3 Beschreibung der exemplarischen Daten im Bereich Gewalt

Von jugendschutz.net wurden für diese Studie Webseiten und Videos zur Verfügung gestellt, deren Inhalt Gewalt explizit darstellt oder die Anwendung von Gewalt verherrlicht. Die meisten der Webseiten enthalten Bilder oder Videos mit expliziten Gewaltdarstellungen.

3.2.3.1 Webseiten

Die uns zur Verfügung stehenden Webseiten im Bereich Gewalt lassen sich grob in die Bereiche *reale Inhalte*, *fiktive Inhalte*[7], *Computerspiele* und *Musik* untergliedern. Auf den Gewalt-Webseiten spielen multimediale Inhalte (Bilder und Videos) eine besonders wichtige Rolle, gewaltverherrlichende Texte sind seltener und haben eine weniger drastische Wirkung auf den Leser als das Betrachten eines Videos, in dem exzessive Gewalt explizit dargestellt wird.

Reale Inhalte

- Informations-Webseiten (z. B. alternative Nachrichtenportale), die in ihren Nachrichten-Artikeln explizite Bilder einsetzen (z. B. Fotos und Berichte von oder über Kriegsopfer),
- Schockbilder als Einleitung zur Argumentation, zum Beispiel gegen Abtreibung oder Tierfolter,
- Webseiten, die »nur« dem Zweck dienen, besonders abstoßende/drastische Bilder zu Unterhaltungszwecken darzustellen.

Fiktive (aber real wirkende) Inhalte

- verstörende Kunst, also zum Beispiel Skulpturen von toten (Fantasie-)tieren, etwa sterbenden Einhörnern oder Hirschen mit aufgerissenem Unterleib. Sammlung von Köpfungsdarstellungen in (z. T. echten) Fotografien oder zeitgenössischen/antiken Kunstgemälden,

[7] Aus rechtlicher Perspektive besteht kein Unterschied zwischen realer und fiktiver Gewalt.

- Spielfilme, wobei hierunter sowohl Trailer der Filme selbst fallen als auch allgemeine Filmdatenbanken[8] oder themenbasierte Datenbanken, die sich zum Beispiel auf Slasherfilme[9] spezialisieren.

Computerspiele Hier sind sowohl Trailer, Ankündigungen oder sonstige Angebote der jeweiligen Hersteller selbst enthalten (siehe auch Abschnitt 3.2.3.2), aber auch Browsergames/Flashfilme etc.

Musik Hier finden sich neben den zum Teil besonders drastischen Liedtexten auch die offiziellen Webangebote der Bands oder Sammlungen von einschlägigen CD-Covern auf zum Beispiel Black-Metal-Foren.

3.2.3.2 Videos

Die vorliegenden Videos lassen sich grob in die drei Kategorien *reale Videos*, *Spielfilme* und *Computerspiele* einteilen.

Reale Videos Reale Videos stellen wahre Begebenheiten dar oder sind von ihnen nicht zu unterscheiden. Oft ist die Aufnahmequalität sehr schlecht, da sie mit einer Handykamera oder einer Handkamera aufgenommen zu sein scheinen. Die Ausleuchtung ist schlecht, insbesondere bei Abend- oder Nachtaufnahmen. Bei Mitschnitten von Überwachungskameras kann die Darstellung auch komplett schwarz-weiß sein. Der Inhalt der Videos ist sehr heterogen und reicht von Prügeleien gegen Wehrlose bis hin zur detailliert dargestellten Enthauptung. Scheint die Situation auch für den Benutzer der Kamera gefährlich zu sein, etwa bei einer Massenschlägerei, ist die Kameraführung entsprechend hektisch und verwackelt. Bei rituellen oder anderweitig vorbereiteten Handlungen wird hingegen mitunter auch eine Standkamera genutzt. Die Audioqualität ist entsprechend der Situation ähnlich durchwachsen. Bei manchen Videos wurde aggressive Musik hinterlegt und das ursprüngliche Tonmaterial verworfen. Manche Videos sind eine Zusammenstellung von anderen, eventuell ursprünglich längeren Videos. Oft wird hierbei ein Logo der Webseite beziehungsweise der Organisation eingeblendet, die sich für den Zusammenschnitt verantwortlich zeigt. Bei einzelnen Videos wird ähnlich einem Wasserzeichen eine Webadresse angegeben.

Spielfilme Ebenfalls als sehr heterogen zu bewerten ist das vorliegende Material zu Gewalt in Spielfilmen. Manche dieser Filme sind indiziert, einige davon zusätzlich in Deutschland beschlagnahmt. Wenn auch die meisten dem Genre Horror zuzuordnen sind, so reicht die Bandbreite von Monsterfilmen bis hin zu sexuell-sadistischen Rachefeldzügen. Dennoch ist die Gewalt nicht immer unmittelbar dargestellt. Auch die Länge der Videos variiert stark. Manchmal sind es nur entweder kurze Ausschnitte aus einzelnen Szenen, Filmtrailer, oder gleich der komplette Film. In Trailern, also zum Beispiel für Kinoankündigungen, wird sehr vieles nur angedeutet, um die Neugierde zu wecken. In besonders schnellen Schnitten wird ein Ausblick auf die Thematik des Films gegeben, Einzelheiten sind aber nicht zu erkennen. Oft werden diese Trailer durch gesprochenen und geschriebenen Text eingeleitet, die zwar eine düstere Stimmung erzeugen, vom Wortlaut her aber harmlos sind (etwa: »es sollte für das junge Ehepaar ein erholsamer Urlaub werden«).

Gewaltspiele Die vorliegenden Gewaltspiele sind allesamt aus dem Umfeld von Ego-Shootern, also von bewaffneten Kampfspielen aus der Ich-Perspektive oder der Über-Schulter-Perspektive. Im unteren Abschnitt des Videos sind üblicherweise die gerade gewählten Waffen zu sehen, die je nach Setting entweder militärisch (z. B. Maschinengewehr), futuristisch (z. B. Laserpistolen) oder provisorisch

[8] zum Beispiel die Online-Filmdatenbank (http://www.ofdb.de, Stand 25.10.2012)

[9] Zu den Slasherfilmen (engl.: slash »aufschlitzen«) werden einige Horrorfilme der 1970er und frühen 1980er Jahre gerechnet. Die meisten der diesem Subgenre zugeordneten Filme folgen einem starren Handlungsschema [Dika, 1990]. Einige Produktionen seit Mitte der 1990er Jahre werden ebenfalls diesem Subgenre zugerechnet.

(z. B. Nagelstock) wirken. Handelt es sich bei dem Video um offiziell zum Spiel herausgegebene Demonstrationen seitens der Hersteller, so sind durchaus Übereinstimmungen zu Spielfilmen zu finden. So wird auch hier durch eingeblendeten und gesprochenen Text zunächst eine düstere Stimmung erzeugt, Gewalt wird aber zunächst nicht explizit dargestellt. Es dauert mitunter Minuten, bis die mitgetragenen Waffen zum Einsatz kommen. Durch die mittlerweile weit fortgeschrittene Computertechnik werden dann Schießereien und Explosionen möglichst spektakulär in Szene gesetzt. Von etwas anderer Natur als die offiziellen Trailer der Hersteller sind die zunehmend beliebter werdenden *Let's Play* Videos. Hier filmt ein Spieler den Verlauf seines Spieles selbst und nimmt dazu Kommentare auf. Es wirkt also, als würde man direkt neben ihm sitzen und auf den gleichen Bildschirm schauen. Vornehmlich beschreibt ein Video dann einen logischen Abschnitt im Spiel, etwa das Erfüllen einer Spielaufgabe (Mission). Jedes *Let's Play* Video ist somit einzigartig.

3.2.4 Zusammenfassung: Charakterisierung von Web-Inhalten im Bereich Gewalt

In den unterschiedlichen Quellen wird Gewalt relativ einheitlich definiert. Zusammenfassend kann die KJM-Definition übernommen werden: »Gewalt ist die physische, psychische oder materielle Schädigung von personalen und dinglichen Objekten durch Subjekte, Naturgewalten oder Unfälle.« [KJM, 2010]. Gewalt lässt sich bezüglich der Akteure (Menschen, Tiere, Fantasiefiguren) beziehungsweise der Art der Gewalt (Foltern, Töten usw.) in Unterthemen unterteilen. Bei der Bewertung muss berücksichtigt werden, was dargestellt wird und wie (z. B. realitätsnah) es dargestellt wird. Weiterhin sind die Gesamtzusammenhänge zu überprüfen, beispielsweise der Umfang der Gewaltdarstellung im Verhältnis zum gesamten Angebot. Eine Abgrenzung zu medienrechtlich unzulässigen, das heißt strafbaren Darstellungen von Gewalt ist erforderlich. Die vorliegenden exemplarischen Webseiten enthalten reale und fiktive Inhalte und sind zum Teil indiziert, das heißt verboten. Verbotene beziehungsweise jugendschutzrelevante Gewaltdarstellungen haben in multimedialen Formaten (Bilder, Videos) eine besondere Relevanz.

3.3 Charakterisierung von Web-Inhalten im Bereich Extremismus

Im folgenden Abschnitt werden zunächst verschiedene Definitionen von Extremismus aufgeführt. Anschließend werden die Kriterien zusammengestellt, die in oben genannten Kriterienkatalogen (vgl. 3.1) zur Identifikation von Extremismus angegeben werden. Dabei wird insbesondere auf die Altersklassifikation eingegangen. Der dritte Teil enthält eine Beschreibung der Daten, die dieser Studie als Beispiele für jugendschutzrelevante Webseiten in Bezug auf Extremismus dienen.

3.3.1 Definitionen

Zur Einführung in das Thema werden im Folgenden verschiedene Definitionen zu Extremismus vorgestellt.

Duden »Extremismus: extreme, radikale [politische] Haltung oder Richtung«[10]

Wikipedia »Die Attribute ›extrem‹ und ›extremistisch‹ sind vom lateinischen Wort extremus abgeleitet, dem Superlativ von ›außen‹ (exterus), laut Stowasser übersetzbar als ›das Äußerste‹, ›das Entfernteste‹ oder ›das Ärgste‹. Der Begriff geht von der Vorstellung eines ›politischen Spektrums‹ aus, das

[10] siehe http://www.duden.de/rechtschreibung/Extremismus, Stand 26.9.2012

aus einer normativen Mitte und ›Rändern‹ (›links außen‹ und ›rechts außen‹) besteht. Eine ›extreme‹ Position wird demgemäß als Rand im Verhältnis zur angenommenen Mitte, zugleich als Minderheit im Verhältnis zur derzeitigen Mehrheit und als Gegensatz zu deren politischer Orientierung und zur herrschenden politischen Ordnung betrachtet. Damit wird diese Ordnung und die Mehrheitshaltung dazu zugleich als ›normaler‹, gültiger und zu schützender Zustand akzeptiert und bewertet. Als ›extremistisch‹ lassen sich also jene Teile eines politischen Spektrums kennzeichnen, von denen aktive Gefährdungen der Grundwerte der zur Zeit herrschenden politischen Ordnung ausgehen, prinzipiell unabhängig von der weltanschaulichen Ausrichtung und den konkreten politischen Zielen der Vertreter ›extremer‹ Positionen.«[11]

Strafgesetzbuch Aus dem Strafgesetzbuch sind im Kontext von Extremismus vornehmlich §86 *Verbreiten von Propagandamitteln verfassungswidriger Organisationen*, §86a *Verwenden von Kennzeichen verfassungswidriger Organisationen* zu nennen, sowie »§ 130 Volksverhetzung

(1) Wer in einer Weise, die geeignet ist, den öffentlichen Frieden zu stören,
   1. gegen eine nationale, rassische, religiöse oder durch ihre ethnische Herkunft bestimmte Gruppe, gegen Teile der Bevölkerung oder gegen einen Einzelnen wegen seiner Zugehörigkeit zu einer vorbezeichneten Gruppe oder zu einem Teil der Bevölkerung zum Hass aufstachelt, zu Gewalt- oder Willkürmaßnahmen auffordert oder
   2. die Menschenwürde anderer dadurch angreift, dass er eine vorbezeichnete Gruppe, Teile der Bevölkerung oder einen Einzelnen wegen seiner Zugehörigkeit zu einer vorbezeichneten Gruppe oder zu einem Teil der Bevölkerung beschimpft, böswillig verächtlich macht oder verleumdet, wird mit Freiheitsstrafe von drei Monaten bis zu fünf Jahren bestraft.

(2) Mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe wird bestraft, wer
   1. Schriften (§ 11 Absatz 3), die zum Hass gegen eine vorbezeichnete Gruppe, Teile der Bevölkerung oder gegen einen Einzelnen wegen seiner Zugehörigkeit zu einer vorbezeichneten Gruppe oder zu einem Teil der Bevölkerung aufstacheln, zu Gewalt- oder Willkürmaßnahmen gegen sie auffordern oder ihre Menschenwürde dadurch angreifen, dass sie beschimpft, böswillig verächtlich gemacht oder verleumdet werden,
      a) verbreitet,
      b) öffentlich ausstellt, anschlägt, vorführt oder sonst zugänglich macht,
      c) einer Person unter achtzehn Jahren anbietet, überlässt oder zugänglich macht oder
      d) herstellt, bezieht, liefert, vorrätig hält, anbietet, ankündigt, anpreist, einzuführen oder auszuführen unternimmt, um sie oder aus ihnen gewonnene Stücke im Sinne der Buchstaben a bis c zu verwenden oder einem anderen eine solche Verwendung zu ermöglichen, oder
   2. eine Darbietung des in Nummer 1 bezeichneten Inhalts durch Rundfunk, Medien- oder Teledienste verbreitet.

(3) Mit Freiheitsstrafe bis zu fünf Jahren oder mit Geldstrafe wird bestraft, wer eine unter der Herrschaft des Nationalsozialismus begangene Handlung der in § 6 Absatz 1 des Völkerstrafgesetzbuches bezeichneten Art in einer Weise, die geeignet ist, den öffentlichen Frieden zu stören, öffentlich oder in einer Versammlung billigt, leugnet oder verharmlost.

(4) Mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe wird bestraft, wer öffentlich oder in einer Versammlung den öffentlichen Frieden in einer die Würde der Opfer verletzenden Weise dadurch stört, dass er die nationalsozialistische Gewalt- und Willkürherrschaft billigt, verherrlicht oder rechtfertigt.

(5) Absatz 2 gilt auch für Schriften (§ 11 Absatz 3) des in den Absätzen 3 und 4 bezeichneten Inhalts.

(6) In den Fällen des Absatzes 2, auch in Verbindung mit Absatz 5, und in den Fällen der Absätze 3 und 4 gilt § 86 Absatz 3 entsprechend.«

[11] siehe http://de.wikipedia.org/wiki/Extremismus, Stand 27.11.2012

Jugendschutzgesetz Das JuSchG bezieht sich in der Auflistung der schwer jugendgefährdenden Trägermedien auf § 130 StGB. Als jugendgefährdend »zählen vor allem unsittliche, verrohend wirkende, zu Gewalttätigkeit, Verbrechen oder Rassenhass anreizende Medien«, sowie Trägermedien, die »den Krieg verherrlichen« [JuSchG, 2008].

Jugendmedienschutz-Staatsvertrag »§ 4 Unzulässige Angebote

(1) Unbeschadet strafrechtlicher Verantwortlichkeit sind Angebote unzulässig, wenn sie
1. Propagandamittel im Sinne des § 86 des Strafgesetzbuches darstellen, deren Inhalt gegen die freiheitliche demokratische Grundordnung oder den Gedanken der Völkerverständigung gerichtet ist,
2. Kennzeichen verfassungswidriger Organisationen im Sinne des § 86 a des Strafgesetzbuches verwenden,
3. zum Hass gegen Teile der Bevölkerung oder gegen eine nationale, rassische, religiöse oder durch ihr Volkstum bestimmte Gruppe aufstacheln, zu Gewalt- oder Willkürmaßnahmen gegen sie auffordern oder die Menschenwürde anderer dadurch angreifen, dass Teile der Bevölkerung oder eine vorbezeichnete Gruppe beschimpft, böswillig verächtlich gemacht oder verleumdet werden,
4. eine unter der Herrschaft des Nationalsozialismus begangene Handlung der in § 6 Absatz 1 und § 7 Absatz 1 des Völkerstrafgesetzbuches bezeichneten Art in einer Weise, die geeignet ist, den öffentlichen Frieden zu stören, leugnen oder verharmlosen,
5. grausame oder sonst unmenschliche Gewalttätigkeiten gegen Menschen in einer Art schildern, die eine Verherrlichung oder Verharmlosung solcher Gewalttätigkeiten ausdrückt oder die das Grausame oder Unmenschliche des Vorgangs in einer die Menschenwürde verletzenden Weise darstellt; dies gilt auch bei virtuellen Darstellungen,
6. […]
7. den Krieg verherrlichen,

[…]«

Bundesamt für Verfassungsschutz (BfV) Das deutsche Bundesamt für Verfassungsschutz definierte im Jahr 2000 Extremismus in Form einer *definitio ex negativo* als ›fundamentale Ablehnung des demokratischen Verfassungsstaats‹. Darunter fallen für ihn alle Bestrebungen, die sich gegen den ›Kernbestand‹ des Grundgesetzes beziehungsweise der FDGO[12] insgesamt richten.

Kommission für Jugendmedienschutz der Landesmedienanstalten (KJM) In den *Kriterien für die Aufsicht im Rundfunk und in den Telemedien* [KJM, 2010] fällt Extremismus unter »Eigenverantwortlichkeit und Gemeinschaftsfähigkeit« und ist in dem Punkt »Weltanschauliche, religiöse und politische Extremismen« zu finden.

Freiwillige Selbstkontrolle Multimedia-Diensteanbieter (FSM) »Darstellungen von Extremismus sind mediale Angebote, in denen politisch-weltanschauliche Totalitarismen oder religiöse Fundamentalismen angedeutet, dargestellt oder propagiert werden.« [FSM, 2011d]

Uwe Backes Professor Dr. Uwe Backes, Professor für vergleichende Diktaturforschung an der TU Dresden, definierte Extremismus als politische Diskurse, Programme und Ideologien, die sich implizit oder explizit gegen grundlegende Werte und Verfahrensregeln demokratischer Verfassungsstaaten richten, wobei – im Gegensatz zu radikal-egalitären Strömungen (Linksextremismus) – der Rechtsextremismus das Ethos fundamentaler Menschengleichheit nicht anerkennt [Backes, 2001].

[12] Freiheitlich demokratische Grundordnung

3.3.2 Kriterien und Altersklassifikation

Extremismus kann unterteilt werden in politischen, weltanschaulichen und religiösen Extremismus. Der Fragebogen der FSM[13] fragt hierzu nach »demokratiefeindlichen oder politisch-extremistischen Anschauungen, die nicht sanktioniert oder kritisiert werden« und »die einseitig positive Darstellung von Nationalsozialismus/Faschismus«. Eine besondere Rolle im Kontext von Extremismus spielen Foren und andere Kommunikationsmöglichkeiten zwischen den Nutzern oder den Nutzern und den Anbietern.

Die KJM unterscheidet beim Extremismus nicht zwischen Entwicklungsbeeinträchtigung und Entwicklungsgefährdung. Als Kriterium für gefährdende oder beeinträchtigende Angebote wird dort angegeben

- »wenn das Gewaltmonopol des Staates negiert wird«,
- wenn »rechtsstaatliche und demokratische Instrumentarien zur Durchsetzung gesellschaftlicher Interessen als untauglich bewertet werden« und
- wenn »die religiöse Freiheit oder freie Religionsausübung in Frage gestellt werden«. [KJM, 2010]

Unter politischen Extremismus fallen beispielsweise »Angebote, die

- die grundsätzliche Freiheit und Gleichheit jedes Individuums verneinen und stattdessen vom Prinzip der Volksgemeinschaft ausgehen, in der der Einzelne nichts und das Volk alles ist,
- in diskriminierender Weise Ausländer für Missstände und Probleme verantwortlich machen,
- das grundgesetzlich festgelegte Mehrparteienprinzip bekämpfen und stattdessen ein Führerprinzip propagieren,
- offen Gewalt als Mittel zur Durchsetzung politischer Ziele bejahen,
- sich an der Ideologie des Nationalsozialismus orientieren,
- den Wert von Menschen aus deren Rasse oder Abstammung ableiten,
- der Demokratie als Staatsform der Bundesrepublik Deutschland grundsätzlich ablehnend gegenüberstehen und z. B. die Wiedereinführung des Deutschen Reiches fordern«. [KJM, 2010]

Bei der Bewertung extremistischer Inhalte ist zunächst abzuprüfen, ob die Inhalte nicht im Sinne § 4 Absatz 1 JMStV unzulässig sind. Für die Erkennung politisch extremistischer, unzulässiger Inhalte listet [KJM, 2010] detaillierte Kriterien auf. Auf der anderen Seite muss festgestellt werden, ob die Inhalte unter den Schutz der Meinungs- oder Pressefreiheit oder der Berichterstattung zum politischen Zeitgeschehen fallen.

In den Prüfgrundsätzen der FSM [FSM, 2011d] werden unter schwer entwicklungsgefährdend folgende Kriterien genannt:

- »Darstellungen im Sinne demokratiefeindlicher, rassistischer, völkischer oder nationalistischer Ideologien«,
- »Verbreitung von destruktiv-sektiererischen Vorstellungen des Satans- oder Hexenglaubens oder anderer destruktiv-extremistischen Glaubensrichtungen« und
- »Beschimpfen von Glaubensbekenntnissen, Religionsgemeinschaften oder Weltanschauungsvereinigungen«.

Alterseinstufung Eine besondere Risikogruppe für extremistische Inhalte ist die Altersgruppe ab 14 (bis 18) Jahren in der Phase der psychosozialen Persönlichkeits- und Identitätsentwicklung. Neben der

[13] siehe https://www.altersklassifizierung.de/, Stand 28.9.2012

Bewertung der Inhalte ist daher festzustellen, wer die Zielgruppe ist und inwieweit die Darstellungen Identifikationsmöglichkeiten für Jugendliche bieten. Bei jüngeren Kindern besteht hauptsächlich die Gefahr »einer prekären Sprachentwicklung/-verwendung […] und einer Ängstigung durch ggf. bedrohlich wirkende extremistische Inhalte« [FSM, 2011d]. In den Kriterien der KJM für die Aufsicht im Rundfunk und in den Telemedien [KJM, 2010] werden die Kriterien für extremistische Inhalte keinen Altersstufen zugeordnet.

[Liesching, 2011] nennt in der Zusammenfassung die Indizien für eine Zuordnung zu einer Altersstufe. Dies sind bezüglich Extremismus folgende:

- »Stufe 6 offensichtlich schwer jugendgefährdend:«
  - »das Werben oder positive Darstellen von Sekten oder destruktiv-extremistischen Glaubensrichtungen«
  - »Beschimpfen von Glaubensbekenntnissen, Religionsgemeinschaften oder Weltanschauungsvereinigungen«
- »Stufe 5 einfach jugendgefährdend:«
  - »zum Rassenhass anreizende Medien«
  - »Diskriminierung von Menschengruppen«
  - »Positive Darstellung der NS-Ideologie«
- »Stufe 4 entwicklungsbeeinträchtigend ab 18:«
  - »Inhalte […], welche indes derart subtile Botschaften enthalten und hinsichtlich ihrer Botschaften noch nicht so drastisch erscheinen, dass sie schon als (offensichtlich schwer) jugendgefährdend eingestuft werden müssen.«
- »Stufe 3 entwicklungsbeeinträchtigend ab 16:«
  - »diskriminierende oder frauenverachtenden Inhalte«
- »Stufe 2 entwicklungsbeeinträchtigend ab 12:«
  - »diskriminierende oder frauenverachtenden Inhalte«

In der Stufe 1 gibt es keine Indizien, die eindeutig dem Extremismus zugeordnet werden können.

3.3.3 Beschreibung der exemplarischen Daten im Bereich Extremismus

Als Beispiele für den Bereich Extremismus wurden für die vorliegende Studie ausschließlich Webseiten mit rechtsextremen Inhalten zur Verfügung gestellt. Anders als bei den oben beschriebenen Gewaltinhalten spielen hier jugendschutzrelevante rechtsextreme *Propaganda-Texte* eine wichtige Rolle. Allerdings gibt es auch eine Reihe von jugendschutzrelevanten Videos in diesem Bereich.

3.3.3.1 Webseiten

Die gesichteten Webseiten lassen sich in die Kategorien

- Pseudo-wissenschaftliche Informationsportale,
- Hetz-Seiten,
- Organisationswebseiten,
- Kommerzielle Webseiten und
- Rassistische Unterhaltungs-Webseiten

unterteilen.

Pseudo-wissenschaftliche Informationsportale Auf diesen Webseiten wird zum Beispiel geschichtsrevisionistisches Gedankengut vermittelt. Dabei wird einem wissenschaftlichen Anspruch gefolgt, das heißt es wird versucht, Aussagen durch Beweisführung und/oder Quellenangaben zu untermauern; dabei wird auch Wert auf strukturelle Homogenität gelegt (vereinheitlichte Zitierweise o.ä.). Diese Webseiten sind im Allgemeinen sehr textlastig und nicht stark bebildert. Das darin vermittelte geschichtsrevisionistische Gedankengut erschließt sich dem Leser daher oft erst nach mehreren Absätzen der Lektüre. Bei manchen, zum Teil enzyklopädisch gehaltenen Webseiten ist der Versuch erkennbar, auf englische Lehnwörter zu verzichten (etwa: »Weltnetz« statt »Internet«).

Hetz-Seiten Während die pseudo-wissenschaftlichen Portale zumeist dem rechtsextremistischen Gedankengut zuzuordnen waren, sind die Hetz-Seiten wesentlich heterogener. Was sie eint, ist der Versuch, durch kurze Parolen, aus dem Kontext gerissene Zitate oder kurz dargebotene historische Ereignisse eine gewisse Geisteshaltung gegenüber einer Bevölkerungsgruppe abzuleiten. Dies kann islamistisch, islamfeindlich oder rechtsextrem sein. So werden zum Beispiel auf einer Seite Thalmud-Zitate wiedergegeben, die belegen sollen, dass der Glaube der Juden menschenverachtende Züge habe.

Organisationswebseiten In diese Kategorie fallen Webauftritte von Organisationen, die rechtsextremem Gedankengut nahe stehen.

Kommerzielle Webseiten Darunter werden Webseiten verstanden, auf denen primär Produkte vertrieben werden, die eindeutig einem extremistischen Hintergrund entspringen. Das sind überwiegend englischsprachige Online-Shops, die diverse Gegenstände mit Bezug zu Nazi-Deutschland vertreiben. Die Spanne reicht hierbei vom einfachen Buchversand, zum Beispiel *Mein Kampf*, bis hin zu T-Shirts, CDs oder Modeschmuck. Die Aufmachung der Webseiten ist dabei sehr heterogen. Neben behelfsmäßig wirkenden, einfachen Shopsystemen wird häufig ein professionelles Auftreten beobachtet, vor allem im amerikanischen Bereich, wo der Vertrieb der Symbole nicht unter Strafe steht. Die Webseiten besitzen Empfehlungsfunktionen, Einkaufswagen, ordentlich aufgemachte Kategorien und erweiterte Suchfunktionen.

Rassistische Unterhaltungs-Webseiten Auf diesen Webseiten befindet sich eine Ansammlung von Comics, mit Text unterlegten Fotos oder Witzen, die einen rassistischen Tenor besitzen (etwa: unkommentierte Zeichnungen von Juden aus Zeitungen des dritten Reiches, Sammlungen von Polenwitzen etc).

3.3.3.2 Videos

Die zur Verfügung gestellten Videos stammen vornehmlich aus dem Bereich

- Propaganda-Berichte,
- Propaganda-Interviews und
- Rechtsextreme Musik.

Propaganda-Berichte Bei den vorliegenden Videos aus dem Ressort Propaganda-Berichte handelt es sich um geschichtsrevisionistische, dokumentarisch anmutende Filmbeiträge. Nur ein Teil der Videos ist durchgehend auf deutsch, oft wird über weite Strecken die deutsche Übersetzung als Untertitel eingeblendet oder von einem zusätzlichen Sprecher hinzugefügt. Die Tonqualität ist insbesondere bei den offensichtlich von Videokassetten digitalisierten Beiträgen über weite Strecken sehr schlecht. Die Farbgebung ist stark verändert, der Text oft nur durch genaues Hinhören noch verständlich. Auch ist bei vielen Beiträgen an historischen Orten, wie etwa Gaskammern in Konzentrationslagern, eine sehr schlechte Tonqualität durch Halleffekte oder unprofessionelle Mikrophone zu

beobachten. Da die Videos sehr lang sind und die meisten in voller Länge gezeigt werden, ist ein Aufteilen des Videomaterials bei Videoplattformen, die nur begrenzte Länge erlauben (etwa 10 Minuten), über mehrere Teile beobachtbar. Teilweise wird ihnen ein Disclaimer vorangeschickt, der besagt, dass der Beitrag der wissenschaftlichen Aufarbeitung dient und »unter besonderer Berücksichtigung der bundesdeutschen Holocaust-Schutzgesetze«[14] entstanden ist. Die im Video sprechenden Personen werden oft mit vollem Namen und mit ihrer Funktion durch Bauchbinden eingeführt. Aussagen werden mit zum Teil offenkundig zweckentfremdetem Archivmaterial belegt.

Propaganda-Interviews Die vorliegenden Interviews mit Propaganda-Inhalt unterscheiden sich von den Berichte dadurch, dass die Interview-Situationen oft spontaner wirken. Entweder in Monologen oder durch einen Moderator befragt beschreiben die Hauptpersonen ihre Sicht auf historische Gegebenheiten. Als Kulisse dienen hierbei so unterschiedliche Orte wie etwa historische Bauten, Gerichtsflure oder Studios. Der Text ist häufig flüssig und eloquent und wird selten durch den Moderator unterbrochen, so dass der Sprechende längere, bisweilen eingeübt wirkende Passagen am Stück redet. Parolen werden weitestgehend vermieden, die generelle Unmut über beispielsweise die deutsche Gerichtsbarkeit entpuppt sich erst beim genauen Hinhören als rechtsextremistisches Verschwörungskonstrukt. Bei manchen Videos werden zunächst Passagen von (teilweise indizierten) geschichtsrevisionistischen Liedern eingeblendet und dabei durch Bildmaterial ergänzt, das teilweise auf das Lied abgestimmt ist. Einige Videos setzen stark auf emotionalisierende Bilder, wie etwa hingerichtete Nationalsozialisten in raschem Wechsel mit im Krieg zerbombten deutschen Städten, gefolgt von deutschen Politikern der Neuzeit, die mit israelischen Politikern an einem runden Tisch sitzen.

Rechtsextreme Musik Bei den vorliegenden Videos aus dem Ressort rechtsextremer Musik war kein Musikvideo im engeren Sinne enthalten (also ein Video, dass z. B. von der Band selbst zur Vermarktung des entsprechenden Liedes produziert wurde). Stattdessen wurden von Anhängern der jeweiligen Band Videos erstellt, die die Musik durch Fotos, Schrift oder Logos begleiten. Diese können hochauflösend sein, häufig sind die im Video eingebetteten Bilder aber auch sehr verpixelt und zum Teil selbst für das menschliche Auge nur aus dem Kontext als solche zu erkennen. Häufig vorhanden waren die CD-Cover des jeweiligen Albums, die zum Teil eindeutige Merkmale wie etwa strafrechtlich relevante Symbole (Hakenkreuze, Triskele, SS-Runen, …) enthielten, zum anderen die in der Szene oft zur Verschleierung benutzten Zahlencodes (etwa: 18 oder 88) als Indizien aufwiesen. Ebenfalls häufig fanden sich Logos von Organisationen und Vereinigungen, die ebenfalls der rechten Szene zugeordnet werden können und deren Erkennungszeichen nach einem Verbot in Deutschland ebenfalls strafbar sind (zum Beispiel *Blood & Honour*). Andere Bilder waren Aufnahmen von rechtsextremen Graffitisprüchen oder von Versammlungen, in der die Anwesenden mit Hitlergruß posierten. Bei manchen Videos sind zudem Fotos der Bandmitglieder eingeblendet, die etwa mit Hakenkreuzflagge posieren oder Musikinstrumente spielen. Zum Teil sind dabei ihre Gesichter jedoch unkenntlich gemacht.

Der musikalische Charakter des Materials ist nicht einheitlich. Zwar finden sich auch getragene, fast ausschließlich instrumentale Stücke, die eher dem *Industrial*-Genre zuzuordnen sind, der überwiegende Teil gehört aber entweder zur Kategorie der Punkmusik oder zu einer Variante des Black Metal (National Socialist Black Metal), wo zu schneller Gitarrenmusik entweder gegrölt oder geschrien wird. Der Text ist selten Deutsch, oft ist nicht einmal die Sprache klar zu verstehen; allerdings wird zu den Videos oft eine Transkription geliefert. Bei manchen Videos wird sogar unten im Bild die englische Transkription und oben die spanische Übersetzung gleich mitgeliefert. Bei diesen unterstreicht der Text die menschenverachtende, rassistische Tendenz.

In manchen Stücken finden sich hineingeschnittene Zitate aus anderen Audio- oder Filmwerken. Diese Passagen werden aus dem Kontext gerissen und erzeugen so zum Teil sogar die gegenteilige

[14] gemeint ist §86 Absatz 3 StGB

Aussage. So findet sich in einem Video mit rechtsextremer Musik auch eine längere Sequenz aus dem britischen Propagandafilm *49th Parallel*. Dort versuchen drei Nationalsozialisten, im Rahmen einer Dorfversammlung die Gemeinschaft für ihre Sicht zu gewinnen, der Aufruf zum gemeinsamen »Heil Hitler«-Ruf bleibt aber wirkungslos; alle Dorfbewohner bleiben sitzen. Im Musikvideo wird diese eigentlich peinlich wirkende Stille gezielt durch einsetzende Gitarrenriffs kaschiert.

3.3.4 Zusammenfassung: Charakterisierung von Web-Inhalten im Bereich Extremismus

Während die Definitionen von Duden und Wikipedia den Schwerpunkt auf den politischen Extremismus legen, schließen FSM und KJM den weltanschaulichen und religiösen Extremismus ein. Der Fokus dieser Studie liegt jedoch sowohl aus der Sicht der Kriterienkataloge als auch auf der des Beispielmaterials auf dem politischen (Rechts-) Extremismus. Gesetze und Erläuterungen in Form von Kriterien regeln eindeutig, welche rechtsextremistischen Inhalte unzulässig sind. Zur Bewertung von jugendschutzrelevanten, extremistischen Inhalten liegen ebenfalls Kriterien vor, wobei stets die Abgrenzung zu verbotenem Material vorzunehmen ist. Besonders gefährdet sind Jugendliche zwischen 14 und 18 Jahren. Bei der Bewertung muss stets mit dem Recht auf Meinungs- und Pressefreiheit und dem § 86 Absatz 3 StGB abgewogen werden. Als exemplarische Web-Inhalte liegen Webseiten und Videos vor. Auffällig ist, dass hier geschriebene oder gesprochene Texte eine große Rolle spielen und ausgewertet werden müssen.

3.4 Charakterisierung von Web-Inhalten im Bereich Selbstgefährdung

Im folgenden Abschnitt wird zunächst ausgearbeitet, was im Allgemeinen und im Kontext dieser Studie unter *Selbstgefährdung* verstanden wird. Anschließend werden die Kriterien zusammengestellt, die in oben genannten Kriterienkatalogen (vgl. 3.1) zur Identifikation von selbstgefährdenden Inhalten und zur Altersklassifikation angegeben werden. Die unterschiedlichen Kategorien innerhalb der Selbstgefährdung werden erläutert. Der dritte Teil beschreibt die Daten, die dieser Studie als Beispiele für jugendschutzrelevante Webseiten in Bezug auf Selbstgefährdung dienen.

3.4.1 Definitionen

Duden Der Duden kennt den Begriff *Selbstgefährdung* nicht.[15]

Wikipedia Selbstgefährdendes Verhalten: »Unter selbstgefährdendem Verhalten versteht man Handlungen, die bewusst oder unbewusst eine Schädigung der eigenen Gesundheit verursachen könnten. Diese Schädigung kann sowohl physisch als auch psychisch sein.«[16]

Jugendschutzgesetz § 18 Absatz 1: »Träger- und Telemedien, die geeignet sind, die Entwicklung von Kindern oder Jugendlichen oder ihre Erziehung zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit zu gefährden, sind von der Bundesprüfstelle für jugendgefährdende Medien in eine Liste jugendgefährdender Medien aufzunehmen. Dazu zählen vor allem unsittliche, verrohend wirkende, zu Gewalttätigkeit, Verbrechen oder Rassenhass anreizende Medien …«

[15] siehe http://www.duden.de/suchen/dudenonline/Selbstgefährdung, Stand 29.11.2012
[16] siehe http://de.wikipedia.org/wiki/Selbstgefährdendes_Verhalten, Stand 4.10.2012

Jugendmedienschutz-Staatsvertrag § 5 Absatz 1: »Sofern Anbieter Angebote, die geeignet sind, die Entwicklung von Kindern oder Jugendlichen zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit zu beeinträchtigen, verbreiten oder zugänglich machen, haben sie dafür Sorge zu tragen, dass Kinder oder Jugendliche der betroffenen Altersstufen sie üblicherweise nicht wahrnehmen.«

Weitere Quellen Wird von selbstgefährdenden Web-Inhalten gesprochen, so geht es dabei sowohl um die Darstellung selbstgefährdenden Verhaltens (z. B. Bilder von geritzten Körpern) als auch um den Aufruf zu eben solchem Verhalten. Die diesbezüglich ausgewerteten Quellen ordnen Verhaltensweisen der Selbstgefährdung unterschiedlichen Überschriften zu. In den Kriterien der KJM für die Aufsicht im Rundfunk und in den Telemedien [KJM, 2010] ist diesbezüglich unter der Überschrift »Eigenverantwortlichkeit und Gemeinschaftsfähigkeit« der Punkt »Risikoverhalten: physische und psychische Integrität« zu finden. In den Prüfgrundsätzen der FSM [FSM, 2011d] fallen entsprechende Webseiten unter die sozial-ethisch desorientierenden Angebote; [Liesching, 2011] verwendet den Begriff »dem sozial-ethischen Wertebild zuwiderlaufende Inhalte«.

Begriffsdefinition im Rahmen der vorliegenden Studie Die verschiedenen Quellen verwenden für den Begriff Selbstgefährdung unterschiedliche Definitionen und Beschreibungen. Aufgrund der im Rahmen der vorliegenden Studie untersuchten Beispieldaten werden im Folgenden diese Punkte unter dem Begriff Selbstgefährdung verstanden:

- Verherrlichung von bzw. Aufforderung zu gesundheitsschädigenden Handlungen: Pro Anorexia nervosa (Pro-Ana)/Pro Bulimia nervosa (Pro-Mia), Selbstverletzung (Ritzen), Suizid, Ohnmacht,
- Verherrlichung von bzw. Aufforderung zum (übermäßigen) Genuss von Suchtmitteln, ggf. mit Handlungsanleitungen,
- Anleitungen zur Durchführung krimineller Handlungen bzw. Aufforderung zu einer Straftat.

3.4.2 Kriterien und Altersklassifikation

In den Prüfgrundsätzen der FSM [FSM, 2011d] werden im Abschnitt 9.2.1 Kriterien für schwer entwicklungsgefährdende Angebote folgende unter Selbstgefährdung fallende Punkte genannt:

- Verführung zum Erwerb oder Gebrauch von Suchtmitteln,
- Aufforderung zu Straftaten, die von § 130a StGB nicht erfasst werden,
- Aufruf zum Suizid, zur Selbstverletzung oder Selbstgefährdung.

Als Kriterien für entwicklungsbeeinträchtigende Angebote werden in den Prüfgrundsätzen der FSM die sozial-ethisch desorientierenden Angebote genannt. Bezüglich der Selbstgefährdung heißt es hier:

»Nicht zuletzt sind auch Angebote auf eine mögliche sozial-ethische Desorientierung hin zu prüfen, die fragwürdige Identitätsentwürfe als anstrebenswerte Persönlichkeitsideale vermitteln. Solche Angebote sind dann als sehr bedenklich einzustufen, wenn mit den spezifischen Darstellungen nahe gelegt oder dazu aufgefordert wird, dass Kinder und Jugendliche ihr Selbstbild aktiv einem als Ideal dargestellten Identitätsentwurf, der wünschens- und nachahmungswert ist, anpassen und dabei gesundheitliche Risiken bezogen auf Körper und Psyche eingehen sollen (z. B. durch Schönheitsoperationen), über die nur unzureichend aufgeklärt wird. Solche Darstellungen sind in der Lage, die Entwicklung einer stabilen Ich-Identität nachhaltig zu stören, insbesondere dann, wenn auch im sozialen Umfeld der Jugendlichen entsprechende Persönlichkeitsideale kaum kritisch hinterfragt und als anstrebenswert propagiert werden.«

Die KJM [KJM, 2010] führt unter »Eigenverantwortung und Gemeinschaftsfähigkeit« den Punkt »Risikoverhalten: physische und psychische Integrität« auf. Hier geht es um Angebote

- » die ästhetisch motivierte Eingriffe am eigenen Körper (Schönheitsoperationen, Piercings, Tätowierungen) zeigen,
- die riskante und waghalsige Aktionen wie Mutproben oder Kunststücke enthalten, bei denen Verletzungen des eigenen Körpers billigend in Kauf genommen werden oder sogar beabsichtigt sind,
- die psychologisch motivierte Schädigungen am eigenen Körper thematisieren (z. B. Anorexie, Bulimie oder selbst verletzendes Verhalten),
- die unkontrollierten und exzessiven Drogen- und Alkoholkonsum zum Inhalt haben.«

Die verschiedenen Ausprägungen der Selbstgefährdung haben jeweils ihre eigenen Definitionen und sind in den Quellen sehr unterschiedlich beschrieben. Daher gehen die folgenden Absätze kurz einzeln darauf ein.

Pro-Ana/Pro-Mia Anorexia Nervosa (Magersucht) und Bulimia Nervosa sind Essstörungen, die vermehrt bei Jugendlichen in der Pubertät auftreten. Laut der Bundeszentrale für gesundheitliche Aufklärung (BZgA) kommt Magersucht am häufigsten bei 14-Jährigen[17] vor. Webseiten, die diese Krankheiten verherrlichen oder als Lebensentwürfe propagieren, haben als Zielgruppe dementsprechend Jugendliche. Bei der Erkennung solcher Seiten geht es hauptsächlich um die Abgrenzung von beratenden Angeboten, die betroffenen Personen Auswege aus der Krankheit aufzeigen möchten. Dazu dient auch das Beurteilungsraster für Internet-Angebote zu Pro-Anorexie-Angeboten [Rauchfuß, 2008] von jugendschutz.net. Das Raster formuliert teilweise exakte Kriterien, die einfach zu überprüfen sind, wie beispielsweise das Vorkommen eines »Glaubensbekenntnisses«. Dennoch geht in die Gesamtbewertung der einzelnen Kriterien immer auch die eher subjektive Bewertung »wie drastisch ist es« mit ein. Das Bewertungsraster macht keine Aussagen über die Altersein-stufung. Liesching nennt als Verstoß gegen § 4 Absatz 2 Nummer 3 JMStV folgenden Einzelfall: »Ein öffentliches Internet-Forum zum Thema Essstörungen, wenn es den mit einer Essstörung verbundenen körperlichen Zustand verherrlicht und die zum Teil katastrophalen Folgen verharmlost.« [Liesching, 2011]

Selbstverletzung: Ritzen Ritzen als Form des autoaggresiven Verhaltens bezeichnet das Aufschneiden, Aufkratzen oder Aufritzen der Haut mit spitzen oder scharfen Gegenständen. Häufig werden Rasierklingen verwendet, manchmal aber auch Glasscherben oder Messer. Laut einer Heidelberger Jugendgesundheitsstudie unter 14- bis 16-jährigen Jugendlichen [Haffner et al., 2006] geben 10,1 % der Jungen in dem Alter an, sich schon einmal verletzt zu haben, 2 % häufiger als 3mal im Jahr, bei den Mädchen sind es sogar 19,9 % beziehungsweise 5,9 %, die dies häufiger als 3mal im Jahr tun. Dem gegenüber stehen in der Studie 1,8 % beziehungsweise 5,1 % der Eltern, die angeben, dies bei ihrem Kind zu wissen.

Suizid Zur Beurteilung von Internet-Angeboten, die zum Suizid aufrufen beziehungsweise diesen verherrlichen, hat jugendschutz.net ebenfalls ein Beurteilungsraster entwickelt, das jedoch nicht öffentlich vorliegt. Zum Thema Suizid sind zahlreiche Foren im WWW zu finden. Laut dem Vorsitzenden der Deutschen Gesellschaft für Suizidprävention (DGS), Prof. Felber, führen diese Foren nicht generell zu vermehrten Selbstmordfällen, sondern können sich durch die Möglichkeit einer anonymen Kommunikation sogar eher positiv auswirken (siehe [Fiedler und Lindner, 2001]).

Ohnmacht Bei sogenannten Würgespielen wird durch kurzzeitige Unterbrechung der Sauerstoffzufuhr des Gehirns etwa durch Strangulation oder durch Würgegriffe bewusst eine Ohnmacht herbei-

[17] siehe http://www.bzga-essstoerungen.de/index.php?id=57, Stand 10.10.2012

geführt. Erwachsene, aber auch Kinder und vor allem Jugendliche würgen sich selbst oder gegenseitig, um eine kurze Bewusstlosigkeit und ein damit verbundenes Rauschempfinden zu erleben.

Aufforderung zum (übermäßigen) Genuss von Suchtmitteln, ggf. mit Handlungsanleitungen Mit dem Begriff Suchtmittel werden hier nicht nur Tabak, Alkohol und harte Drogen gemeint, sondern auch andere Mittel, die Rauschzustände herbeiführen können wie zum Beispiel Klebstoff oder Pilze. Entwicklungsgefährdend sind entsprechende Darstellungen, wenn es sich um »Verherrlichung, konkrete Aufforderung oder Bewerbung eines Konsums handelt« [FSM, 2011d]. Zur Alterseinstufung »ab 16« oder »ab 12« kann bereits die »unreflektierte, tendenziell identifikatorische Darstellung problematischer Verhaltensweisen (z. B. Alkohol- und Drogenkonsum als normale, unproblematische oder ›coole‹Lebensweise)« [Liesching, 2011] führen.

Anleitungen zur Durchführung krimineller Handlungen beziehungsweise Aufforderung zu einer Straftat »Eine explizite Aufforderung zu einer Straftat ist schwer entwicklungsgefährdend« [FSM, 2011d], während das entsprechende Kriterium eines entwicklungsbeeinträchtigenden Angebots ebenda wie folgt beschrieben wird:

» […] Weitere, bei einer Überprüfung auf eine mögliche sozial-ethische Orientierung hin besonders zu berücksichtigende Inhalte, sind auch Darstellungen, in denen problematische Vorstellungen von Ethik und Moral einseitig propagiert werden. Als bedenklich einzustufen ist vor allem, wenn die vermittelten ethisch-moralischen Vorstellungen die Notwendigkeit gesellschaftlicher Normen, Werte, Regeln und Konventionen für das Zusammenleben negieren oder das Überschreiten der dort definierten und allgemein anerkannten moralischen Grenzen zum handlungsleitenden Ziel erklären oder wenn sie ein asoziales, einzig an persönlichen Interessen und Bedürfnissen orientiertes Handeln als Erfolg versprechende oder einzig mögliche Strategie zum Erreichen von persönlichem Erfolg propagieren.«

In den Kriterien der KJM für die Aufsicht im Rundfunk und in den Telemedien [KJM, 2010] heißt es: »Durch die positive Darstellung von Straftaten oder Straftätern als Identifikationsfiguren kann die Entwicklung des Rechtsempfindens nachhaltig gestört werden.«

Alterseinstufung Bezüglich selbstgefährdender Inhalte ist die Altersgruppe von 14 bis 18 Jahren besonders gefährdet. Entsprechende Angebote sind daher als besonders problematisch zu sehen, wenn die Zielgruppe Jugendliche ist. [Liesching, 2011] nennt in der Zusammenfassung die Indizien für eine Zuordnung zu einer Altersstufe. Dies sind bezüglich Selbstgefährdung folgende:

- »Stufe 6 offensichtlich schwer jugendgefährdend:«
  - »Anreizen zum Drogenkonsum sowie die Verführung zum Erwerb oder Gebrauch von Suchtmitteln«
  - »Aufforderungen zu Straftaten, die von §130a StGB nicht erfasst werden«
  - »Verherrlichung oder Aufruf zum Suizid, zur Selbstverletzung oder zur Selbstgefährdung«
- »Stufe 5 einfach jugendgefährdend:«
  - »zu Verbrechen anreizende Medien«
  - »Propagierung oder Verherrlichung des Drogenkonsums und Drogenanbaus«
  - »Verherrlichung des Alkoholkonsums«
  - »Suizid anleitende oder propagierende Medien«
  - »Pro-Anorexie-Inhalte«
- »Stufe 4 entwicklungsbeeinträchtigend ab 18:«
  - »[…] unkritische, indes noch nicht positiv übersteigerte Behandlung von Themen wie Anorexie, Selbstmord, sonstige Selbstgefährdung, Drogen- einschließlich Alkohol- und Tabakkonsum.«
- »Stufe 3 entwicklungsbeeinträchtigend ab 16:«

    - »[…] unreflektierte, tendenziell identifikatorische Darstellung problematischer Verhaltensweisen […] im Hinblick auf mögliche Nachahmungseffekte […] «
- »Stufe 2 entwicklungsbeeinträchtigend ab 12:«
    - »[…] unreflektierte, tendenziell identifikatorische Darstellung problematischer Verhaltensweisen […] im Hinblick auf mögliche Nachahmungseffekte […] «

In der Stufe 1 gibt es keine Indizien, die eindeutig der Selbstgefährdung zugeordnet werden können.

3.4.3 Beschreibung der exemplarischen Daten im Bereich Selbstgefährdung

Die vorliegenden exemplarischen Daten zur Selbstgefährdung lassen sich in Webseiten und Videos unterteilen. Der Schwerpunkt des Materials liegt auf der Kategorie *Aufforderung zu gesundheitsschädigenden Handlungen*.

3.4.3.1 Webseiten

Das vorliegende Material gliedert sich in die Teilbereiche

- Pro-Ana/Pro-Mia,
- Selbstverletzendes Verhalten,
- Anleitung zum Selbstmord,
- Verherrlichung von Rauschzuständen,
- Gefährliche Handlungsanregungen und
- Kommerzielle Webseiten zu Suchtmitteln.

Pro-Ana/Pro-Mia In den Plattformen, in denen Anorexia Nervosa und Bulimia Nervosa einseitig propagiert oder verherrlicht wird, sind einige wiederkehrende Elemente sehr häufig anzutreffen. Die Foren selbst sind oft erst durch Anmeldung erreichbar, die Anmeldeseite ist oft mit eindeutigen Hinweisen versehen, etwa »Dieses Forum ist nur für Betroffene gedacht« oder »ab einem bestimmten Rang sind regelmäßige Fotobeweise Pflicht«.

Auf den Seiten finden sich sogenannte »heilige Texte«. Dies sind Briefe von einer personifizierten Anorexie-Störung (»Ana«), die den Leser zu einer strikten Einhaltung von Hungerkuren und Abmagerung auffordert. Der Text ist dabei als Manifest weitestgehend standardisiert. Ebenfalls wiederkehrend sind weitere religiös wirkende Textelemente wie etwa ein Glaubensbekenntnis zum Hungern, zehn Gebote oder Psalme. Ergänzt werden diese durch Ratschläge, wie eine Essstörung vor der Familie oder vor Ärzten geheim gehalten werden kann. Auch gibt es fiktive Dialogszenen zwischen besagter Ana und einem zweifelnden, normalgewichtigen Mädchen, dass aufgrund seines »schwachen« Charakters beschimpft wird. Ebenfalls auf den Seiten anzutreffen sind Tabellen zum Body-Mass-Index, wo Werte im Normbereich als besonders fett angefeindet werden. In manchen Kategorien finden sich Partnersuche oder Hunger-Wettbewerbe zum gemeinsamen Abnehmen, häufig werden auch Tagebücher angelegt, wo detailliert über den Gewichtsverlauf des Verfassers berichtet wird.

Viele Bilder werden angezeigt, manche davon als Einzelbild in einem Beitrag, andere in einer Galerie. Bei den Bildmotiven gibt es prinzipiell drei Hauptkategorien: (a) die sogenannten »Thinspirations« (auch: »Thinspos«), also Fotos von dünnen oder magersüchtigen Menschen (häufig bekannte Schauspielern oder Models), (b) »Real Thinspos«, also eigene Bilder oder Bilder von Menschen aus der eigenen Umgebung – etwa durch Handyfotos oder von privaten Webseiten entnommen – und

(c) »Anti-Thinspos«, wo übergewichtige Menschen als abschreckendes Beispiel verwendet werden. Eine spezielle Unterkategorie von Bildern ist zweigeteilt: auf der linken Seite ist eine (zumeist an sich schon relativ dünne) Person zu sehen, auf der rechten Seite ist sie phototechnisch soweit bearbeitet, dass sie nochmals dünner erscheint.

Bilder auf Pro-Ana/Pro-Mia Webseiten sind häufig am Computer nachbearbeitet. Oft wird eine Schwarz-Weiß- oder Sepia-Färbung nachträglich durchgeführt, manchmal werden Personen nachträglich mit Engelsflügeln oder ähnlichem geschmückt. Auf den Bildern selbst sind häufig triggernde Sprüche zu finden, die von den Betroffenen auch als »thinlines« bezeichnet werden. Darunter fallen oftmals englische Motivationssprüche (etwa: »nothing tastes as good as thin feels«). Die Schriftart und Farbe variiert hierbei sehr stark, jedoch scheint es einen generellen Trend zu geschwungenen Schriftformen zu geben.

Selbstverletzendes Verhalten Bei den Selbstverletzungen sind hauptsächlich sich selbst zugefügte Schnittwunden (»Ritzen«), etwa durch Rasierklingen oder Scherben, sowie die selbstdurchgeführte Blutabnahme mittels medizinischer Spritzen oder Kanülen zu nennen. Die Bilder gliederten sich in drei große Gruppen: (a) reale Bilder von anderen, die meist vom Grad der Verletzung besonders bemerkenswert waren, (b) reale Bilder von den Verletzungen am eigenen Körper, zum Teil über eine große Zeitspanne aufgenommen, und (c) zum Thema passende Zeichnungen.

Suizid: Selbstmordforen und Anleitungen zum Selbstmord Im Gegensatz zu fachlich betreuten Beratungsforen, die eine lebensbejahende Möglichkeit der Krisenbewältigung aufzeigen und unterstützen wollen, behandeln die im Rahmen der vorliegenden Studie gesichteten Forumsbeiträge vornehmlich prosuizidale Tendenzen, indem sie den Betroffenen eine Austauschplattform für geeignete Methoden bieten. Ankündigungen zum Suizid sind aber durchaus auch in Nicht-Suizidforen untergebracht, etwa in der allgemeinen Rubrik »Hilfe« von normalen Hollywood-Filmeforen.

In den prosuizidalen Foren werden – im Gegensatz zu Beratungsforen – Anleitungen, Ankündigungen und Verabredungen zum Suizid ausdrücklich erwünscht. Die in den Foren verwendeten Avatare (also Icons an der Seite des Beitrags) sowie die Signaturen sind zwar kaum von provozierenden Bildern in anderen, harmloseren Foren zu unterscheiden. Allerdings ist schon die Art der Forenthemen sehr auffällig (etwa: »Drogen«, »Gas«, »Pulsader«, »Abschiedsbriefe« etc.). Bei den Anleitungen werden Themen wie die Methode selbst, die dafür nötigen Suizidmittel sowie geeignete Orte besprochen und von Forumsteilnehmern mit Tipps kommentiert.

Häufig anzutreffen sind sogenannte *Trigger*, *Tasteless-Bilder* oder Videos, die selbstverletzendes Verhalten oder zum Teil schwer entstellte, tote Menschen zeigen. Diese sind manchmal mit einer Warnung versehen. Einige Bilder sind real, viele sind professionell stilisiert und offenbar gestellt, bei einigen Bildern ist der Kontext auf den ersten Blick nicht ersichtlich. Manche sind vermutlich als schlechter Scherz angelegt (etwa: eine Medikamentenschachtel mit dem Aufdruck »Suizid – hilft gegen Depressionen« etc.).

Verherrlichung von Rauschzuständen Unter dem Begriff Rauscherfahrung werden mehrere Aspekte von selbtsgefährdendem Verhalten zusammen gefasst, die in sich zwar grundlegend verschieden sind, deren Datenmaterial aber trotzdem ähnlichen Mustern folgt: (a) Alkoholmissbrauch, (b) Drogenmissbrauch und (c) Würge-»Spiele«, also das bewusste Herbeiführen einer Ohnmacht durch kurzzeitige Unterbrechung der Sauerstoffzufuhr des Gehirns, etwa durch Strangulation oder durch Würgegriffe. Die betrachteten Seiten enthalten oft ein Sammelsurium von Tipps, Tricks und Bauanleitungen. Auf Drogenseiten wird etwa der Anbau und Konsum von bewusstseinsverändernden Pilzen oder die Züchtung von Cannabis-Pflanzen erklärt. Auf Alkoholseiten werden Gasmasken gezeigt, die mittels Schlauch an einem Trichter befestigt sind, oder es werden Anleitungen gegeben, wie man sich in möglichst kurzer Zeit viel Alkohol in den Hals schießen lassen kann.

Anleitung zur Durchführung einer Straftat Auf diesen Webseiten finden sich zumeist Bastelanleitungen für Gegenstände oder Maschinerien, die ein hohes Selbstverletzungsrisiko bergen beziehungsweise auch als Waffe gegen andere benutzt werden können (z. B. Sprengstoff-Theorien oder Geschossbausätze aus Alltagsgegenständen wie etwa Plastikrohre in Kombination mit Kartoffeln).

Kommerzielle Webseiten zu Suchtmitteln Weiterhin wurden uns Seiten von Web-Shops zur Beschaffung von legalen Suchtmitteln, die jedoch für Kinder und Jugendliche nicht geeignet sind (Zigaretten, Alkohol), zur Verfügung gestellt.

3.4.3.2 Videos

Das vorliegende Videomaterial gliedert sich in die Teilbereiche

- Pro-Ana/Pro-Mia,
- Selbstverletzendes Verhalten und
- Verherrlichung von Rauschzuständen.

Pro-Ana/Pro-Mia Die der vorliegenden Studie zur Verfügung gestellten Videos aus dem Genre »Pro-Ana« bestanden hauptsächlich aus sich abwechselnden Passagen mit Text und Fotos, zumeist mit Musik unterlegt. Fotos und Texte fielen dabei in ähnliche Kategorien wie bei den oben beschriebenen Forumsbeiträgen (Thinspirations, Motivations-Slogans).

Selbstverletzendes Verhalten Die vorliegenden Videos aus dem Bereich der Selbstverletzung zeigten hauptsächlich durch scharfe Gegenstände wie etwa Rasierklingen hervorgerufene Schnittwunden am Körper, vornehmlich am Arm. Fast alle waren mit Musik unterlegt, die häufig aus dem Bereich melancholischer Rockmusik stammten. Oft wechselten sich Bilder und geschriebener Text ab, wobei auf Textseiten selten mehr als zehn Wörter gleichzeitig zu sehen waren. Lange Sätze wurden über mehrere Bilder heruntergebrochen. Der Inhalt der Texte schwankte zwischen einer Warnung an mögliche Nachahmer bis hin zur Aussage, dass die Selbstverletzung die einzig logische Schlussfolgerung aus der jeweiligen Lebenssituation sei.

Verherrlichung von Rauschzuständen Ein Großteil der Bilder und Videos von Rauschzuständen und selbst herbeigeführter Ohnmacht sind von den Teilnehmergruppen selbst angefertigt. Hier wird während des Rausches beziehungsweise während der Ohnmacht aufgenommenes Videomaterial gezeigt. Für die Würgespiele gibt es zum Teil durch statisch montierte Kameras festgehaltene Szenen, in denen die Technik ausprobiert wird. Aufwendigere Zusammenstellungen besprechen zudem die jeweils angewandte Methode, unterlegen diese mit Text und zeigen dann die jeweilige Wirkung als zusätzliches Video.

3.4.4 Zusammenfassung: Charakterisierung von Web-Inhalten im Bereich Selbstgefährdung

Es gibt keine einheitliche Definition des Begriffes Selbstgefährdung. Auf der Basis der unterschiedlichen Beschreibungen wird in dieser Studie unter Selbstgefährdung verstanden:

- Verherrlichung von bzw. Aufforderung zu gesundheitsschädigenden Handlungen: Pro-Ana/Pro-Mia, Selbstverletzung (Ritzen), Suizid, Ohnmacht,
- Verherrlichung von bzw. Aufforderung zum (übermäßigen) Genuss von Suchtmitteln, ggf. mit Handlungsanleitungen,

- Anleitungen zur Durchführung krimineller Handlungen bzw. Aufforderung zu einer Straftat.

Jeder dieser Punkte hat völlig eigene Kriterien, die in den vorliegenden Katalogen unterschiedlich ausführlich und konkret aufgeführt werden. Bei allen Punkten ist die Altersgruppe von 14 bis 18 Jahren ist besonders gefährdet. Das vorliegende Beispielmaterial legt den Schwerpunkt auf gesundheitsschädigende Handlungen. Es liegen Webseiten und Videos vor, bei denen je nach Thema vermehrt Texte oder Bilder vorkommen, zum Teil spielen auch Foren eine große Rolle. Die unterschiedlichen Themen innerhalb der Selbstgefährdung erfordern eine jeweils eigene Analyse der Daten.

3.5 Zusammenfassung

Es liegen mehrere Dokumente vor, die Kriterien zur Bewertung und Altersklassifizierung von Web-Inhalten aufzeigen. Aus den Bereichen Fernsehen/Kino und Computerspiele stammen eigene Bewertungsschemata, die zum Teil für Web-Inhalte gleichermaßen gelten.

Zum Thema Gewalt sind einheitliche Definitionen und detaillierte Kriterien verfügbar, die eine relativ sichere redaktionelle Bewertung ermöglichen. Für die Altersklassifikation ist es neben der Bewertung einzelner Elemente (z. B. eines Bildes) notwendig, größere Zusammenhänge des Angebotes (z. B. den Spannungsaufbau) zu betrachten und mit den entwicklungspsychologischen Aufnahmefähigkeiten der verschiedenen Altersgruppen abzugleichen. Die zur Verfügung gestellten Web-Inhalte im Bereich Gewalt folgen den dargestellten Kriterien. Zwar wurden auch textuelle Inhalte in diesem Bereich zur Verfügung gestellt (z. B. besonders drastische Liedtexte), allerdings spielen Bilder und Videos im Bereich Gewalt eine wichtigere Rolle, da sie den Betrachter je nach Brutalität und Detailgrad der Gewaltdarstellung nachhaltig verstören können. Die Bild- und Videodaten lassen sich prinzipiell in die Bereiche *fiktive* Gewalt und *reale* Gewalt einteilen. Im Bereich fiktive Gewalt enthielten die Videodaten insbesondere Ausschnitte aus FSK18-Filmen sowie Trailer oder Let's-Play-Videos von USK18-Spielen. Reale Gewaltbilder zeigten besonders abstoßende Schocker-Aufnahmen (z. B. von Unfällen), die zum Teil auch zur Meinungsbildung eingesetzt werden (z. B. tote Robbenbabys im Bereich Tierschutz). Reale Gewaltvideos zeigten Aufnahmen von echten, teils äußerst brutalen Gewalttaten.

Unter den Begriff Extremismus fallen im Kontext des Jugendmedienschutzes sowohl politischer als auch weltanschaulicher und religiöser Extremismus. Im Bereich des politischen Extremismus ist in den Quellen genau und relativ einheitlich beschrieben, welche Inhalte unzulässig sind. Zur Altersklassifizierung für zulässige Inhalte muss vor allem auf die verwendete Sprache, aber auch auf den Gesamtzusammenhang geachtet werden. Bei der Bewertung von Web-Inhalten bezüglich Extremismus ist stets mit dem Recht auf Meinungs- und Pressefreiheit abzuwägen. Bei den zur Verfügung gestellten jugendschutzrelevanten Daten handelt es sich ausschließlich um Daten im Bereich Rechtsextremismus. Hier spielen im Gegensatz zu den oben beschriebenen Gewaltseiten auch Texte eine große Rolle. Auf den entsprechenden Hetz- und Propagandawebseiten stehen umfangreiche extremistische Texte bereit, die sich prinzipiell für eine automatische Textklassifikation eignen. Allerdings gibt es auch eine Reihe von relevanten multimedialen Inhalten. Dazu gehören Propaganda-Berichte und Interviews, aber auch Videos zu rechtsextremer Musik. Zu den wiederkehrenden visuellen Eigenschaften zählt insbesondere eine häufig verwendete rechtsradikale *Symbolik* (z. B. Hakenkreuze, SS-Runen oder Logos von relevanten Bands und Organisationen).

Auf der Basis der vorliegenden Quellen zum Jugendmedienschutz fallen in die Kategorie Selbstgefährdung eine Reihe sehr unterschiedlicher Themenbereiche (z. B. Pro-Ana/Pro-Mia, Selbstverletzung, Suizid, Aufforderung zum (übermäßigen) Genuss von Tabak/Alkohol/Drogen, Anleitungen zur Durchführung krimineller Handlungen bzw. Aufforderung zu einer Straftat). Jeder dieser Punkte hat

völlig eigene Kriterien, die in den vorliegenden Katalogen unterschiedlich ausführlich und konkret aufgeführt werden. In dieser Kategorie wurden der vorliegenden Studie vor allem Daten zu den Themen *Pro-Ana/Pro-Mia* und *Selbstverletzung durch Ritzen* zur Verfügung gestellt. Im Fall von Pro-Ana sind die entsprechenden Webseiten (Blogs und Foren) vom Inhalt und Aufbau her sehr homogen und bieten gute Ansatzpunkte für Methoden zur automatischen Textklassifikation. Darüber hinaus wurden vielfach wiederkehrende Thinspiration-Bilder beobachtet, diese traten sowohl in Videos als auch direkt auf Webseiten auf. Im Bereich *selbstverletzendes Verhalten durch Ritzen* wurden insbesondere wiederkehrende Bilder mit Aufnahmen von Selbstverletzungen beobachtet (insbesondere an Armen und Beinen), ebenfalls gleichermaßen in Webseiten und Videos eingebettet.

Der gemäß der dargestellten Kriterien ausgewählte Datensatz deckt naturgemäß nur einen Teil aller möglichen jugendschutzrelevanten Daten ab. Trotz der großen Heterogenität möglicher Web-Inhalte (insbesondere im Bereich Gewalt) bilden die Daten jedoch eine gute Basis zur Identifikation von wiederkehrenden Eigenschaften in den einzelnen Inhaltskategorien, welche dann potentiell von automatischen Verfahren detektiert werden können. In den nun folgenden Kapiteln werden solche Verfahren zur automatischen Detektion der beschriebenen Inhalte und wiederkehrenden Charakteristika ausgewählt, beschrieben und bewertet. Eine besondere Herausforderung liegt hierbei in der automatischen Altersklassifikation, da die redaktionellen Kriterien für diese feinere Kategorisierung abstrakter sind und größere Zusammenhänge mit einbeziehen (z. B. den Spannungsbogen des Videos).

TEIL II

TECHNISCHE VERFAHREN IM JUGENDMEDIENSCHUTZ

KAPITEL 4

ANALYSE TECHNISCHER VERFAHREN ZUM JUGENDMEDIENSCHUTZ

Die größte Herausforderung im technischen Jugendmedienschutz besteht darin, für einen großen Anteil relevanter Daten herauszufinden, ob beziehungsweise in welchem Umfang diese als jugendschutzrelevant einzustufen sind. Anhand dieser Information können dann leicht Blacklists für problematische Dokumente erstellt werden, die durch eine geeignete Vorgehensweise dem Zugriff Jugendlicher entzogen werden können.

Da hier eine gewaltige Datenflut zu bewältigen ist, die zudem ständig wächst und auf viele unterschiedliche Wege zugreifbar ist, erscheint eine Lösung ohne technische Hilfsmittel hoffnungslos. Daher wird in diesem Kapitel der Stand der Technik ausgewählter Verfahren dargestellt. Bereits klassifizierte Daten lassen sich dazu verwenden, zunächst Verfahren zu identifizieren, die besonders vielversprechend für diese Aufgabe sind. Dies erfolgt in Abschnitt 4.1 auf Basis von Webseiten, die gemäß der Klassifikation durch Experten jugendschutzrelevante Inhalte in Wort, Ton oder Bild enthalten.

Wir beginnen danach zunächst die Untersuchung dieser Verfahren mit einigen allgemeinen Grundlagen des maschinellen Lernens (vgl. Abschnitt 4.2) und den notwendigen Techniken zur Vorverarbeitung von Webseiten (vgl. Abschnitt 4.3), die es ermöglichen, die für die Analyse geeigneten Bestandteile zu identifizieren. Dann diskutieren wir die wichtigsten Verfahren getrennt nach den drei im Internet relevanten Informationstypen:

1. Textanalyse: Ein Großteil der Informationen im Netz besteht entweder aus Text oder ist in einen Textkontext eingebettet. Textuelle Informationen liefern den deutlichsten Hinweis auf jugendschutzrelevantes Material, da hier die Informationen bereits explizit sind. In Abschnitt 4.4 werden moderne Verfahren zur Analyse von textuellen Daten vorgestellt.
2. Multimedia-Analyse: Bei dem untersuchten jugendschutzrelevanten Material haben Bilder und Videos meist den verstörendsten Effekt. Obwohl sie mitunter von hinreichend aussagekräftigem beschreibenden Text begleitet sind, finden sie sich, zum Beispiel auf Videoplattformen, auch häufig isoliert oder in einen eher unbedenklichen Kontext eingebettet. Daher ist es wichtig, solches Material auch anhand des audio-visuellen Inhalts erkennen zu können. Dies kann mit Hilfe von Techniken der Multimedia-Analyse geschehen (vgl. Abschnitt 4.5).
3. Umfeldanalyse: Durch die starke Verlinkung ist das Internet keine lose Sammlung von Daten, sondern hoch strukturiert. Diese Struktur kann ausgenutzt werden, um die Natur online verfügbarer Daten zu erschließen. So kann zum Beispiel die Tatsache, dass eine bestimmte Webseite von vielen als jugendschutzrelevant bekannten Seiten verlinkt wird, einen deutlichen Hinweis auf ihren ebenfalls jugendschutzrelevanten Inhalt geben. Techniken zur Umfeldanalyse von Webseiten anhand der Linkstruktur werden in Abschnitt 4.6 untersucht.

Schließlich gibt Abschnitt 4.7 anhand eines Beispielsystems eine Vorstellung, wie die in diesem Kapitel vorgestellten Einzeltechniken in einem System zusammengeführt werden können.

4.1 Auswahl technischer Verfahren auf Grundlage von redaktionell klassifizierten Daten

Bei einer ersten Sichtung der von jugendschutz.net zur Verfügung gestellten Daten wurde schnell deutlich, dass eine getrennte Betrachtung der inhaltlichen Kategorien (Gewalt, Extremismus, Selbstgefährdung)[1] sinnvoll erscheint. Während sich manche Techniken zwar prinzipiell kategorienübergreifend anwenden lassen, sind dabei in den meisten Fällen doch spezielle Eigenschaften der einzelnen Kategorien gesondert zu berücksichtigen.

In diesem Abschnitt werden zunächst zwei grundsätzlich verschiedene Ansätze zur Detektion jugendschutzrelevanter Web-Inhalte motiviert: die erneute Detektion von bereits bekannten Dokumenten (zum Beispiel bereits blockierte Videoclips, die auf einer neuen Plattform wieder auftauchen), sowie die Detektion unbekannter Dokumente, die bestimmte jugendschutzrelevante Merkmale aufweisen. Im Anschluss werden nacheinander mögliche Verfahren zur Detektion von Daten aus den Bereichen Gewalt (Abschnitt 4.1.2), Extremismus (Abschnitt 4.1.3) und Selbstgefährdung (Abschnitt 4.1.4) beschrieben.

4.1.1 Unterscheidung: Detektion identischer und ähnlicher Dokumente

In der digitalen Welt ist das »Spiegeln«, also das Kopieren von Web-Inhalten auf andere Webadressen, sehr einfach. Ein solches Vorgehen ist zum Beispiel bei Schockerseiten häufig vorzufinden. Viele dieser Seiten werden ab einem gewissen Bekanntheitsgrad wieder gelöscht. Dabei sind die Inhalte der gelöschten Seiten jedoch nicht aus dem Internet verschwunden, sondern zumeist in Zwischenspeichern oder auf Spiegelseiten weiterhin verfügbar (insbesondere beim sogenannten Streisand-Effekt). Aus Sicht des Menschen handelt es sich bei den gespiegelten Web-Inhalten um eine identische Seite, aus Computersicht können sich aber kleine Aspekte verändern. Auf solche kleinen Änderungen müssen automatisierte Verfahren robust reagieren. Zum Beispiel könnte ein jugendschutzrelevanter Text auf zwei optisch und inhaltlich sonst völlig verschiedenen Webseiten inhaltlich identisch enthalten sein. Bei Multimediamaterial, also etwa bei Bildern oder Videos, kann zum Beispiel die Größe verändert worden sein. Auch können die relevanten Webobjekte unkommentiert in neuem Kontext vorkommen, etwa als Zusammenschnitt der »besten« Szenen eines Horrorfilms. Der Kern des Web-Inhalts bleibt dabei jedoch der Gleiche. In einem solchen Fall sprechen wir im Folgenden von Identität.

Im Gegensatz hierzu beschreibt ein Merkmal einen einzelnen Aspekt eines Web-Inhalts, der auf einen jugendschutzrelevanten Inhalt hindeutet. Hierzu zählt zum Beispiel eine charakteristische Wortwahl und Ausdrucksweise in Texten, oder das Auftreten bestimmter Symbole in Bildern und Videos (zum Beispiel Häufung nationalsozialistischer Symbole in rechtsextremen Musikvideos). Auf Basis solcher Merkmale kann dann automatisch entschieden werden, ob ein unbekanntes Dokument als potentiell jugendschutzrelevant einzustufen ist.

Beispiel Es sei ein Musikvideo mit rechtsextremem Inhalt bereits als jugendschutzrelevant identifiziert. Wenn nun ein neues Video überprüft wird, das das gleiche Lied enthält, aber verrauscht klingt und mit einem 5-sekündigen Vorspann versehen wurde, so sollte dies trotzdem als Kopie des bereits identifizierten Videos detektiert werden (Identität). Handelt es sich bei einem Video jedoch nicht um das gleiche, sondern um ein anderes Lied der gleichen Band, so kann zumindest auf einzelne gemeinsame Merkmale zurückgegriffen werden. Ein mögliches Merkmal wäre in diesem Fall die wiedererkannte Stimme des Sängers. Für ein noch völlig unbekanntes Video mit rechtsextremer Musik könnten zum Beispiel folgende Merkmale herangezogen werden: Frakturschrift auf dem CD-Cover,

[1] Die Auswahl dieser Themen wird im Vorwort dieser Studie motiviert.

eingeblendete Bilder von Adolf Hitler, Bilder der Band mit Hitlergruß oder nationalsozialistische Symbolik. Eine Häufung von vielen solchen Merkmalen legt den Verdacht nahe, dass das Video in eine jugendschutzrelevante Kategorie fällt.

In den folgenden Abschnitten werden nun die von jugendschutz.net zur Verfügung gestellten Daten noch einmal aus technischer Sicht betrachtet, um mögliche Verfahren zur Detektion jugendschutzrelevanter Web-Inhalte im Text- und Multimediabereich zu identifizieren.

Im Bereich der Klassifikation unbekannter Texte gibt es verschiedene Arten von generellen vielversprechenden Techniken, um aus Webseiten Rückschlüsse auf ihre Jugendschutzrelevanz zu ziehen. Diese Techniken, die im weiteren Verlauf des Kapitels detaillierter beschrieben und bewertet werden, können in allen drei behandelten Inhaltskategorien eingesetzt werden. In den folgenden Abschnitten wird für den Bereich Textanalyse (vgl. Abschnitt 4.4) insbesondere dargestellt, welche zusätzlichen Besonderheiten und Regularitäten in den jeweiligen Inhaltsbereichen beobachtet und gegebenenfalls von entsprechenden Verfahren ausgenutzt werden können.

Im Bereich Multimedia-Analyse (vgl. Abschnitt 4.5) werden insbesondere Möglichkeiten des Einsatzes von Fingerprinting-Verfahren, die es erlauben leicht veränderte Kopien von Multimedia-Inhalten anhand sogenannter Fingerabdrücke zu identifizieren, geprüft. Außerdem werden für die einzelnen Bereiche wiederkehrende Symbole und andere visuelle Eigenschaften identifiziert, die auf eine Jugendschutzrelevanz hindeuten.

4.1.2 Auswahl von Verfahren zur Detektion von Web-Inhalten im Bereich Gewalt

Wir betrachten in diesem Abschnitt die Verfahren, die potentiell in der Lage sind, Web-Inhalte im Bereich Gewalt zu detektieren.

4.1.2.1 Textanalyse

Zusätzlich zu universell einsetzbaren Klassifikationsverfahren (vgl. Abschnitt 4.4.2) könnte im Bereich Gewalt gegen Menschen zusätzlich der emotionalere Ton von jugendschutzrelevanten Seiten erkannt werden. Dazu gehören die Erkennung von Imperativen (»Schaut Euch das mal an!«) und die Analyse von Kommentaren zu Videos (»Das hat bestimmt weh getan.«).

Ähnliches gilt für den Bereich Gewalt gegen Tiere. Auch hier spielt der emotionale Ton sowie der Imperativ eine wichtige Rolle. Der Fokus könnte darauf gerichtet werden, den subjektiveren Ton einer Aktivistenwebseite vom objektiveren Ton einer Nachrichtenseite zu trennen, um Indizien auf verdächtige Seiten zu erhalten. Die Annahme dahinter ist, dass redaktionell erarbeitete Nachrichtenartikel auf die Verwendung sogenannter Schockbilder im Allgemeinen verzichten.

4.1.2.2 Multimedia-Analyse

Generell ist bei den zur Verfügung gestellten Gewaltvideos festzuhalten, dass die Anzahl der Stammvideos, also des Urmaterials, das die Basis für die hochgeladenen Filme und Bilder bildet, begrenzt erscheint. Zwar gibt es natürlich eine immense Anzahl an Spielfilmen, die Gewaltszenen beinhalten, dennoch ist es nicht jedem Besitzer einer Kamera ohne Weiteres möglich, einen neuen Film mit aufwändigen Spezialeffekten zu erstellen und im Netz zur Verfügung zu stellen. Bei realer Gewalt ist bei

besonders drastischen Formen die Zahl der vorhandenen Videos ebenfalls überschaubar. Enthauptungsvideos beispielsweise stellen insgesamt gesehen nur eine kleine Anzahl der zu betrachtenden Videos.

Da die prominentesten Vertreter der Kategorie Gewaltvideos daher bereits bekannt sein dürften, ist es die Hauptaufgabe von technischen Verfahren, neu hochgeladene Videos auf diese Inhalte zu überprüfen. Dies wird mit Fingerprinting-Verfahren erreicht. Diese Verfahren speichern eindeutige Charakteristika eines Videos, die auch dann auftreten, wenn das Video leicht verändert worden ist, etwa durch eine andere Bildauflösung oder eine stärkere Kontrastierung der Farben. Fingerprinting-Verfahren gibt es sowohl für die Audio- als auch für die Videospur. Diese sind bei Gewaltvideos und Spielfilmen gleichermaßen wichtig, da die drastische Wirkung zumeist von den Bildern ausgeht und der Ton sich oftmals ändert. Beispiele hierfür sind unterschiedliche Synchronfassungen bei Spielfilmen oder die musikalische Unterlegung bei Gewaltfilmen aus makabren Gründen. Ebenfalls mit Fingerprints erfassbar sind Ausschnitte oder Compilations, wenn also »die besten« Szenen eines Films zusammengelegt wurden oder mehrere kleinere Gewaltfilme, zum Beispiel Prügelszenen, aneinandergefügt werden.

Weiter gefächert in Zahl und Form sind solche Daten, die ohne großen Aufwand von Nutzern erstellt werden können. Bei Gewaltvideos sind dies beispielsweise Happy Slapping Videos. In diesem Fall die Videos einzeln zu markieren erscheint deutlich schwieriger, allerdings ist davon auszugehen, dass auch hier die prominentesten Vertreter ihren Weg als Compilation in die Videoportale finden werden. Über diese quasi vorsortierten Videozusammenstellungen kann man, wiederum mit Hilfe von Fingerprints, bei sicher erkannten Teilvideos die neuen Daten zur Menge der bekannten jugendschutzrelevanten Inhalte hinzunehmen und so Schritt für Schritt eine Datenbank von Material aus dieser Kategorie aufbauen. Übertragen lässt sich dieses Vorgehen für Videos auch auf entsprechende Bilder aus der Kategorie Gewalt.

Einen weiteren Bereich in dieser Kategorie nehmen die sogenannten »Let's Play«-Videos ein. Hier werden von Nutzern Aufzeichnungen von Verläufen von Computerspielen hergestellt und kommentiert. Ein konventionelles Fingerprinting wird nur dann greifen können, wenn bestimmte wiederkehrende Szenen von hinreichender Dauer (mindestens einige Sekunden) enthalten sind. Das können etwa sogenannte Cut-Scenes sein, also Abschnitte im Spiel, die vorgegebene Filmsequenzen verwenden.

Neben Fingerprinting kann auch die Logoerkennung die Klassifikation unterstützen, da viele der Videos eindeutig gekennzeichnet sind. Bei offiziellen Spielfilmen wird beispielsweise die Altersfreigabe der FSK zum Teil am Anfang eingeblendet und kann so erkannt werden. Gleiches gilt für offizielle, bereits eingestufte Teaser und Trailer von Gewaltspielen. Ebenfalls für Gewaltspiele geeignet erscheint eine Erkennung von typischen Bedienelementen auf dem Bildschirm, etwa das Fadenkreuz, die gewählte Waffe am unteren Rand des Bildes oder die Munitionsanzeige.

Bei Videos mit realer Gewalt wird häufig ein Vorspann der Webseite vorangestellt, die das Video zur Verfügung gestellt hat. Dieser liefert insbesondere bei einschlägigen Webseiten einen wichtigen Hinweis. Oft wird auch die URL der entsprechenden Seite über Teile des Bildes gelegt. Zwar wäre hier auch Texterkennung denkbar, allerdings ist die URL oft stilisiert, so dass Logoerkennung sinnvoller erscheint. Bei Bildern werden manche besonders drastische Exemplare für Collagen verwendet. Wenn diese nur einen Teil des neuen Bildes ausmachen, ist hier Fingerprinting nicht das Mittel der Wahl, sondern ebenfalls die Logoerkennung, insbesondere dann, wenn das Bild zusätzlich stark geändert wurde (etwa rotiert/gespiegelt).

Tabelle 4.1
Technische Verfahren zur Erkennung von Multimedia-Beiträgen mit Gewalt.

	Fiktive Gewalt		Reale Gewalt	
Verfahren	Filme	Spiele	Video (Gewalttat)	Foto (Gewalttat)
Fingerprinting (Abschnitt 4.5.1)	Film Trailer Compilation Ausschnitte	Trailer Cut-Scenes	Video Neu vertontes Video Compilation	Bild Ausschnitt
Logoerkennung (Abschnitt 4.5.4)	FSK/PEGI	USK/PEGI spezifische Controls	Webseiten-Primer URL	URL Collage
Spracherkennung (Abschnitt 4.5.2)		Let's Play		

Bei den »Let's Play« Videos kann auch die automatische Spracherkennung zielführend sein. Aufgabe der Spracherkennung ist es dann, den Namen und/oder den Kontext des Spiels automatisch anhand des gesprochenen Inhalts zu ermitteln.

Tabelle 4.1 stellt die möglichen Verfahren und ihre Ziele kurz zusammen. Die Verfahren werden in Abschnitt 5.3 qualitativ und quantitativ bewertet.

4.1.3 Auswahl von Verfahren zur Detektion von Web-Inhalten im Bereich Extremismus

Wir betrachten in diesem Abschnitt die Verfahren, die potentiell in der Lage sind, Web-Inhalte im Bereich Extremismus zu detektieren.

4.1.3.1 Textanalyse

Bei der Klassifikation extremistischer Inhalte bezüglich ihrer Jugendschutzrelevanz geht es vor allem darum, diejenigen Texte, die sich kritisch mit der Materie auseinandersetzen von jenen zu unterscheiden, die die Zusammenhänge einseitig oder verherrlichend darstellen. Beide Arten von Texten bedienen sich häufig einer ähnlichen Terminologie, so dass der Einsatz einfacher Schlagwort-basierter Ansätze nicht ausreichend sein wird, um die politische Ausrichtung und Intention der Inhalte ermitteln zu können.

Es liegt die Annahme nahe, dass statt der reinen Auswahl und Anzahl der verwendeten Wörter die gewählte Reihenfolge besser zu einer Trennbarkeit der Dokumente beitragen kann, da auf diese Weise der transportierte Gehalt der Nachricht besser erfasst wird. Dabei wird eine linguistische Analyse durch das Vorkommen wörtlicher Zitate mit extremistischem Inhalt erschwert. Es lässt sich aber vermuten, dass sich, verglichen mit allgemeinem Textmaterial, in tatsächlich jugendschutzrelevanten Texten entsprechende Phrasen signifikant häufen.

Eine weitere mögliche Hilfe zur maschinellen Themenerkennung ist die Erkennung von Eigennamen von Personen, Orten oder Organisationen, die sogenannte Named Entity Recognition (NER). Alleine

über eine Liste der im Text enthaltenen Entitäten lässt sich ein Bild des beschriebenen Inhalts gewinnen. Darüber hinaus bietet dieser Ansatz viele weitere Indikatoren. Dazu bietet sich einerseits der Kontext der Namensnennung an. Andererseits kann in Verbindung mit der Disambiguierung der Entitäten, also der Zuordnung der gefundenen Namen zu einer von mehreren gleichnamigen Entitäten, durch Verknüpfung mit entsprechenden Wikipedia-Artikeln Zusatzinformation gewonnen werden.

Ab einem Umfang von etwa mehreren tausend Webseiten zum Thema Rechtsextremismus ließen sich vermutlich automatisch entsprechende Merkmale extrahieren. Dabei werden die relativen Häufigkeitsverteilungen der Phrasen in diesen Dokumenten mit jenen aus einer inhaltlich unbedenklichen Textkollektion verglichen.

Als ein weiteres wichtiges Kriterium stellt sich die Anzahl der nach *Chilling Effects*[2] bei Google als illegal zurückgehaltenen, auf die Zielseite verweisenden Webseiten dar. Die Liste dieser zurückgehaltenen Webseiten basiert im Wesentlichen auf Meldungen rechtlich bedenklicher Inhalte, zum Beispiel auf den Listen der BPjM. Eine Google-Suche nach »link:<URL>« liefert Webseiten, die Links auf die gesuchte Webseite enthalten. Sind diese als illegal bekannt, erscheint außerdem der Hinweis: »Aus Rechtsgründen hat Google X Ergebnis(se) von dieser Seite entfernt.« In einzelnen Proben aus dem Bereich »Rechtsextremismus« wurden auf diese Weise bei sehr vielen als jugendschutzrelevant markierten Webseiten bis zu hundert Links aus dem Suchergebnis gesperrt. Diese Größenordnung ist weit überdurchschnittlich und somit ein guter Indikator für die automatische Erkennung. Weitere Informationen zur allgemeinen Umfeldanalyse von Webseiten findet sich in Abschnitt 4.6, während das hier vorgeschlagene Vorgehen in Abschnitt 5.4.4 qualitativ evaluiert wird.

4.1.3.2 Multimedia-Analyse

Wie auch schon beim Thema Gewalt erscheinen im Kontext von extremistischen Videos und Bildern Fingerprinting-Verfahren äußerst vielversprechend. Zum Beispiel gibt es einige bekannte geschichtsrevisionistische Videoreportagen, die aufgrund des Aufwandes der Herstellung nur in überschaubarer Anzahl vorkommen. Da die Qualität erheblich schwanken kann und mehrere Versionen zirkulieren, können hier robuste Fingerprinting-Verfahren die Suche gezielt erleichtern. Bei solchen Videos erscheint, stärker als bei Gewaltvideos, generell die Audiospur wichtiger, da hier der Hauptinhalt liegt. Dies gilt insbesondere für extremistische Musik, die von den Fans oft mehrfach ins Netz gestellt wird, um einer Löschung entgegenzuwirken. Die Musik wird zum Teil auch in den oben genannten Reportagen eingeblendet und kann daher auch bei unbekannten Filmen wertvolle Informationen liefern.

Aber auch auf der visuellen Ebene sind viele wiederkehrende Bilder vorhanden, die eine relativ eindeutige Zuordnung erlauben. So sind etwa bei Musik oft die Cover der zugrundeliegenden CDs oder Fotos der Band zu sehen. Häufig finden sich auch Bilder von verehrten Märtyrern der Szene. Neben historischen Schwarz-Weiß-Aufnahmen aus dem dritten Reich sind dies auch populäre Vertreter aus der Neonazi-Szene wie etwa der 1993 verstorbene Ian Stuart Donaldson (vgl. Abschnitt 5.4.5).

Auf der Ebene von Logos ist insbesondere eine Detektion von eindeutigen, zum Teil strafrechtlich relevanten Symbolen wie etwa dem Hakenkreuz, der SS-Rune oder der Triskele des in Deutschland verbotenen Netzwerks Blood & Honour interessant. Insbesondere in der rechtsextremen Szene werden bestimmte Wörter bewusst zur Verschleierung vermieden und durch Abkürzungen oder Zahlencodes ersetzt. Diese sind häufig stilisiert und verziert, so dass eine einfache Texterkennung hier nicht zielführend ist und statt dessen ebenfalls eine Logoerkennung sinnvoll erscheint.

[2] siehe http://www.chillingeffects.org, Stand: 16.11.2012

Tabelle 4.2
Technische Verfahren zur Erkennung von extremistischen Multimedia-Inhalten

Verfahren	Musik	Interviews	Dokumentation	UGC
Fingerprinting (Abschnitt 4.5.1)	Lieder CD-Cover Band-Fotos »Märtyrer«-Bilder	bekannte Filme unterlegte Musik	Video Compilation	unterlegte Bilder
Logoerkennung (Abschnitt 4.5.4)	Symbole Zahlencodes			Symbole Zahlencodes
Spracherkennung (Abschnitt 4.5.2)		Inhaltsanalyse		Inhaltsanalyse
Texterkennung (Abschnitt 4.5.3)	Slogans Namen	Bauchbinden URLs Slogans Zitate		eingeblendeter Text
Sonstiges	Personenerkennung Hitlergruß	Sprechererkennung		

Texterkennung in Bildern ist vielmehr dann sinnvoll, wenn es um längere Textpassagen geht, etwa bei eingeblendeten Untertiteln (die z.T. auch bei Musikstücken vorkommen), eingeblendeten Slogans oder Zitaten.

Prinzipiell möglich, aber mit Vorsicht zu betrachten ist die Personenerkennung, etwa bei wissenschaftlich wirkenden Propaganda-Reportagen oder bei Interviews mit dem Rechtsextremismus nahestehenden Personen. Eine Personenerkennung kann durch Gesichtserkennung, durch Texterkennung von eingeblendeten Bauchbinden sowie durch Sprechererkennung erfolgen. Allerdings ist insbesondere bei dieser Herangehensweise zu prüfen, inwieweit die jeweilige Aussage inhaltlich als extremistisch zu bewerten ist. Um auf inhaltlicher Ebene weitere Merkmale zu erlangen, ist je nach Qualität der Aufnahme eine automatische Spracherkennung denkbar.

Vorstellbar ist auch die automatische Erkennung von relevanten Posen wie etwa dem Hitlergruß.

4.1.4 Auswahl von Verfahren zur Detektion von Web-Inhalten im Bereich Selbstgefährdung

Wir betrachten in diesem Abschnitt die Verfahren, die potentiell in der Lage sind, Web-Inhalte im Bereich Selbstgefährdung zu detektieren.

4.1.4.1 Textanalyse

Generell ist der Bereich Selbstgefährdung in bestimmte Unterbereiche gegliedert, zum Beispiel Pro-Ana/Pro-Mia oder Suizid. Für alle diese Unterbereiche müssen angepasste Techniken entwickelt werden. Dies erscheint aus mehreren Gründen sinnvoll. Erstens wird häufig eine spezielle, in der Allgemeinheit wenig verbreitete Sprache verwendet. Zweitens gibt es in jedem dieser Unterbereiche

bestimmte Standardtexte, die referenziert werden. Drittens wird häufig mit einem Schema aus Belohnung und Strafe gearbeitet. Viertens können strukturelle Aspekte, wie zum Beispiel eine Tagebuchstruktur, analysiert werden.

Im Folgenden werden die einzelnen Unterbereiche genauer betrachtet. Bei Verherrlichung von Alkohol- und Drogenkonsum gestaltet sich die Erkennung vermutlich etwas schwieriger, da diese Themen auf vielen Seiten und insbesondere auf allgemeinen Nachrichtenseiten sehr präsent sind. Auch ist das Vokabular (»Komasaufen«, bestimmte Drogenarten) auf nicht-jugendschutzrelevanten Seiten durchaus ähnlich. Allerdings könnte man über die Identifikation von Anleitungen und Tipps zur Anwendung von Alkohol und Drogen entsprechende Seiten erkennen.

Einfacher stellt sich die Erkennung von Pro-Ana/Pro-Mia-Inhalten dar, da diese häufig mit quasi identischen Disclaimern oder Briefen von Ana oder Mia versehen sind. Also ganz konkret: »Dies ist eine Pro Ana [sic] Homepage, die mir helfen soll abzunehmen und auf der ich Euch etwas über mich erzählen werde. Warnung: Die Inhalte können Triggern! Wenn ihr nicht wisst, was Pro Ana ist, gerade in Therapie seid oder etwas gegen Pro Ana habt, dann verlasst diese Seite bitte.« Oder: »Erlaube mir, mich vorzustellen. Mein Name, oder wie ich von so genannten ›Ärzten‹ genannt werde, ist Anorexie […], aber du kannst mich Ana nennen. Ich hoffe, wir werden gute Freunde. In der nächsten Zeit werde ich viel Zeit in dich investieren und ich erwarte das Gleiche von dir. […] Du darfst nicht viel essen. Es wird langsam anfangen […] Aber es wird nicht lange dauern, dann werde ich dir sagen, dass das nicht genug ist. […] Ich werde dich an deine Grenzen treiben.« Desweiteren enthalten diese Seiten Gebote und Glaubensbekenntnisse (»Du bist NIE zu dünn!«). Auch Extremformen (»ATTE (Ana till the end)«) sind über einheitliche Klauseln identifizierbar. Zu beachten ist allerdings wie auch in den anderen Bereich, dass ggf. entsprechende Berichterstattung in Medien möglich ist, die identische Begriffe verwendet. Daher sollten einzelne Techniken nicht allein eingesetzt werden, sondern mit anderen Techniken, wie zum Beispiel der Bilderkennung, kombiniert werden.

In den Unterbereichen des selbstverletzenden Verhaltens und der Würgespiele verhält es sich ähnlich wie bei Alkohol und Drogen: Es gibt auch viele legale Seiten, zum Beispiel mit Hilfsangeboten oder Betroffenenberichten. Pro-Selbstverletzung-Seiten und Pro-Würgespiele-Seiten enthalten aber wieder Anleitungen und Methodendiskussionen. Außerdem wird zum Beispiel Ritzen als Lifestyle verherrlicht und verharmlost.

Im Unterbereich Suizid gibt es eine große Bandbreite an Foren, Communities und Informationsangeboten. Sie reicht von ausschließlich lebensbejahenden Inhalten zur Unterstützung in Lebenskrisen bis hin zu lebensverneinenden, prosuizidalen Inhalten. Dabei kommen auch zahlreiche Mischformen vor. Bei den lebensbejahenden Inhalten gelten üblicherweise strikte Regeln, die es insbesondere verbieten, Ratschläge oder Anleitungen zur Durchführung des Suizids zu veröffentlichen. Ein erkennbares Merkmal sind Links zu bekannten Therapieeinrichtungen oder professionellen Beratern. Des Weiteren ist der Zugang zu den Hilfsangeboten wesentlich einfacher und offener gestaltet. Demgegenüber zeichnen sich lebensverneinende Webseiten durch einen kleinen öffentlichen Bereich sowie einen restriktiven Zugang zu den Hauptinhalten aus. Hier sind Suizidanleitungen sowie Verabredungen zum Suizid die Regel; es herrscht ein destruktiver Grundton.

4.1.4.2 Multimedia-Analyse

Häufiger als in den Bereichen Gewalt oder Extremismus finden sich bei Inhalten zu selbstgefährdendem Verhalten selbst erstellte Videos, die Privatpersonen oder gestalterisch veränderte Bilder thematisch passender Vorlagen zeigen. Hier reicht es daher nicht, die Inhalte mit bereits bekannten Videos zu vergleichen; vielmehr muss auf feingranularer Ebene nach bestimmten wiederkehrenden Elementen gesucht werden. Dies sind in vielen Fällen wiederkehrende Bilder von Personen oder Ereignissen,

Tabelle 4.3
Technische Verfahren zur Erkennung von Multimedia-Inhalten im Bereich Selbstgefährdung. Würgespiele sowie Alkohol- und Drogenverherrlichung stellen sich als zu heterogen dar, um Gemeinsamkeiten ableiten zu können.

Verfahren	Pro-Ana/Pro-Mia	Selbstverletzung	Selbstmord
Fingerprinting (Abschnitt 4.5.1)	Lieder Thinspos	Lieder Ritzbilder	Trigger
Logoerkennung (Abschnitt 4.5.4)			Webseitenlogo
Spracherkennung (Abschnitt 4.5.2)			
Texterkennung (Abschnitt 4.5.3)	Text in Videos	Text in Videos	
Sonstiges	Konzepterkennung Anorexie	Konzepterkennung Selbstverletzung	

die von den betroffenen Personen verherrlicht oder als besonders abstoßend empfunden werden. Für Pro-Ana/Pro-Mia sind dies insbesondere schlanke bis dürre Personen, die allerdings nur in Extremfällen eindeutig sind und zum Teil auch in Diät-Magazinen vorkommen könnten. Hier müssten also über eine aktuell gehaltene Datenbank typische Fotos durch Fingerprints identifiziert werden, um Zusammenstellung aus solchen Bildern als eindeutigen Hinweis zu werten.

Dies funktioniert jedoch kaum bei den sogenannten »Real-Thinspos«, also Bildern von sehr dünnen Menschen aus dem direkten Umfeld der Betroffenen. Diese Bilder sind einzigartig und daher, sofern sie nicht mehrfach verwendet werden, durch Fingerprints nicht zu finden. Zu beachten ist auch, dass Bilder in diesem Bereich oft umgestaltet werden, also mit hinzugefügtem Text, Sepia-Farben oder gestalterischen Elementen wie angefügten Engelsflügeln versehen werden. Auf diese Veränderung müssen automatische Verfahren robust reagieren können.

Ähnliche Überlegungen gelten auch für Bilder im Bereich der anderen Selbstgefährdungsthemen, also etwa für Ritzbilder bei selbstverletzendem Verhalten oder für sogenannte Trigger-Bilder im Bereich Suizid (zum Beispiel Fotos von Leichen nach erfolgtem Selbstmord). Sowohl bei selbstverletzendem Verhalten als auch bei Suizid-Themen gibt es zusätzlich eine Reihe von künstlerisch bearbeiteten und ebenfalls wiederkehrenden Fotos.

Bei Tagebuch-ähnlichen Videos, die von Betroffenen erstellt werden, wird häufig ein Musikstück im Hintergrund abgespielt, welches ebenfalls als zusätzliches Indiz über Audio-Fingerprints erkannt werden kann. Da die Bandbreite dieser Lieder von allgemeinen Rocksongs bis hin zu Magersucht propagierenden Liedern reicht, muss die Aussagekraft der Stücke aber gesondert untersucht werden.

Logoerkennung scheint in einigen Fällen sinnvoll zu sein, etwa bei Schockerbildern, die oft mit einem Logo der veröffentlichenden Schocker-Webseite versehen werden.

Häufig sind die Videos mit geschriebenem Text angereichert, der von persönlichen Lebensgeschichten, Anleitungen, wiederkehrenden Motivationsslogans bis hin zu Texten der unterlegten Musik reichen kann. Auch wenn manche Texte eher eine allgemein depressive Stimmung widerspiegeln, sind viele hinreichend eindeutig. Die Texterkennung muss dabei mit zum Teil stark geschwungenen Schriftarten, die über andere Bilder gelegt worden sind, umgehen können.

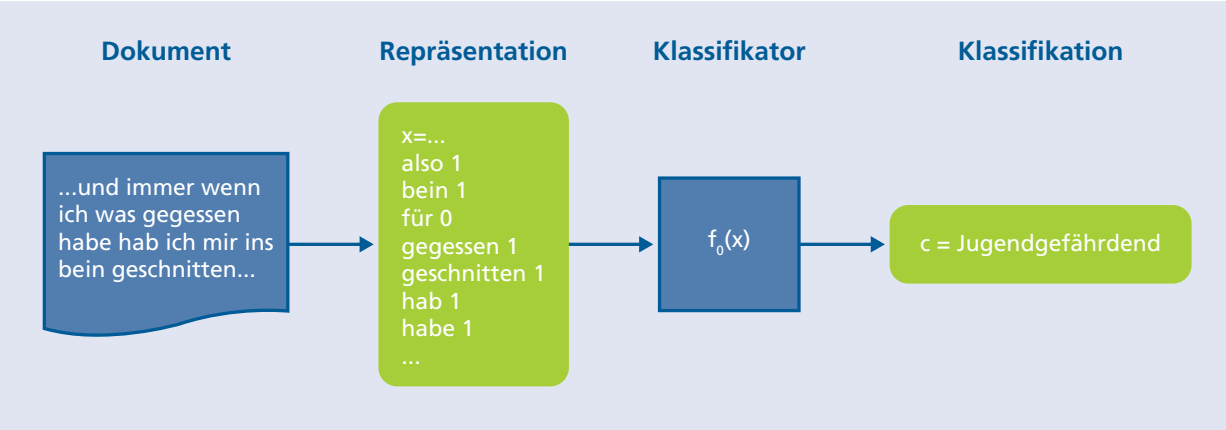


Abbildung 4.1
Informationsfluss bei der Anwendung eines Klassifikators. Das betrachtete Dokument wird in einen Vektor x überführt, der die relevanten Merkmale des Dokumentes enthält. Darauf wird die Klassifikatorfunktion angewendet, welche die wahrscheinlichste Klasse des Dokumentes berechnet.

4.2 Ziel, Vorgehensweise und Qualitätsbewertung eines Klassifikators

Verfahren der Textklassifikation [Sebastiani, 2002] und der Multimedia-Klassifikation [Troncy et al., 2011] ordnen jedes Dokument in eine vorgegebene Klasse (zum Beispiel »Jugendschutzrelevant« und »Unbedenklich«) ein. Moderne Klassifikatoren werden nicht manuell erstellt, sondern lernen aus Beispielen und können nach Beendigung der Lernphase verallgemeinern. D.h., sie lernen nicht einfach die Beispiele (mit den Klassen »Jugendschutzrelevant« und »Unbedenklich« annotierte Dokumente) auswendig, sondern »erkennen« Gesetzmäßigkeiten in den Lerndaten. So kann das System auch unbekannte Dokumente beurteilen. Derartige Verfahren des Maschinellen Lernens verwenden Techniken der Statistik und des Data Mining.

Voraussetzung für das Lernen ist eine Trainingsmenge $D = \{d_1, \ldots, d_n\}$ von Dokumenten, die schon mit den betrachteten Kategorien c (zum Beispiel »Jugendschutzrelevant«, »Unbedenklich«) manuell annotiert wurden. Ein Dokument d wird im Allgemeinen durch einen Vektor von Merkmalen $x = (x_1, \ldots, x_M)$ repräsentiert, zum Beispiel durch die Häufigkeit jedes Wortes im Dokument. Ein Klassifikator kann nun als eine Funktion, die Klassifikatorfunktion, definiert werden, welche aus x die zugehörige Klasse c berechnet.

$$f_b : x \to c$$

Abbildung 4.1 zeigt den Informationsfluss bei der Anwendung eines Klassifikators.

Die Klassifikatorfunktion f_b hat einen Vektor b von freien, reellwertigen Parametern, das heißt veränderlichen Stellgrößen. Kennt man den Wert des Parametervektors, so ist der Klassifikator eindeutig festgelegt. Der optimale Parameterwert b^* wird durch Optimierungsverfahren so berechnet, dass f_{b^*} die manuell annotierten Klassen der Dokumente d in der Trainingsmenge optimal reproduzieren kann. Dieser Optimierungsprozess wird auch als Training bezeichnet. Anschließend kann der Klassifikator auf neue, bisher unbekannte Dokumente d^{neu} aus der gleichen Domäne angewendet werden, um deren Klasse c durch $f_{b^*}(d^{neu}) = c$ zu prognostizieren.

Es gibt eine große Anzahl von Klassifikatorfunktionen. Die wichtigsten Varianten werden in den folgenden Abschnitten vorgestellt und ihre Eignung für die Klassifikation jugendschutzrelevanter Inhalte diskutiert.

4.2.1 Bewertung eines Klassifikators

Um die Qualität eines Klassifikators zu bewerten gibt es verschiedene Maße, von denen die gebräuchlichsten im Folgenden kurz vorgestellt werden sollen. Prinzipiell gibt es bei einer ja/nein Entscheidung wie bei der Frage »Beinhaltet diese Webseite Medienobjekte, die erst ab 18 sichtbar gemacht werden sollen« zwei Arten von Fehlern: der Fehler 1. Art, auch α-Fehler oder Rate der Falsch-Positiven genannt, bewertet eine Seite als »Jugendschutzrelevant«, obwohl sie es nicht ist (Overblocking). Der Fehler 2. Art, auch β-Fehler oder Rate der Falsch-Negativen genannt, stuft eine jugendschutzrelevante Seite als harmlos ein (Underblocking). Üblicherweise gibt man diese Werte als Wahrscheinlichkeit, also als Wert zwischen 0 und 1 an.

Erst die kombinierte Betrachtung beider Größen ergibt ein genaueres Bild von der Klassifizierungsperformanz, jedoch können diese Größen insbesondere bei seltenen Ereignissen auch irreführend sein. Als Beispiel betrachten wir einen medizinisch-diagnostischen Test für eine Krankheit A, unter der 80 000 Menschen der Bundesrepublik leiden. Angenommen, die Fehlerwahrscheinlichkeiten α und β liege bei erfreulichen 0,1 %. Bei einer stark gerundeten Einwohnerzahl Deutschlands von 80 Millionen würde der Test zwar nur 80 von den 80 000 kranken Menschen irrtümlich für gesund halten, aber auch 79 920 gesunde Menschen als krank fehldiagnostizieren.

Im Bereich der computergestützten Klassifizierung ist es hingegen üblich, die Trefferquote (engl.: *recall*) und die Präzision (engl.: *precision*) anzugeben. Die Präzision entspricht dem Anteil der richtigen an den positiven Diagnosen, also $1 - \alpha$. Die Trefferquote misst die Vollständigkeit der identifizierten Dokumente, also den Anteil der korrekt klassifizierten an den relevanten Dokumenten. Ein Jugendschutzfilter mit einer hohen Trefferquote bei einer geringen Präzision wird dazu neigen, vermehrt Seiten irrtümlich zu sperren, aber im Gegenzug nur wenige gefährdende Seiten auszulassen. Ist hingegen die Präzision hoch, aber die Trefferquote gering, so wird es kaum Seiten geben, die irrtümlich gesperrt werden, andererseits aber relativ viele Seiten, die trotz relevanter Inhalte nicht identifiziert werden.

Zu jeder Klassifikationsaufgabe muss man explizit angeben, wie hoch die beiden Fehlerarten gewichtet werden, das heißt wie viele Fehler 1. Art man für einen Fehler 2. Art in Kauf nehmen will. Ist für den Nutzer die Trefferquote β-mal wichtiger als die Präzision, so kann man dies mit folgender Maßzahl berücksichtigen:

$$F_\beta = \frac{(1-\beta^2) * \text{Präzision} * \text{Trefferquote}}{\beta^2 * \text{Präzision} + \text{Trefferquote}}$$

Für $\beta = 1$ erhält man den F-Wert $F = F_1$, der zur Messung der Genauigkeit des Klassifikators die Trefferquote und die Präzision als gleich wichtig betrachtet. Für den Fall, dass Präzision $=$ Trefferquote gilt auch, dass $F =$ Präzision.

4.2.1.1 Klassifikationsschwelle und »unsichere« Dokumente

Die meisten Klassifikatoren berechnen für jede Klasse einen reellwertigen Indikatorwert. Ist dieser Indikatorwert (zum Beispiel eine Wahrscheinlichkeit) über einer Klassifikationsschwelle, so wird das Dokument der Klasse zugeordnet, andernfalls wird eine Alternativklasse gewählt. Durch die Veränderung der Klassifikationsschwelle lassen sich die Präzision und die Trefferquote verändern. Der Zusammenhang wird in Abbildung 4.2 für ein Beispiel gezeigt. Bei einer hohen Klassifikationsschwelle werden nur wenige Dokumente der Klasse zugeordnet, die Präzision ist hoch, aber die Trefferquote, die Zahl der identifizierten Klassenmitglieder ist gering. Ist hingegen die Klassifikationsschwelle niedrig, so werden viele Dokumente der Klasse zugeordnet (hohe Trefferrate), wobei aber die Präzi-

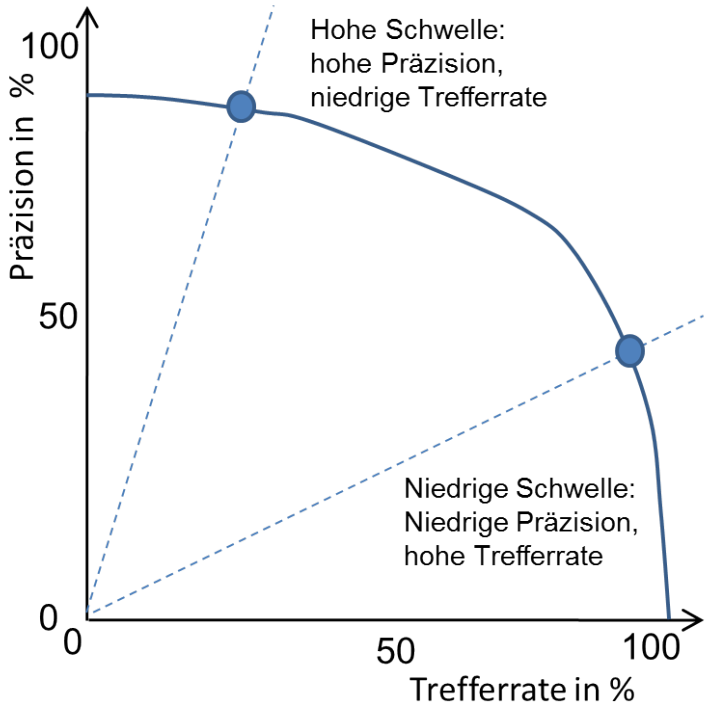


Abbildung 4.2
Präzision und Recall der Klassifikation in Abhängigkeit von der Klassifikationsschwelle. Bei einer hohen Klassifikationsschwelle ist die Präzision hoch und die Trefferrate gering. Bei einer niedrigen Klassifikationsschwelle ist die Präzision niedrig und die Trefferrate hoch.

sion nur gering ist. Durch die Auswahl der Klassifikationsschwelle kann man daher zwischen einer hochgenauen und einer weniger genauen Klassifikation wählen.

Definiert man für jede Klasse eine solche Klassifikationsschwelle mit hoher Präzision, so gibt es eine Reihe von »unsicheren« Dokumenten, welche keine dieser Schwellen erreichen und nicht klassifiziert werden. Bei der Anwendung des Klassifikators kann man mit diesen Dokumenten gesondert verfahren und sie zum Beispiel manuell einordnen.

Bei vielen Klassifikatoren wurde der Indikatorwert allein in Hinblick auf die Klassifikation optimiert und ist als Wahrscheinlichkeitswert nicht besonders zuverlässig. Alternative Indikatorwerte erfassen die Klassifikationwahrscheinlichkeiten oft mit einer höheren Zuverlässigkeit, zum Beispiel ein mit Hilfe des Bagging-Verfahrens berechneter Indikator (vgl. Abschnitt 4.4.2.7), welcher für beliebige Klassifikatoren bestimmt werden kann.

4.2.2 Überanpassung

Beim Training eines Klassifikators wird die Klassifikatorfunktion f_b so angepasst, dass in der Trainingsmenge $D = \{d_1, \ldots, d_n\}$ die korrekten Klassen $c(d_1), \ldots, c(d_n)$ der Dokumente möglichst gut reproduziert werden, zum Beispiel gemessen durch den F-Wert. Hierbei kann es allerdings geschehen, dass der Klassifikator nicht nur systematische Zusammenhänge zwischen den Eingaben x und den Klassen $c(d_i)$ erfasst, sondern auch Zufallsfehler »lernt«.

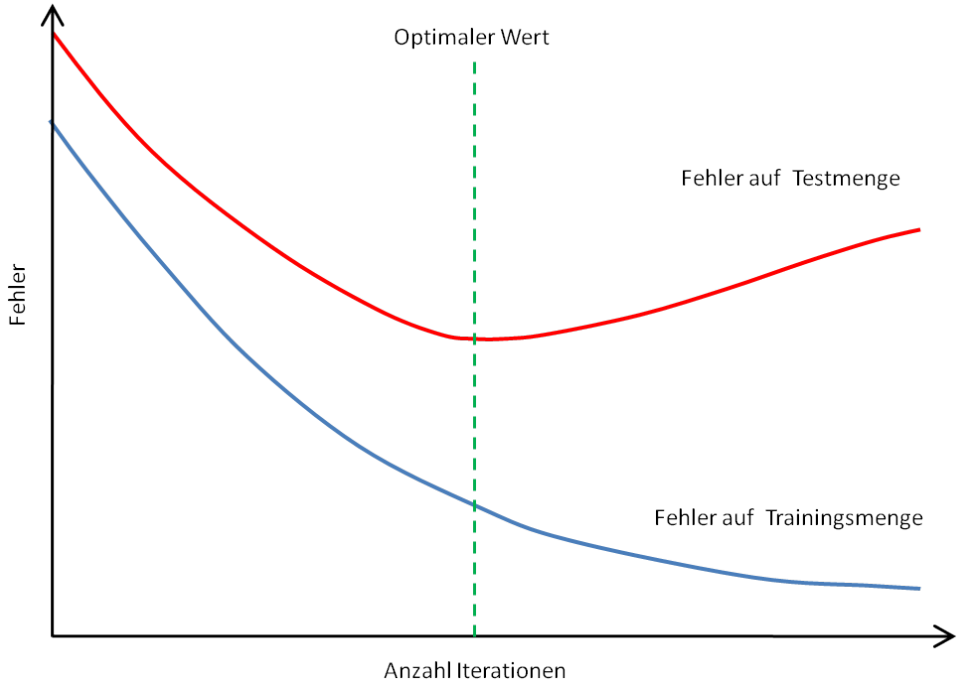


Abbildung 4.3
Fehlerwerte im Verlauf des Trainings eines Klassifikators auf der Trainingsmenge und einer unabhängigen Testmenge. Ab einer gewissen Anzahl von Iterationen kann das Ausmaß der Fehler auf der Testmenge wieder steigen.

Abbildung 4.3 zeigt für ein Beispiel das Ausmaß der Fehler im Verlauf der Optimierung auf der Trainingsmenge und auf einer unabhängigen Testmenge $\tilde{D} = \{\tilde{d}_1, \ldots, \tilde{d}_m\}$ von Dokumenten, welche nicht beim Training verwendet werden. Die Optimierung findet schrittweise statt. In jedem Schritt (Iteration) nähert sich der Parameterwert dem besten Parameterwert. Zu Beginn der Optimierung sinkt der Fehler auf beiden Datenbeständen. Nach einer gewissen Anzahl von Iterationen aber steigt der Fehler auf der Testmenge an, während er auf der Trainingsmenge noch weiter sinkt.

Das Phänomen der Überanpassung lässt sich durch ein Beispiel veranschaulichen. Gehen wir einmal von der Aufgabe aus, dass Nachrichten in die Kategorien »Sport« und »Politik« eingeordnet werden sollen und dass allein das Wort »Merkel« als Merkmal vorliegt. Stehen nun lediglich Nachrichten eines Tages als Trainingsmaterial zur Verfügung und fand an diesem Tag ein Länderspiel statt, das von Bundeskanzlerin Merkel besucht wurde, so kann es sein, dass 8 von 10 Nachrichten mit dem Wort »Merkel« in die Kategorie »Sport« eingeordnet wurden. Ein entsprechender Klassifikator würde »Merkel« als ein Indiz für Sport werten. Offenbar liegt hier eine »seltene« Konstellation der Beobachtungen vor. Stammen die Dokumente in der Testmenge von einem anderen Tag, an dem »Merkel« eher in Politiknachrichten auftaucht, so würde der Klassifikator zu einem erhöhten Fehler auf der Testmenge führen. Offenbar enthält die Trainingsmenge zu wenige Beobachtungen. Bei einem Klassifikator mit hunderttausend Merkmalen ergeben sich aber »seltene« Konstellationen auch bei vielen tausend Beobachtungen. Daher tritt dort fast immer der Effekt der Überanpassung auf, wenn man keine Gegenmaßnahmen ergreift.

Wichtig ist, dass der Fehler auf der Testmenge eine im Mittel korrekte Abschätzung des Klassifikationsfehlers auf neuen Daten erlaubt. Der Fehler auf der Testmenge ist also diejenige Größe, die ein Klassifikator eigentlich optimieren sollte. Würde die Testmenge aber bei der Optimierung verwendet, so würden die Parameter an die Testmenge angepasst und erlauben keine Abschätzung der Genau-

igkeit auf neuen Daten. Daher muss man die oben aufgeführten Bewertungsmaße immer auf einer unabhängigen Testmenge bestimmen, welche beim Training in keiner Phase verwendet wurde.

Es gibt eine Reihe von Techniken, wie der Fehler auf der Testmenge minimiert werden kann:

- Beim Early Stopping wird die Optimierung nicht bis zum Ende durchgeführt, sondern man stoppt nach einer Reihe von Iterationen. Hierbei kann man den Fehler auf einer weiteren unabhängigen Menge von annotierten Beispielen, der Validationsmenge, als ein Stoppkriterium verwenden.
- Bei der Merkmalsauswahl werden irrelevante Merkmale entfernt (vgl. Abschnitt 4.4.2.2).
- Bei der Regularisierung werden große Parameterwerte mit einem Strafterm bewertet. Der Klassifikator bevorzugt daher im Mittel kleinere Parameterwerte. Es lässt sich zeigen, dass hierdurch die Überanpassung reduziert wird (vgl. Abschnitt 4.4.2.2).
- Bei der Verringerung der Modellkapazität wird die effektive Anzahl der freien Parameter reduziert. Damit kann der Klassifikator nur noch weniger komplexe Zusammenhänge darstellen. Details werden in Abschnitt 4.4.2.5 diskutiert.

4.2.3 Kreuzvalidierung

Wie in Abschnitt 4.2.1 diskutiert, ist der auf einer unabhängigen Testmenge bestimmte Klassifikationsfehler ein erwartungstreuer Schätzwert für den zu erwartenden Klassifikationsfehler auf neuen Dokumenten. Typischerweise wird eine Testmenge von 30 % der verfügbaren manuell annotierten Dokumente verwendet, damit der geschätzte zu erwartende Klassifikationsfehler hinreichend genau ist. Allerdings verringert sich hierdurch die erreichbare Genauigkeit des Klassifikators, da ja nur noch 70 % des Trainingsmaterials verwendet wird.

Um bei der Abschätzung des zu erwartenden Klassifikationsfehlers eine möglichst große Trainingsmenge zu nutzen, wurde das Verfahren der k-fachen Kreuzvalidierung entwickelt. Wie in Abbildung 4.4 gezeigt, wird hierbei die Menge der annotierten Dokumente in k etwa gleich große Teilmengen zerlegt. Es werden anschließend k verschiedene Klassifikatoren geschätzt, wobei jeweils eine Teilmenge als Testmenge dient und die restlichen Teile zusammengenommen als Trainingsmenge verwendet werden. Für jeden Klassifikator lässt sich dann der Klassifikationsfehler auf der jeweiligen Testmenge bestimmen. Der Mittelwert dieser Klassifikationsfehler liefert dann einen erwartungstreuen Schätzwert des tatsächliche Klassifikationsfehlers. Der Schätzwert ist zuverlässiger als bei einer einzelnen Testmenge, da er ja mit Hilfe einer wesentlich größeren Anzahl von Dokumenten berechnet wurde. Eine k-fach Kreuzvalidierung wird immer dann durchgeführt, wenn nur eine beschränkte Anzahl von annotierten Dokumenten zur Verfügung steht.

4.3 Vorverarbeitung von Webseiten: Identifikation zu klassifizierender Objekte

Aus technischer Sicht sind bei der Bereitstellung von Webseiten zur weiteren Verarbeitung einige wichtige Details zu beachten. Im Gegensatz zu in sich geschlossenen und homogenen Textkorpora erhält man hier Daten aus einer Vielzahl unterschiedlicher Quellen mit einer ebenso großen Bandbreite möglicher Probleme.

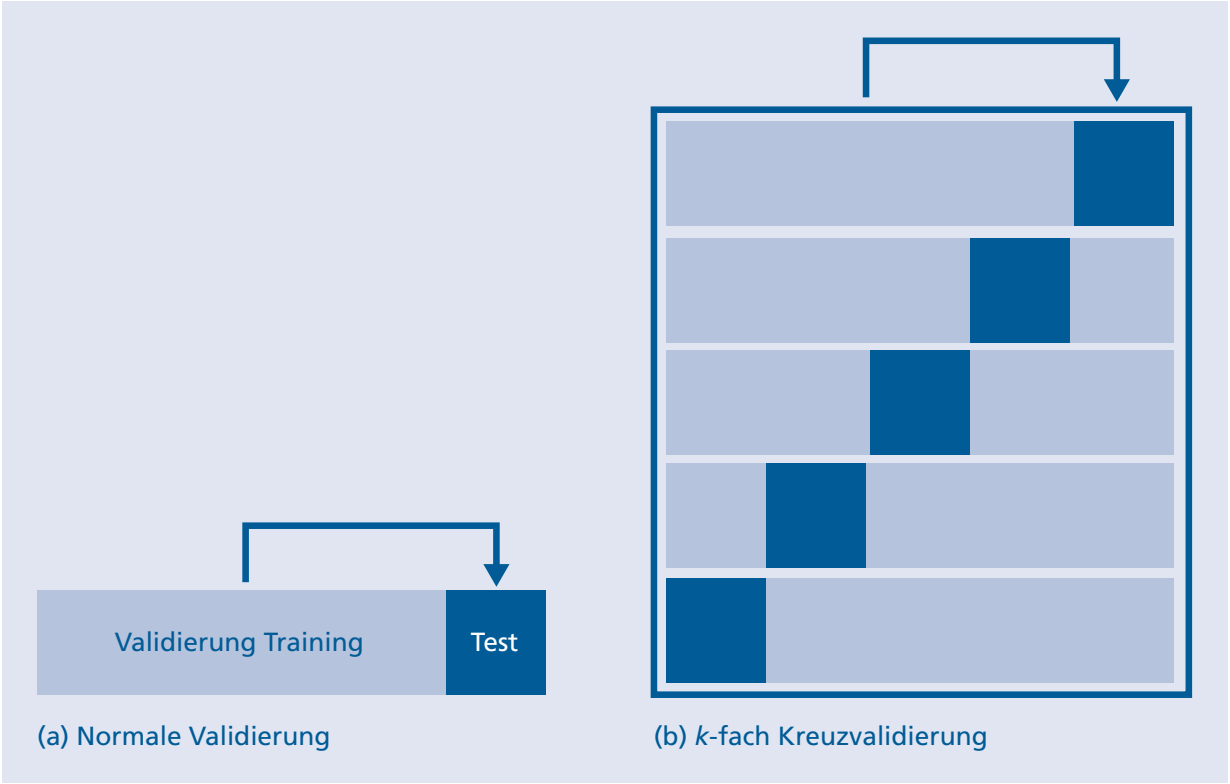


Abbildung 4.4
Vergleich zwischen normaler Validierung und k-fach Kreuzvalidierung: Bei der Kreuzvalidierung wird in k Durchgängen eine jeweils andere Teilmenge des Materials zum Test und der Rest der annotierten Daten als Trainingsmenge verwendet.

4.3.1 Datennormalisierung

Beim Crawlen von Webseiten, also dem automatischen Herunterladen von Webseiten und Verfolgen der darin enthaltenen Hyperlinks, stößt man unweigerlich recht bald auf fehlende oder fehlerhafte Angaben in Hypertext Transfer Protocol (HTTP)-Headern oder den Webseiten selbst. Die Aufgabe der Datennormalisierung besteht darin, die heterogenen Inhalte der vielen unterschiedlichen Quellen zu vereinheitlichen.

4.3.1.1 Automatische Fehlerkorrektur

Bei der Analyse von Webseiten muss beachtet werden, dass häufig Fehler in diesen Seiten zu finden sind, zum Beispiel ungültige HyperText Markup Language (HTML)-Strukturen.

Da die meisten Seiten trotz fehlerhafter Strukturen im sogenannten »Quirks-Mode« vom Browser gerendert (dargestellt) werden können, bleiben die Fehler vom Benutzer meist unbemerkt. Um eine Webseite mit fehlerhafter HTML-Struktur auch ohne Browser in einer Filtersoftware verarbeiten zu können, lassen sich entsprechende Softwarebibliotheken[3] zur HTML-Bereinigung voranstellen.

Typisch ist außerdem, dass der verwendete Zeichensatz oder der unter einer Webadresse erreichbare Datentyp, der »MIME-Type«, falsch oder gar nicht angegeben wird. Hier helfen Softwarebibliotheken zur automatischen Zeichensatz- und Datentyp-Erkennung.[4] Die Erkennungsgenauigkeit dieser

[3] zum Beispiel HTML Tidy Library Project, siehe http://tidy.sourceforge.net, Stand 30.11.2012
[4] zum Beispiel Apache Tika – a content analysis toolkit, siehe http://tika.apache.org, Stand 30.11.2012

Bibliotheken ist gemeinhin sehr hoch, so dass sich durch deren Einsatz keine Folgefehler erwarten lassen.

4.3.1.2 Textkodierungen vereinheitlichen

Nach diesen grundlegenden Schritten sind einige weitere und häufig vernachlässigte Datennormalisierungsschritte dringend zu empfehlen. Um die nachfolgende, in Abschnitt 4.4 näher beschriebene Verarbeitung zu vereinfachen, sollten alle nicht sichtbaren Zeichen entfernt werden, damit rein optisch identische Wörter auch bei einem syntaktischen Vergleich als identisch erkannt werden. Ein Beispiel eines solchen Zeichens ist das weiche Trennzeichen[5], das in Browsern nur am Zeilenende bei der automatischen Silbentrennung sichtbar ist, sich aber als unsichtbares Zeichen in jedem Wort einer Webseite befinden kann.

Dasselbe Problem ergibt sich auch bei dem Vergleich von Texten ohne vorherige Unicode-Normalisierung. Werden Texte in einem Unicode-Zeichensatz angeliefert (zum Beispiel UTF-8 oder UTF-16), so können sie zusammengesetzte oder kombinierte Zeichen enthalten. Das bedeutet, in Unicode lässt sich ein »ä« als sogenanntes Kompatibilitätszeichen als ein einzelner kombinierter Buchstabe oder äquivalent als »a« mit einem kombinierenden horizontalen Doppelpunkt »¨« darstellen. Werden zwei auf diese Weise unterschiedlich kodierte Wörter »töten« und »to¨ten« von der Filtersoftware verglichen, werden sie als ungleich angesehen, obwohl sie im Browser identisch angezeigt werden.

Diese Vorverarbeitung ist extrem wichtig, da alle daran anschließenden Schritte nur dann zuverlässig funktionieren, wenn die Daten einheitlich und korrekt angeliefert werden. Allein ein falscher Zeichensatz könnte dazu führen, dass eine Webseite nicht als jugendschutzrelevant erkannt wird. Angenommen, eine Webseite enthält jugendschutzrelevanten Inhalt und der Zeichensatz ist angegeben als ISO 8859-5 (kyrillisch). Der eigentliche Zeichensatz dieser Webseite ist allerdings UTF-8. Wird dieser Fehler nicht automatisch erkannt und behoben, führt es dazu, dass die aus der Webseite extrahierten Wörter für das zur Filterung verwendete Modell fast ausschließlich unbekannt sind und sich damit jeder Erkennung entziehen.

Problematisch ist der geschilderte Fall zusätzlich dadurch, dass sämtliche modernen Browser eine recht gute automatische Zeichensatz-Korrektur vornehmen, so dass der Benutzer von diesen Fehlern nichts bemerkt und nur die korrigierten Inhalte zu sehen bekommt. Es ist also unverzichtbar, dass die Filtersoftware die Web-Inhalte in ähnlicher Weise aufbereitet bekommt, in der auch Browser die Inhalte aufbereiten, bevor sie dem Benutzer angezeigt werden.

Eine weitere Schwierigkeit besteht in diesem Zusammenhang durch den Einsatz von Java-Script in Webseiten. Wird in dynamisch aufgebauten Webseiten der Inhalt mittels Java-Script vollständig geändert, so passiert dies bei der Betrachtung im Browser häufig völlig unbemerkt vom Benutzer. Viele einfache Webcrawler interpretieren diesen Code allerdings nicht, so dass die Inhalte unverändert bleiben. In diesem Fall wird die Einstufung nach Jugendschutzrelevanz nicht auf Basis dessen vorgenommen, was der menschliche Benutzer am Browser tatsächlich zu sehen bekommt.

4.3.2 Extraktion von HTML-Merkmalen

Verglichen mit rein textuellen Daten lassen sich aus der HTML-Struktur einer Webseite vielfältige zusätzliche Informationen gewinnen. So geben viele Webseiten beispielsweise eine inhaltliche Zusammenfassung und eine Liste von Stichwörtern in Form von META-Tags im Kopfteil der Seite an.

[5] siehe http://de.wikipedia.org/wiki/Weiches_Trennzeichen, Stand 30.11.2012

Das Ergebnis von [Golub und Ardö, 2005] legt allerdings nahe, dass die Einbeziehung von expliziten Metadaten allein keinen besonderen Einfluss auf die Klassifikationsqualität hat. Es lassen sich allerdings besonders für die im Rahmen dieser Studie betrachteten Inhalte weitere interessante Merkmale aus der HTML-Struktur gewinnen.

In Kombination mit einer inhaltlichen Klassifikation nach Themen (zum Beispiel »Gewalt« oder »selbstverletzendes Verhalten«) erscheint es hilfreich, eine funktionale Klassifikation ganzer Bereiche eines Webangebots durchzuführen. Damit gemeint ist die Erkennung von Linklisten, Forenstrukturen, soziale Netzwerkfunktionen oder Listen von Video-Thumbnails mit Links zu Videoseiten. Die so gewonnenen zusätzlichen Merkmale können die Erkennung des Gefährdungspotenzials einer Webseite durchaus sinnvoll unterstützen.

Einen guten Überblick über die allgemeine Klassifikation von Webseiten und die verschiedenen Möglichkeiten, Merkmale aus HTML zu gewinnen, bietet [Qi und Davison, 2009]. Wie in [Kan und Thi, 2005] gezeigt wird, lassen sich Webseiten bis zu einem gewissen Grad sogar bloß anhand ihrer Webadresse kategorisieren, ohne deren Inhalte selbst zu betrachten. Auch dies erscheint in Anbetracht der zu filternden Datenmengen durchaus sinnvoll, da sich so die Notwendigkeit des Herunterladens jeder einzelnen Webseite erübrigt. Experimente dazu finden sich in den Tabellen 5.9, 5.11 und 5.19, jeweils in Kapitel 5.

Die wichtigsten Merkmale und Kriterien zur Entscheidung, ob eine Webseite jugendschutzrelevante Inhalte enthält oder nicht, finden sich aber für gewöhnlich im Inhalt der Webseite und nicht in deren Metainformationen.

4.4 Technische Verfahren für die Textanalyse

In den letzten Jahren wurden eine Reihe automatischer computergestützter Verfahren zur Erkennung von spezifischen Inhalten in Texten und Webseiten entwickelt. Beispiele sind die Klassifikation von E-Mails als »Spam« oder »Kein Spam« oder die Kategorisierung von Zeitungsartikeln in mehrere Kategorien, zum Beispiel »Politik«, »Sport«, »Wirtschaft« usw. Die Klassifikation von Internetinhalten als »Jugendschutzrelevant« und »Unbedenklich« stellt ähnliche Anforderungen und kann daher viele der schon entwickelten Verfahren übernehmen.

4.4.1 Vorverarbeitung von Texten

Steht der Text eines Dokumentes zur Verfügung, gibt es eine Reihe von Arbeitsschritten zur Vorverarbeitung, die in der Regel nacheinander angewendet werden.

- Die Layoutanalyse interpretiert die Gestaltung des Dokuments, um daraus strukturelle Eigenschaften wie Titel, Navigation, Überschriften, Abschnitte usw. zu erkennen. Für die weitere Verarbeitung werden HTML-Anweisungen entfernt.
- Die Sprachidentifikation erkennt die Sprache eines Dokumentes (Deutsch, Englisch etc.). Dies geschieht mit hoher Zuverlässigkeit anhand charakteristischer Buchstabenfolgen. [Gottron und Lipka, 2010] erzielen beispielsweise eine Genauigkeit von 99,4 % bei der Erkennung der Sprache von Überschriften.
- Die Tokenisierung erkennt Token (Wörter, Zahlen, Satzzeichen) und Sätze im Text. Hierbei wird der Text an den Zwischenräumen in Token zerlegt. Problematisch ist jedoch die Erkennung und Behandlung von Abkürzungen.

- Die Wortarterkennung (Part-of-speech Tagging) ordnet die Token eines Textes einer Wortart (engl.: part-of-speech) zu. Hierbei werden sowohl die Buchstaben des Wortes als auch der Kontext (zum Beispiel angrenzende Wörter oder Wortarten) berücksichtigt.
- Die Grundformreduktion (Lemmatization) ordnet die Wörter des Textes der Grundform (engl.: lemma) zu (zum Beispiel lief $\rightarrow$ laufen). Als Alternative werden oft Stemming-Verfahren angewendet, welche jedes Wort durch Abschneiden der Endungen auf den Wortstamm reduzieren [Porter, 1980].
- Das Parsing analysiert den Text durch sogenannte Parser und bestimmt die Funktion einzelner Wörter und Phrasen. Ziel ist es, in einem Satz Subjekt, Prädikat, Objekt, Nebensatz, etc. zu erkennen. In vielen Fällen wird die Satzstruktur durch einen Syntaxbaum beschrieben. Leider ist die Zuverlässigkeit dieser Verfahren oft nicht sehr hoch.

Allerdings werden nicht immer sämtliche Arbeitsschritte benötigt. Es schließt sich die semantische Analyse an, die zum Beispiel durch eine Klassifikation oder eine Erkennung von charakteristischen Phrasen erfolgen kann. Die hierzu verfügbaren Techniken werden in den folgenden Abschnitten diskutiert.

Während der Vorverarbeitung können eine Reihe weiterer Merkmale bestimmt werden, welche für die Klassifikation verwendet werden können.

- n-Gramme sind Folgen von jeweils n Wörtern. Da die Anzahl der unterschiedlichen Wortkombinationen sehr schnell wächst, betrachtet man in der Regel nur 2-Gramme (Bigramme) und 3-Gramme (Trigramme) von Wörtern. Diese charakterisieren typische Folgen von Wörtern, die zusammen auftauchen. Beispielsweise erscheint das Bigramm »Schwein gehabt« meist in ganz anderen Kontexten auf als die einzelnen Wörter »Schwein« und »gehabt«. Alternativ können auch n-Gramme von Buchstaben als Merkmal verwendet werden.
- Das Merkmalsgewicht ist ein nicht-negativer numerischer Faktor, der den Informationsgehalt eines Wortes oder eines anderen Merkmals misst. Während eine Reihe von gängigen Wörtern in fast allen Dokumenten auftaucht, erscheinen andere Wörter nur in wenigen inhaltlichen Zusammenhängen. Das tf-idf-Maß (von engl. term frequency – inverse document frequency) eines Wortes w quantifiziert diesen Informationsgehalt. Es ist definiert als

  $$tfidf(w,d) = tf(w,d) * \log \frac{N}{df(w)} .$$

  Die Termfrequenz $tf(w,d)$ gibt an, wie häufig Wort w in Dokument d erscheint. N ist die Gesamtzahl der Dokumente im Korpus und $df(w)$ die Dokumentfrequenz, also die Anzahl der Dokumente, in denen w vorkommt.

Ergebnis der Vorverarbeitung ist in der Regel ein großer Vektor x von M Merkmalen des Dokuments d. Typischerweise gehört dazu für jedes Wort in der Dokumentensammlung die Häufigkeit seines Auftauchens in Dokument d. Der Eingabevektor x hat daher meist mehrere zehntausend bis zu mehreren Millionen Komponenten. Nur die Merkmale, die in dem Dokument auftauchen, erhalten einen Wert ungleich Null; alle anderen Komponenten haben den Wert Null. Eine solche Darstellung eines Dokuments wird Vektorraum-Darstellung (engl.: vector space model) genannt. Da in der Regel nur relativ wenige Merkmale in einem Dokument zutreffen, so kann man eine spärliche Vektordarstellung (engl.: sparse vector) verwenden.

4.4.2 Wichtige Verfahren zur Klassifikation von Textdokumenten

Für die Klassifikation von Textdokumenten steht eine ganze Reihe von Verfahren zur Verfügung, die wir im Folgenden vorstellen.

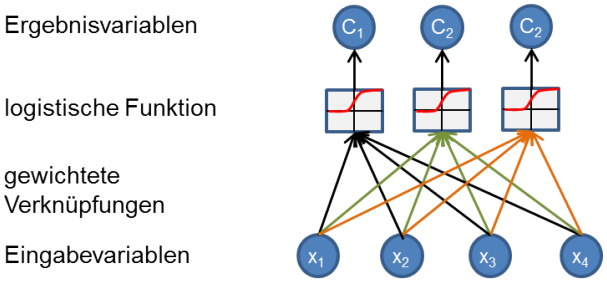


Abbildung 4.5
Der logistische Klassifikator bildet gewichtete Summen aus den Eingabevariablen x_i, welche durch die logistische Funktion in Wahrscheinlichkeiten für die Ergebnisvariablen c_j transformiert werden.

4.4.2.1 Bayes-Klassifikator

Ein einfacher Klassifikator ist der Bayes-Klassifikator. Er berechnet für ein Dokument d für jede der möglichen Klassen $c_1, \ldots, c_k$ die bedingte Wahrscheinlichkeit $p(c_i|d)$, dass d zu Klasse c_i gehört. Natürlich müssen sich die Wahrscheinlichkeiten zu eins addieren, das heißt $p(c_1|d)+\cdots+p(c_k|d) = 1$. Kennt man diese Wahrscheinlichkeiten, so kann d der Klasse mit der höchsten Wahrscheinlichkeit zugeordnet werden.

Der Bayes-Klassifikator trifft nun die stark vereinfachende Annahme, dass in den Dokumenten, die zu einer Klasse gehören, alle Wörter unabhängig auftreten. Diese Annahme ist in der Regeln nicht erfüllt, da zum Beispiel das Wort »Merkel« wesentlich wahrscheinlicher für ein Dokument wird, wenn das Wort »Angela« im Dokument auftaucht. Die Berechnung des Bayes-Klassifikators ist für ein Computerprogramm einfach: Es müssen lediglich die relativen Häufigkeiten jedes Merkmals x_i (zum Beispiel Wörter) in den Trainingsdokumenten der einzelnen Klassen (zum Beispiel »Jugendschutzrelevant« und »Unbedenklich«) bestimmt werden. Trotz der stark vereinfachenden Annahmen ist der Bayes-Klassifikator in vielen realen Anwendungen recht leistungsfähig. Ein Vorteil des Bayes-Klassifikators ist, dass er mit relativ wenigen Trainingsdaten auskommt. Darüber hinaus ist es möglich, einen Bayes-Klassifikator inkrementell an neue Daten anzupassen, da ja lediglich die relativen Häufigkeiten zu adaptieren sind. Aus diesem Grund wird er etwa im Spamfilter des Thunderbird E-Mail-Programms verwendet.

4.4.2.2 Logistischer Klassifikator

Der logistische Klassifikator [Hastie et al., 2011, S. 119] ist ein einfach zu implementierender aber leistungsfähiger Klassifikator, welcher den Eingabevektor x einer der möglichen Klassen $c_1, \ldots, c_k$ zuordnet. Um die Wahrscheinlichkeit für eine Klasse c_i zu berechnen, werden die Eingabemerkmale x eines Dokuments d mit spezifischen Gewichten $b_{i,k}$ der Klasse c_i multipliziert und anschließend addiert. Es ergibt sich eine Maßzahl

$$s_i = b_{i,0} + b_{i,1}x_1 + b_{i,2}x_2 + \cdots + b_{i,M}x_M$$

die einen beliebigen reellen Zahlenwert annehmen kann. Eine solche Transformation nennt man auch ein lineares Modell. Durch die sogenannte logistische Funktion kann man die Zahlenwerte $s_1, \ldots, s_k$ in Wahrscheinlichkeiten $p(c_1|d), \ldots, p(c_k|d)$ für die einzelnen Klassen überführen, so dass höhere Werte von m_i auch höheren Werten $p(c_i|d)$ entsprechen. Abbildung 4.5 zeigt die Struktur eines logistischen Klassifikators an einem kleinen Beispiel.

Steht eine Trainingsmenge $D = \{d_1, \ldots, d_n\}$ von Dokumenten mit den annotierten Klassen zur Verfügung, so kann man für den gegebenen Parametervektor b die Wahrscheinlichkeit der gesamten Trainingsmenge berechnen. Diese Wahrscheinlichkeit wird auch Likelihood genannt.

Mit Hilfe eines Optimierungsverfahrens lassen sich nun die Parameter b so anpassen, dass alle Dokumente der Trainingsmenge eine hohe Wahrscheinlichkeit erhalten. Hierzu wird der Gradient von b benötigt, also die Richtung des steilsten Anstiegs der Likelihood. Die Optimierung wird durch schrittweises Anpassen durchgeführt, also zum Beispiel mit der Broyden-Fletcher-Goldfarb-Shanno (BFGS) Methode [Nocedal und Wright, 2006].

Gibt es k Klassen und M Merkmale, so hat der logistische Klassifikator oft mehrere Millionen Parameter. In diesem Fall ist eine Regularisierung der Parameter erforderlich, um eine Überanpassung zu vermeiden. Dies sind meist zusätzliche Strafterme, die die Parameterwerte in Richtung 0 ziehen und den Einfluss überflüssiger Merkmale verringern. Die Größe dieser Strafterme kann man durch Kreuzvalidierung berechnen.

Es zeigt sich, dass der logistische Klassifikator im Allgemeinen leistungsfähiger ist als der verwandte Bayes-Klassifikator. Ein Grund hierfür ist, dass der Bayes-Klassifikator alle Wahrscheinlichkeiten direkt aus der Trainingsmenge berechnet. Meist ist aber die Annahme der Unabhängigkeit der Merkmale x_j verletzt und einzelne Merkmale x_j sind stark korreliert, sagen also das Gleiche aus. Daher müsste das Gewicht dieser Merkmale insgesamt verringert werden. Dies erfolgt automatisch beim logistischen Klassifikator, indem die Merkmale so angepasst werden, dass die Prognose direkt optimiert wird.

Exkurs: Maximum Likelihood Ansatz Der logistische Klassifikator ist ein Beispiel für den Maximum-Likelihood Ansatz [Hastie et al., 2011, S. 265]. Dabei wird derjenige Parameter als Schätzung b^* ausgewählt, für den die Wahrscheinlichkeit der Trainingsdaten (Likelihood) maximal ist. Unter wenig einschränkenden Annahmen lässt sich beweisen, dass die gewonnenen Schätzwerte die Informationen in den Trainingsdaten in optimaler Weise auswerten. Damit haben die geschätzten Parameter die minimale Abweichung von den »wahren« Parametern, die bei den vorliegenden Daten möglich ist. Darüber hinaus haben die Abweichungen die Form einer Normalverteilung. Dies ist die Grundlage für sehr viele statistische Tests.

Schließlich ergibt sich ein allgemeiner Zusammenhang zwischen Höhe der Fehler und Größe der Trainingsmenge. Wird die Anzahl n der Trainingsdokumente größer, so sinkt die mittlere Differenz des geschätzten Parameters vom wahren Parameterwert mit einem Faktor $1/\sqrt{n}$. Dies bedeutet, dass eine viermal größere Trainingsmenge die Differenz zwischen dem wahren Parameter und dem geschätzten Parameter im Mittel näherungsweise halbiert. Diese Beziehung gilt auch für die Klassifikationsgenauigkeit.

Merkmalsauswahl Bei der Klassifikation von Textdokumenten gibt es oft hunderttausende von möglichen Eingabemerkmalen x_i, zum Beispiel die Häufigkeit, mit der ein Wort in dem jeweiligen Dokument vorhanden ist. Für jedes Merkmal muss in der Regel ein Parameter bestimmt werden. Man kann nun für viele Klassifikatoren zeigen, dass die Streuung der Prognose wächst, falls man Merkmale verwendet, welche keine Beziehung zu der Dokumentenklasse haben. Daher ist es bei vielen Klassifikatoren hilfreich, die Anzahl der Merkmale auf eine optimale Anzahl zu reduzieren.

Wie oben diskutiert, werden die Abweichungen zwischen dem geschätzten Parameter b^* und dem wahren Parameter durch eine Normalverteilung beschrieben. Dies erlaubt die Ableitung allgemeiner Tests zur Prüfung von Klassifikatoren und Koeffizienten. Anhand des Likelihood-Quotienten-Tests kann man erkennen, ob ein Klassifikator mit weniger Parametern signifikant schlechter ist als der Klassifikator mit mehr Parametern. Mit dem Wald-Test kann man prüfen, ob bestimmte Parameter erforderlich sind oder weggelassen werden können. Diese Tests können als ein Kriterium verwendet

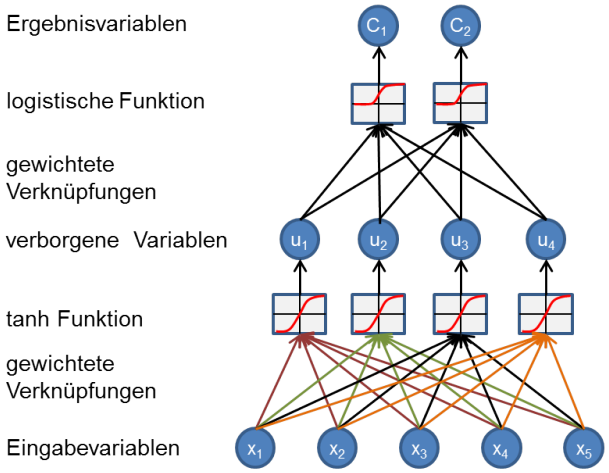


Abbildung 4.6
Das Neuronale Netz besteht aus mehreren Schichten. Bei den hier gezeigten drei Schichten werden zunächst gewichtete Summen aus den Eingabevariablen x_i gebildet, die mit der $\tanh$-Funktion nichtlinear in die verborgenen Variablen u_j transformiert werden, welche Werte in $[-1, 1]$ annehmen können. Es folgt ein logistischer Klassifikator. Er bildet gewichtete Summen aus den verborgenen Variablen u_i, welche durch die logistische Funktion in Wahrscheinlichkeiten für die Ergebnisvariablen c_j transformiert werden.

werden, um zu entscheiden, ob bestimmte Eingabemerkmale verwendet werden sollen oder nicht. Viele Ansätze suchen schrittweise dasjenige Merkmal, welches den Klassifikator in optimaler Weise verbessert und nehmen es in die Menge der verwendeten Merkmale auf [Hastie et al., 2011, S. 57f].

Um den Rechenaufwand zu reduzieren, werden auch einfachere Strategien zur Merkmalsauswahl verfolgt. Eine Möglichkeit ist, die Korrelation zwischen der Klassenvariablen und allen verbliebenen Merkmalen zu berechnen und ihre statistische Signifikanz durch Tests (zum Beispiel Chi-Quadrat-Test) zu bestimmen. Oft wird dann eine größere Anzahl von Merkmalen mit hoher Korrelation der Menge der verwendeten Merkmale hinzugefügt. Dieses Vorgehen ist suboptimal, weil es die Korrelation der Merkmale untereinander ignoriert, führt in der Praxis dennoch oft zu guten Ergebnissen.

4.4.2.3 Künstliches Neuronales Netz

Der logistische Klassifikator hat zwei Schichten, die Eingabevariablen x und die Ergebnisvariablen c. Man kann diesen Klassifikator um weitere Schichten von intermediären Variablen u erweitern, indem man mehrere Lagen von Klassifikatoren hintereinander schaltet. Wichtig ist, dass in jeder Lage eine nichtlineare Transformation durchgeführt wird, etwa mit der logistischen Funktion oder der sigmoiden Funktion $\tanh(x)$. Abbildung 4.6 zeigt die Struktur eines Neuronalen Netzes an einem kleinen Beispiel. Historisch wurden derartige Modelle wegen ihrer Ähnlichkeit zu biologischen Nervennetzwerken als künstliche Neuronale Netze bezeichnet [Hastie et al., 2011, S. 389]. Da für die intermediären Variablen u in der Regel keine Beobachtungswerte verfügbar sind, bezeichnet man sie auch als verborgene Variablen.

Wie beim logistischen Modell kann eine Likelihood-Funktion formuliert werden, welche die Wahrscheinlichkeit der beobachteten Dokumente für die jeweiligen Modellparameter misst. Der Gradient dieser Likelihood-Funktion kann effizient berechnet werden. Hiermit lassen sich die optimalen Parameters des Modells mit Hilfe eines Optimierungsverfahrens berechnen.

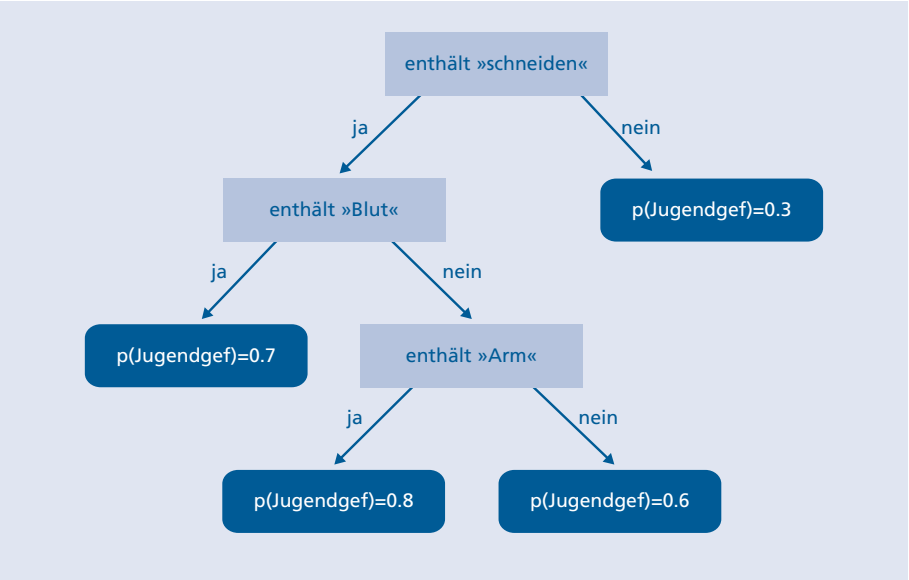


Abbildung 4.7
Einfacher Entscheidungsbaum zur Klassifikation von »jugendschutzrelevant«. In die Entscheidung gehen die Wörter des Dokumentes ein.

Der Vorteil eines künstlichen Neuronalen Netzes liegt darin, dass die verborgenen Variablen eine Transformation der Eingaben durchführen, welche die endgültige Klassifikation verbessern kann. In den letzten Jahren wurden in diesem Bereich neue Ansätze entwickelt, welche die Performanz der Neuronalen Netze stark erhöht haben [Bordes et al., 2011].

4.4.2.4 Entscheidungsbaum und Random Forest

Entscheidungsbäume [Hastie et al., 2011, S. 305] sind eine Methode zur automatischen Klassifikation von Dokumenten. Ein Entscheidungsbaum besteht aus einem Wurzelknoten und beliebig vielen inneren Knoten sowie mindestens zwei Blättern. Dabei repräsentiert jeder Knoten eine logische Regel und jedes Blatt gibt die gesuchte Wahrscheinlichkeit der Klassen an. Ein einfaches Beispiel eines Entscheidungsbaumes gibt Abbildung 4.7.

Ein Baum kann auf einfache Weise gelernt werden. Sei $D_1 = \{d_1, \ldots, d_n\}$ die Trainingsmenge. Für den ersten Schritt $t = 1$ wird nun das Merkmal x_i gesucht, welches als einzelnes Merkmal in D_t die Klassen $c_j \in C$ in optimaler Weise prognostiziert. Als Merkmal wird meist der Informationsgewinn verwendet [Mitchell, 1997]. Anschließend wird D_0 in zwei Teilmengen D_{t_1} und D_{t_0} zerlegt, je nachdem ob Merkmal x_i in dem Dokument vorkommt oder nicht. Anschließend werden die beiden Teilmengen D_{t_1} und D_{t_0} jede für sich in gleicher Weise analysiert, das heißt optimale prognostizierende Merkmale für die jeweilige Teilmengen gesucht. Durch sukzessive Anwendung entsteht so eine Baumstruktur, der Entscheidungsbaum (vgl. Abbildung 4.7). Das Verfahren stoppt, sobald alle Dokumente in der entstehenden Teilmenge die gleiche Klasse haben, oder kein Merkmal die Klassifikationsgenauigkeit verbessern kann. Häufig wird der Baum verkleinert, indem man mit Hilfe der Kreuzvalidierung überflüssige Aufteilungen (Splits) eliminiert (pruning).

Da im Mittel die Trainingsmenge durch jeden Split halbiert wird, ist die Trainingsmenge nach relativ wenigen Merkmalen erschöpft und es können nur wenige Merkmale für einen Entscheidungspfad berücksichtigt werden. Als Alternative wurde das Klassifikationsverfahren des Random Forest entwickelt [Breiman, 2001, Hastie et al., 2011], welches aus mehreren verschiedenen, unkorrelierten

Entscheidungsbäumen besteht. Für eine Klassifikation darf jeder Baum in diesem Wald eine Entscheidung treffen und als endgültige Klassifikation wird die Klasse mit den meisten Stimmen gewählt.

Die Entscheidungsbäume des Random Forests werden im Allgemeinen mit folgenden Verfahren gewonnen:

- Bei n Dokumenten in der Trainingsmenge werden $m < n$ Dokumente zufällig ausgewählt. Dabei können Dokumente auch mehrfach ausgewählt werden.
- Bei M Merkmalen der Dokumente werden an jedem Knoten im Baum $q \ll M$ Merkmale zufällig gewählt, die zur Betrachtung des Splits genommen werden.
- Der Baum wird voll ausgebaut und nicht verkleinert.

Es hat sich gezeigt, dass ein Random Forest meist eine sehr hohe Klassifikationsgenauigkeit hat. Er kann in effizienter Weise tausende von Merkmalen berücksichtigen.

4.4.2.5 Support Vektor Machine

Eine Support Vektor Maschine (SVM) ist ein Klassifikator [Hastie et al., 2011, S. 417] für zwei Klassen. Sie unterteilt eine Menge von Dokumenten so in Klassen, dass um die Klassengrenzen herum ein möglichst breiter Bereich frei von Dokumenten bleibt (Large Margin). Als Klassengrenze wird eine Hyperebene gewählt. Der Abstand derjenigen Vektoren, die der Hyperebene am nächsten liegen, wird dabei maximiert. Dieser breite, leere Rand (Margin) soll später dafür sorgen, dass auch Objekte, die nicht genau den Trainingsobjekten entsprechen, möglichst zuverlässig klassifiziert werden.

Die Klassifikationsfunktion der SVM hat die Form

$$y_i = \text{sgn}\left(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_M x_M\right)$$

Die Signum-Funktion hat die Werte $\text{sgn}(u) = 1$ falls $u > 0$, $\text{sgn}(u) = -1$ falls $u < 0$ und $\text{sgn}(u) = 0$ falls $u = 0$. Die Werte -1 und 1 bedeuten die Zuordnung zur Klasse 1 beziehungsweise 2. Mit Hilfe eines Optimierungsverfahrens werden die Parameter $b_0, b_1, \cdots, b_M$ so modifiziert, dass für die Dokumente der Trainingsmenge der kleinste Abstands zur Hyperebene möglichst groß wird. Nach der statistischen Lerntheorie ist die Komplexität der Klasse aller Hyperebenen mit einem größeren Margin geringer als die der Klasse aller Hyperebenen mit einem kleineren Margin. Die Generalisierungsfähigkeit dieser Klassifikatoren ist also höher.

Diejenigen Dokumente mit dem minimalen Abstand (Margin) von der Hyperebene werden Supportvektoren genannt. Sie bestimmen die Lage der Hyperebene. In der Regel sind nur relativ wenige Dokumente Supportvektoren. Abbildung 4.8 zeigt die Supportvektoren in einem Klassifikationsproblem.

Oft sind die Trainingsbeispiele nicht streng linear separierbar. Für diesen Fall wird das Optimierungsproblem durch die Einführung von Schlupfvariablen so verändert, dass Verletzungen der Nebenbedingungen möglich sind, das Ausmaß der Verletzungen aber so klein wie möglich bleibt. Beide Optimierungsprobleme sind konvex und können mit modernen Verfahren effizient gelöst werden.

Es stellt sich heraus, dass zur Berechnung der optimalen Parameter nur die Skalarprodukte $x^T z = x_1 * z_2 + \cdots + x_M * z_M$ erforderlich sind. Hierbei sind $x = (x_1, \ldots, x_M)$ und $z = (z_1, \ldots, z_M)$ die zu den Dokumenten d_i und d_j gehörenden Vektoren. Daher kann man auch die Eingabevektoren zunächst mit einer nichtlinearen Funktion $\phi(x)$ in einen höherdimensionalen Eingaberaum transformieren, um dann das Skalarprodukt $\phi(x_i)^T \phi(x_j)$ zu berechnen. Die Anzahl der neu entstehenden Merkmale

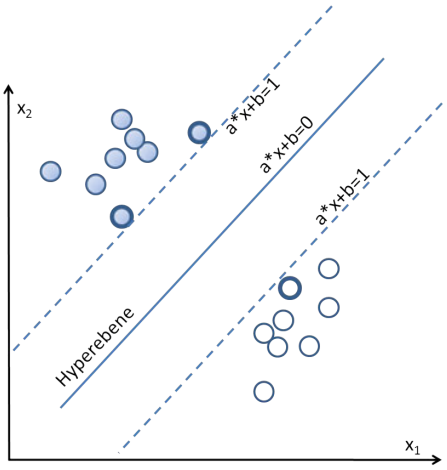


Abbildung 4.8
Support Vektor Maschine mit einer trennenden Hyperebene zwischen den Dokumenten der beiden Klassen. Die Supportvektoren liegen am nächsten an der Hyperebene und sind durch fette Kreise symbolisiert.

ist dabei wesentlich größer als die Anzahl M der ursprünglichen Eingabemerkmale x. In diesem größeren transformierten Merkmalsraum kann nun eine Hyperebene mit dem ursprünglichen SVM-Verfahren konstruiert werden. Durch die erhöhte Anzahl von neuen Merkmalen werden Klassen trennbar, welche sich im ursprünglichen Merkmalsraum nicht durch eine Hyperebene trennen ließen.

Die Abbildung $\phi(x)$ wird allerdings nicht explizit benötigt, da ja nur die Werte des Skalarproduktes $K(x,z) = \phi_1(x) * \phi_1(z) + \cdots + \phi_N(x) * \phi_N(z)$ erforderlich sind. $K(x,z)$ wird auch Kernfunktion genannt [Joachims, 2002]. Folgende Kernfunktionen werden oft verwendet:

- Linearer Kern: $K(x,z) = x^T z$.
- Polynomialer Kern: $K(x,z) = (\gamma x^T z + r)^m$ für $m = 1$, $m = 2$ oder $m = 3$.
- Radiale Basisfunktion: $K(x,z) = \exp(-\gamma \|x - z\|^2)$, für ein $\gamma > 0$.
- Sigmoider Kern: $K(x,z) = \tanh(\gamma x^T z + r)$ für einen Parameter r.

Dabei sind γ, r und m spezielle Kernparameter, welche oft durch Kreuzvalidierung bestimmt werden. Eine Kernfunktion muss lediglich gewissen Symmetriebedingungen genügen. Kann man diese sicherstellen, so können Kernfunktionen sogar auf komplexere Strukturen wie Bäume [Reichartz et al., 2010] oder Graphen [Gärtner et al., 2003] angewendet werden.

Die wichtigste Eigenschaft der SVM ist, dass sie gemäß theoretischer Abschätzungen [Vapnik, 1995] weitgehend unabhängig von der Anzahl der Eingabemerkmale ist. Sie kann mit Millionen von Eingabemerkmalen arbeiten, da sie eine Anzahl von Datenpunkten (Support Vektoren) auswählt, welche die Hyperebene definieren, und ignoriert die restlichen Datenpunkte. Dies gestattet eine gute Generalisierung und macht die SVM besonders performant bei der Klassifikation von Textdokumenten.

Bevor die Support Vektor Maschine trainiert werden kann, müssen die Dokumente der Trainingsmenge eine Vorverarbeitung durchlaufen, also Tokenisierung und Satzzerlegung. Zusätzliche Merkmale werden durch Wortart-Erkennung und gegebenenfalls n-Grammbildung gewonnen. Es ist wichtig, dass diese Merkmale gewichtet werden, zum Beispiel durch tf-idf (vgl. Abschnitt 4.4.1). Zudem müssen die Eingabemerkmale einheitlich auf die Länge 1 normiert werden. Auch ist zu beachten, dass alle neuen Dokumente die gleiche Vorverarbeitung durchlaufen.

4.4.2.6 k-Nächste-Nachbarn Klassifikator

Das k-Nächste-Nachbarn Verfahren (kNN) [Hastie et al., 2011, S. 463] klassifiziert Dokumente entsprechend ihrer Ähnlichkeit zu Dokumenten in der Trainingsmenge. Dabei wird die Ähnlichkeit durch eine Distanzfunktion $Dist(x, z)$ zwischen den Merkmalsvektoren von zwei Dokumenten d_1 und d_2 definiert. Ein Dokument wird in folgenden Schritten klassifiziert:

- Berechne die Entfernung $\delta_i = Dist(x(d), x(d_i))$ zwischen d und den Vektoren aller Dokumente d_i in der Trainingsmenge.
- Sortiere die Distanzen δ_i und bilde die Menge D_k der k Dokumente mit dem geringsten Abstand.
- Bestimme die Klasse c, welche bei den annotierten Dokumenten in D_k am häufigsten vorkommt und verwende sie als Klassifikation für d.

Es gibt eine Reihe von Varianten des Ansatzes (vgl. [Sebastiani, 2002, S. 28]). Um zu vermeiden, dass die Klasse mit der größten Anzahl von Dokumenten zu oft gewählt wird, wird bei der Auswahl der Klasse auch noch die Distanz zu den einzelnen Dokumenten berücksichtigt. Bei der Wahl des k-Wertes muss man beachten, dass ein zu geringes k die Varianz vergrößert. Der konkrete Wert von k und die beste Art der Distanzfunktion wird oft durch Kreuzvalidierung gewählt.

In einer Reihe von Vergleichsexperimenten hat sich gezeigt, dass das kNN Verfahren sehr effektiv ist. Sein größter Nachteil ist der hohe Aufwand zur Berechnung der Distanzen während der Anwendung des Klassifikators.

4.4.2.7 Kombination mehrerer Klassifikatoren

Durch die Kombination unterschiedlicher Klassifikatoren ist es möglich, die Klassifikationsgenauigkeit zu vergrößern. Wir stellen hier das sogenannte Bagging und das Boosting vor.

Bagging Bagging [Breiman, 1996] ist auf beliebige Klassifikatoren anwendbar und erzeugt eine Reihe unterschiedlicher Trainingsmengen. Voraussetzung ist eine Trainingsmenge $D = \{d_1, \ldots, d_n\}$ von Dokumenten, die schon mit den betrachteten Kategorien $c \in C$ (zum Beispiel »Jugendschutzrelevant«, »Unbedenklich«) manuell annotiert wurden. Hieraus werden B neue Trainingsmengen D_r, $r = 1, \ldots, R$, erzeugt. Dies geschieht, indem zufällig $m \leq n$ Dokumente aus der Menge D ausgewählt werden. Dabei können die Dokumente mehrfach selektiert werden. Dieser Vorgang wird B-mal wiederholt (zum Beispiel R=100). Die resultierenden neuen Trainingsmengen $D_1, \ldots, D_R$ enthalten einige Dokumente d_j mehrfach und andere überhaupt nicht (für $m = n$ enthält D_r im Mittel 64 % der Elemente von D).

Für jede Trainingsmenge D_r wird ein Klassifikator $f_r : x \rightarrow c$ trainiert. Diese Klassifikatoren unterscheiden sich in einem Ausmaß, welches die zu erwartenden Variationen bei neuen Trainingsmengen widerspiegelt. Soll ein neues Dokument x^{neu} klassifiziert werden, so berechnet man die Klassifikation $f_b(x^{\text{neu}})$ für jeden Klassifikator und bestimmt die endgültige Klassifikation durch Mehrheitsentscheidung.

Es lässt sich zeigen, dass der Mehrheitsklassifikator des Bagging in den allermeisten Fällen eine höhere Genauigkeit aufweist als der Ausgangsklassifikator [Hastie et al., 2011, S. 285]. Zudem kann man aus der Streuung der einzelnen Klassifikationsergebnisse $f_b(x^{\text{neu}})$ auf die Unsicherheit der Gesamtklassifikation schließen. Bagging kann auf jeden Klassifikator angewendet werden, um dessen Performanz zu erhöhen, wobei allerdings der Rechenbedarf auf ein Vielfaches steigt. Besonders vor-

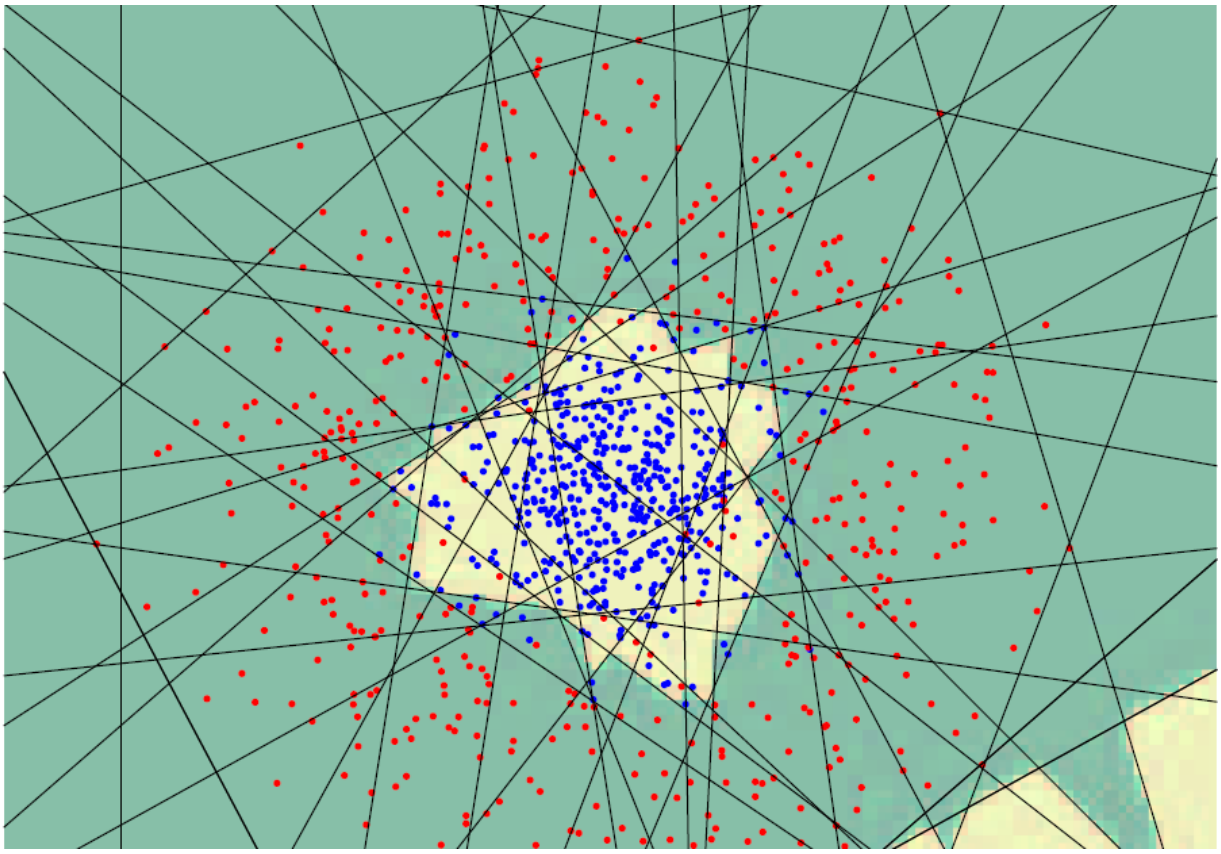


Abbildung 4.9
Von AdaBoost erzeugter Klassifikator als gewichtetes Mittel von $R = 40$ Einzelklassifikatoren. Aufgabe war die Trennung der roten von den blauen Punkten.

teilhaft ist das Ergebnis bei Entscheidungsbäumen und logistischen Klassifikatoren, während der Gewinn bei SVMs meist gering ist.

Boosting Einen ähnlichen Ansatz zur Kombination von Klassifikatoren verfolgt das Boosting [Freund und Schapire, 1997]. Es wird wieder ein Gesamtklassifikator F als gewichtetes Mittel von R Einzelklassifikatoren f_r bestimmt

$$F_R(x) = \text{sgn}\left(w_1 f_1(x) + \cdots + w_R f_R(x)\right)$$

wobei wieder $\text{sgn}(u) = 1$ falls $u > 0$, $\text{sgn}(u) = -1$ falls $u < 0$ und $\text{sgn}(u) = 0$ falls $u = 0$. Die w_b sind reellwertige Gewichte der einzelnen Klassifikatoren f_b.

Bei der Variante AdaBoost [Freund und Schapire, 1995] werden sukzessive die Klassifikatoren f_r bestimmt. Im r-ten Schritt gibt es die Klassifikatoren $f_1(x), \ldots, f_{r-1}(x)$ und die entsprechenden Gewichte. Vor jedem Durchlauf wird beurteilt, welche Dokumente d_i mit dem bislang erstellten Gesamtklassifikator $F_{r-1}(x)$ gut eingeordnet werden können und welche nicht. In jedem Schritt werden die Fallgewichte der falsch klassifizierten Dokumente in der Trainingsmenge erhöht und die Gewichte der korrekt klassifizierten Dokumente erniedrigt. Anschließend wird ein neuer Klassifikator f_r mit der so gewichteten Trainingsmenge trainiert. Der neue Klassifikator konzentriert sich also stärker auf die bisher falsch klassifizierten Dokumente. Das Verfahren bricht ab, wenn sich keine weitere Verbesserung ergibt.

Auf Grund theoretischer Ergebnisse lässt sich zeigen [Hastie et al., 2011, S. 337 ff.], dass AdaBoost gute Generalisierungseigenschaften besitzt und wie die SVM den Margin maximiert. AdaBost kann

als ein Verfahren zur Merkmalsauswahl mit einer theoretisch fundierten Auswahlstrategie betrachtet werden. In der praktischen Anwendung führt es zu durchweg guten Ergebnissen. Abbildung 4.9 zeigt, wie der Gesamtklassifikator relativ komplexe Klassengrenzen nachbilden kann.

4.4.3 Unterstützende Modelle

Klassifikationsverfahren bewerten konstruktionsbedingt den Inhalt eines Dokuments als Ganzes. Häufig möchte man aber auch die einzelnen jugendschutzrelevanten Wörter und Phrasen in einem Text erkennen. Dieses Vorgehen wird Informationsextraktion genannt. Dies schließt auch die Erkennung von inhaltlichen Relationen zwischen Wörtern und Phrasen ein. Beispielsweise geht es in dem folgenden Satz »… genau deswegen schneide ich an den Beinen …« um die Selbstverstümmelung an den Beinen und nicht um das harmlose Schneiden von Gemüse.

Verfahren der Informationsextraktion bewerten die Bedeutung einzelner Wörter und Phrasen. Sie erzeugen Merkmale, welche bei der Klassifikation berücksichtigt werden können.

4.4.3.1 Sprachmodelle

Ein Sprachmodell ordnet einer Folge von Wörtern $w_1, \ldots, w_n$ eine Wahrscheinlichkeit zu, die abschätzt, wie häufig diese Wortsequenz in einer Sprache vorkommt. Oft kommen durchaus plausible Wortfolgen von vier und mehr Wörtern in einer Trainingsmenge nicht vor. Daher ist es schwierig, die Wahrscheinlichkeit langer Wortfolgen abzuschätzen. Ausgangspunkt ist oft das n-Gramm Modell

$$p(w_1, \ldots, w_m) = \prod_{i=1}^{m} p(w_i | w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} p(w_i | w_{i-n+1}, \ldots, w_{i-1})$$

Es werden also nur Wortfolgen der Länge n betrachtet und die bedingten Wahrscheinlichkeiten aus den beobachteten Häufigkeiten dieser n-Gramme geschätzt. Bei $n = 2$ beziehungsweise $n = 3$ spricht man von Bigramm- beziehungsweise Trigramm-Modellen.

Allerdings können die n-Gramm Wahrscheinlichkeiten oft nicht direkt aus den beobachteten Häufigkeiten abgeleitet werden, da viele n-Gramme nicht in der Trainingsmenge enthalten sind, obwohl sie durchaus eine Wahrscheinlichkeit größer als Null haben. In diesem Fall werden unterschiedliche Glättungsverfahren durchgeführt, um auch den plausiblen unbeobachteten n-Grammen eine positive Wahrscheinlichkeit zuzuordnen [Manning und Schütze, 2001, S. 191].

Sprachmodelle unterscheiden sich für unterschiedliche Domänen oder Klassen. Hat man zwei Klassen A und B, so kann man für jede der beiden Klassen ein Sprachmodell erstellen. Ein neues Dokument d kann man dann der Klasse A zuordnen, wenn $p(d|A) > p(d|B)$ ist. Die folgenden Abschnitte beschreiben Varianten dieses Schemas.

4.4.3.2 Dynamische Markov-Ketten (DMK)

Anstelle der Abfolge von von Wörtern kann ein Sprachmodell auch die Abfolge der Bits in der ASCII-Darstellung eines Dokuments repräsentieren. Analog zu Abschnitt 4.4.3.1 kann man die Folge der Bits $(v_1, \ldots, v_n)$ eines Dokuments darstellen durch

$$p(v_1, \ldots, v_n) = \prod_{i=1}^{n} p(v_i | v_1, \ldots, v_{i-1}) \approx \prod_{i=1}^{n} p(v_i | v_{i-n+1}, \ldots, v_{i-1})$$

Der Vorteil dieser Darstellung ist, dass man in der bedingten Verteilung $p(v_i|v_{i-n+1}, \ldots, v_{i-1})$ nur zwei Zustände $v_i = 0$ und $v_i = 1$ prognostizieren muss. Dynamische Markov-Ketten (DMK) sind eine Darstellung dieser bedingten Verteilungen, welche mit zusätzlichen Daten immer genauer wird. Das Verfahren zählt die relativen Häufigkeiten aus, mit der nach einer bestimmten Bitfolge $v_1, \ldots, v_{i-1}$ eine 0 oder eine 1 auftaucht. Stellt es fest, dass zwischen den Häufigkeiten von 0 und 1 ein signifikanter Unterschied besteht, so wird der Wert von $p(v_i|v_1, \ldots, v_{i-1})$ durch die gefundenen Häufigkeiten geschätzt und als neuer Term dem Modell hinzugefügt. Je nach Signifikanzschwelle können die gefundenen Modelle sehr groß werden. Sie wurden erfolgreich auf die Textklassifikation und die Klassifikation von Spam-E-Mails angewendet [Bratko et al., 2006].

Werden sehr viele Trainingsdaten verwendet, können die Modelle sehr groß werden. In [Bergholz et al., 2009] wurde eine Technik zur Verringerung dieses Wachstums entwickelt, welche in der Praxis sehr gute Ergebnisse zeigt. In [Bergholz et al., 2010] wurde die Likelihood der dynamischen Markov-Ketten des Textes als ein zusätzliches Merkmal in der Textklassifikation verwendet und damit hervorragende Ergebnisse erzielt (vgl. Tabelle 4.5). Ein zusätzlicher Genauigkeitsgewinn ergab sich durch DMK-Modelle des Textes, der in Weblinks enthalten war.

4.4.3.3 Themenmodelle

Die Themenanalyse betrachtet ein Dokument als eine Mischung von verschiedenen zugrundeliegenden Themen (engl.: topics). Es wird ein mathematisches Modell formuliert, welches die Wörter eines Dokuments in überlappende Gruppen einteilt.

Latent Dirichlet Allocation Das wichtigste Verfahren in diesem Zusammenhang ist die Latent Dirichlet Allocation (LDA) [Blei et al., 2003], welche ein generatives wahrscheinlichkeitsbasiertes Modell für die Wörter eines Dokuments definiert.

Es geht davon aus, dass k die Anzahl der Themen (Topics) $t_1, \ldots, t_k$ und n_w die Anzahl der Wörter $w_1, \ldots, w_{n_w}$ in einem Korpus ist. Gemäß dem LDA-Verfahren ist die Wahrscheinlichkeit für Wort w_i im Dokument d gegeben durch die Mischungsverteilung

$$p(w_i|d) = p(w_i|t_1)p(t_1|d) + \cdots + p(w_i|t_j)p(t_j|d).$$

$p(t_j|d)$ ist die Wahrscheinlichkeit, dass Topic t_j in Dokument d auftaucht. $p(w_i|t_j)$ ist die Wahrscheinlichkeit, mit der das Wort w_i zu Topic t_j gehört. Die Wahrscheinlichkeitswerte $p(t_j|d)$ und $p(w_i|t_j)$ sind die freien Parameter des Modells, die aus den vorliegenden Dokumenten geschätzt werden müssen. Hierbei handelt es sich um sehr viele Parameter und die Anzahl der Parameter wird um so größer, je mehr Dokumente einbezogen werden. Um eine Überanpassung zu vermeiden, wird eine Regularisierung für die Parameter $p(t_j|d)$ und $p(w_i|t_j)$ durchgeführt. Hierzu wird die Dirichlet-Verteilung verwendet, welche speziell auf die Wahrscheinlichkeitswerte im Intervall $[0, 1]$ zugeschnitten ist und die Wahrscheinlichkeiten in Richtung eines mittleren Wertes »zieht«.

Es ist wichtig, dass im Gegensatz zu den Klassifikationsverfahren die LDA keine annotierten Daten benötigt. Ein stochastisches Trainingsverfahren, das Gibbs Sampling, hat sich als besonders effizient und genau erwiesen. Informell lässt sich sagen, dass diejenigen Wörter w_i zu einen Thema t_j gehören, die einen hohen Wahrscheinlichkeitswert $p(w_i|t_j)$ besitzen. Es zeigt sich, dass die Wörter eines Topics die Tendenz haben, häufig gleichzeitig in einem Dokument aufzutauchen. Es ergeben sich hierdurch oft überraschend intuitive Gruppen von Wörtern, welche auch nach Meinung von Experten zusammengehören.

In der Regel sind die Wahrscheinlichkeiten $p(w_i|t_j)$ für die allermeisten Topics extrem klein. Sind für ein Wort w_i die $p(w_i|t_j)$ in mehreren Topics relativ groß, so handelt es sich oft um die unterschiedli-

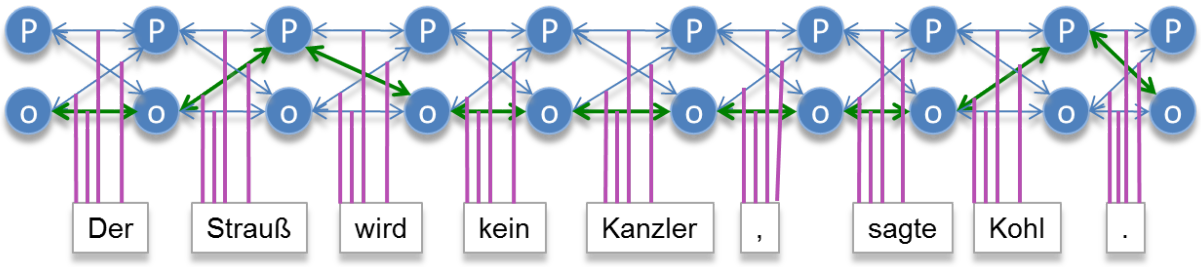


Abbildung 4.10
Verbindungsstruktur eines Conditional Random Fields (CRF) mit zwei Zuständen P (=Person) und o (=sonstiges). Jedes beobachtete Wort (untere Reihe) legt fest, welche »Verbindungsstärke« b_j zwischen den benachbarten Zuständen verwendet wird. Die breiteren Verbindungen entsprechen der Labelkette, welche die Personen »Strauß« und »Kohl« identifiziert.

chen Bedeutungen eines Wortes, zum Beispiel »Bank« im Finanzkontext und »Bank« als Möbel. Zur Charakterisierung eines Topics kann man oft die wahrscheinlichsten drei Wörter nehmen.

[Thongsuk et al., 2011] prognostizieren für jedes Wort das wahrscheinlichste Thema und verwenden dieses als ein zusätzliches Merkmal bei der Klassifikation von Twitter-Nachrichten. Anschließend klassifizieren sie die Nachrichten in die drei Klassen »Fluglinie«, »Essen« und »Computer und Technologie«. Die Trainingsmenge umfasst 1 000 Nachrichten. Mit 50 LDA-Topics ergibt sich ein F-Wert von 95,7 %, während sich ohne die zusätzlichen LDA-Merkmale ein F-Wert von 77 % ergibt. Dies zeigt, dass die Topic-Merkmale direkt als Zusatzmerkmale in die Klassifikation eingehen können.

Klassen-Themenmodell Das LDA-Modell kann in verschiedene Richtungen erweitert werden. Für diese Studie relevant ist die Einbeziehung von möglicherweise verfügbaren Klasseninformationen von Dokumenten, also zum Beispiel »Jugendschutzrelevant« und »Unbedenklich«. In [Bergholz et al., 2008] wurde eine Variante der LDA, das Klassen-Themenmodell (CLTOM), entwickelt. Dieses Verfahren erzeugt Wort-Cluster, welche die Klasseninformationen bei der Bildung der Themen berücksichtigt. Es ist sehr ähnlich zu dem L-LDA Verfahren [Ramage et al., 2009].

4.4.3.4 Extraktion von Phrasen

Oft tauchen Phrasen eines bestimmten Typs innerhalb eines Satzes in einem charakteristischen Kontext auf. Ein Beispiel sind Namen von Entitäten, zum Beispiel Personen, Organisationen und Orte. Zu deren Erkennung wird ein Annotationsproblem für Wörter formuliert: Das Wort, bei dem der Personenname anfängt, erhält das Label B-PER, weitere Wörter des Personennamens erhalten das Label I-PER. Dies wird für alle interessierenden Arten von Entitäten durchgeführt; Token außerhalb von Namen erhalten das Label o. Der Satz »Larry Ellison sagte, dass Oracle Sun übernehmen wird.« kann zum Beispiel sollte also folgende Annotation erhalten:

B-Per	I-Per	o	o	o	B-Org	B-Org	o	o	o
Larry	Ellison	sagte	,	dass	Oracle	Sun	übernehmen	wird	.

Es stellt sich nun die Aufgabe, für ein neues Dokument diese Label zu prognostizieren. Als Modell kann man hierzu ein Conditional Random Field (CRF) nutzen [Lafferty et al., 2001], welches für jedes Wort die Wahrscheinlichkeit der einzelnen Label modelliert. Hierbei bezeichnet man die Wörter eines Satzes zusammen mit ihren Eigenschaften (zum Beispiel Wortart) durch Vektoren x_t, so dass sich der ganze Satz als Folge von Vektoren $x = (x_1, \ldots, x_m)$ darstellen lässt. Die zugehörigen Label

bezeichnet man durch $y = (y_1, \ldots, y_m)$. Weiter unterstellt man, dass sich benachbarte Klassenlabel y_t, y_{t+1} gegenseitig statistisch beeinflussen.

Damit gelangt man zu einem Modell, welches die bedingte Wahrscheinlichkeit der Label $p(y_1, \ldots, y_m|x)$ für einen gegebenen Satz x modelliert. Es ist abhängig von den Faktoren $g_j(y_t, y_{t-1}, x)$, welche die Wahrscheinlichkeiten beeinflussen. Sie hängen von zwei benachbarten Label y_t, y_{t-1} sowie von allen beobachteten Wörtern und Merkmalen x des Satzes ab. Einige der $g_j(y_t, y_{t-1}, x)$ können auch nur vom laufenden Zustand y_t beeinflusst werden. Als Merkmale innerhalb der Merkmalsvektoren können die Wörter selbst, deren syntaktische Wortart, Stammform, sowie Vorsilben und Endungen verwendet werden. Weiter kann man Formeigenschaften (zum Beispiel Großschreibung oder Bindestriche) sowie Vorsilben und Endungen als Merkmale verwenden. Zudem kann man als Merkmal aufnehmen, ob ein Wort in einer Liste, zum Beispiel von Ortsnamen, enthalten ist. Weiterhin kann man diese Merkmale für die benachbarten Wörter berücksichtigen. Damit ergeben sich Millionen unterschiedlicher Merkmale mit jeweils zugehörigen freien Parametern. Abbildung 4.10 zeigt, wie bei dem meist verwendeten sequentiellen Conditional Random Field (CRF) die beobachteten Wörter x_t die Zustandsvariablen y_t, y_{t-1} beeinflussen. Hierbei wurden der Einfachheit halber nur zwei Zustände P und o verwendet.

Nach dem Maximum-Likelihood-Prinzip werden die freien Parameter durch Optimierungsverfahren so eingestellt, dass die Trainingsdaten optimal reproduziert werden. Hierbei werden die Parameterwerte regularisiert, also die Parameter in Richtung Null bewegt, so dass zufällige Schwankungen im Trainingsmaterial weitgehend ignoriert werden. Das Training ist ein aufwändiger Optimierungsprozess, der mehrere Stunden Rechenzeit beanspruchen kann.

Der Vorteil von CRF-Modellen liegt darin, dass die Sequenz der Wörter eines ganzen Satzes genutzt wird, um die Label zu prognostizieren. Für die Klassifikation von jugendschutzrelevanten Dokumenten könnte das Modell auf folgende Weise genutzt werden.

- Bilde Mengen D_J und D_U von annotierten Dokumenten mit den Klassen »Jugendschutzrelevant« und »Unbedenklich«.
- Bilde eine Liste von zur Unterscheidung relevanter Wörter, zum Beispiel
  - durch manuelle Zusammenstellung von charakteristischen jugendschutzrelevanten Ausdrücken,
  - durch die Selektion von Wörtern, die sich hinsichtlich ihrer Vorkommenshäufigkeit in D_J und D_U signifikant unterscheiden. Dies kann zum Beispiel mit einem Chiquadrat-Test erfolgen.
- Annotiere das Vorkommen der Wörter aus der Liste in D_J mit dem Label J und in D_U mit dem Label U. Alle anderen Wörter werden mit o gelabelt.
- Trainiere ein CRF auf dem Trainingskorpus $D_J \cup D_U$. Das Modell lernt dann, ob ein Wort eher in Kontext »Jugendschutzrelevant« oder »Unbedenklich« verwendet wurde.
- Füge die prognostizierten Wahrscheinlichkeit der Label in D_J und D_U als Wortmerkmal hinzu und trainiere einen Klassifikator.

Der Klassifikator kann dann dieses Merkmal als ein zusätzliches Indiz für die Klassifikation verwenden.

4.4.4 Frei verfügbare Verfahren zur Textanalyse

Es gibt mittlerweile eine größere Anzahl von frei verfügbaren Implementationen von Verfahren zur Textanalyse. Im Vordergrund stehen dabei die Programmiersprachen Python und Java.

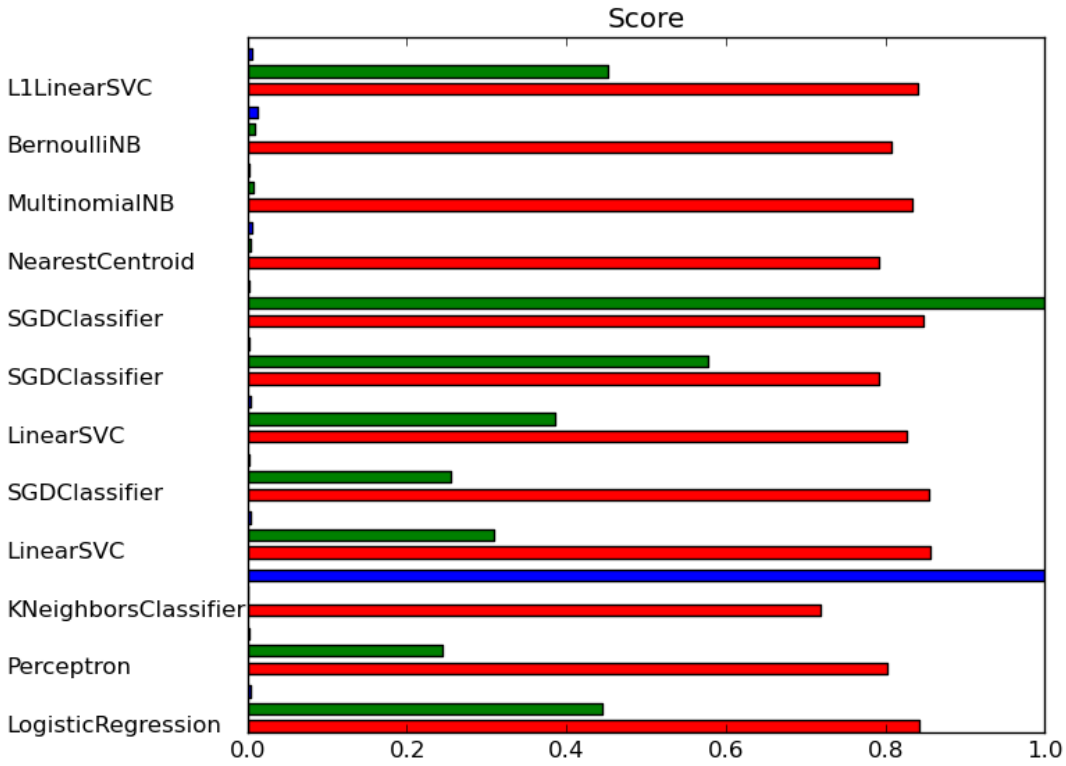


Abbildung 4.11
Klassifikationsergebnisse für den 20-Newsgroups-Korpus mit zwölf verschiedenen Klassifikatoren. Für jedes Verfahren wird angegeben: die zur Klassifikation der Testmenge benötigte Zeit (blauer Balken), die Trainingszeit (grüner Balken) und der F-Wert (roter Balken).

4.4.4.1 Python

Das Natural Language Toolkit (NLTK)[6] ist eine Sammlung von Verfahren der Computerlinguistik und Sprachverarbeitung für die Programmiersprache Python. NLTK ist quelloffen und ist unter der Apache-Lizenz verfügbar. Es enthält Verfahren zur Vorverarbeitung, Tokenisierung, Erkennung von Mustern und Parsing. Es gibt verschiedene Klassifikatoren (Entscheidungsbaum, logistischer Klassifikator). Die Verfahren von NLTK wurden von den Autoren in einem Lehrbuch dokumentiert [Bird et al., 2009].

Zur Vorverarbeitung kann man den Treetagger[7] nutzen, mit dem sich Wortartinformationen, Stammformen und Phraseninformationen (Chunks) ableiten kann. Der Treetagger ist in C programmiert und hat Wrapper für Python und Java. Allerdings ist das Programm nur zur Evaluation, Forschung und Lehre freigegeben. Bei anderen Verwendungen muss man bei dem Autor anfragen.

Scikit-learn ist eine frei verfügbare Bibliothek für die Python Programmiersprache.[8] Sie enthält unter anderem die folgenden Verfahren: Logistischer Klassifikator, Support Vektor Maschine, k-Nächste-Nachbarn, Bayes-Klassifikator, Entscheidungsbäume, Ensemble Methoden (zum Beispiel Bagging, Boosting, Random Forest), Kreuzvalidierung. Die Bibliothek ist Open Source Software und kommerzi-

[6] siehe http://www.nltk.org, Stand 1.12.2012
[7] siehe http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger, Stand 1.12.2012
[8] siehe http://scikit-learn.org, Stand 1.12.2012

ell frei verwendbar (BSD-Lizenz). Es gibt ein instruktives Demoprogramm, welches zehn verschiedene Textklassifikatoren auf einen großen Korpus anwendet und die Ergebnisse vergleicht.[9] Die Ergebnisse werden in Abbildung 4.11 dargestellt.

CRF++ ist eine Open Source Implementation[10] von Conditional Random Fields in C++. Sie kann für Aufgaben wie Wortart-Erkennung oder Phrasenextraktion verwendet werden. Es gibt Wrapper für Python und Java. Gensim ist ein freies Python Framework[11] zur Berechnung von Topic Modellen. Gensim ist unter einer LGPL-Lizenz verfügbar.

4.4.4.2 Java

Die OpenNLP Bibliothek ist eine Java-Programmsammlung[12] zur Verarbeitung von Textdokumenten. Sie unterliegt der liberalen Apache Lizenz. Sie unterstützt die häufigsten NLP-Aufgaben, wie Tokenisierung, Satzzerlegung, Wortart-Erkennung und die Extraktion von Phrasen. Die Module sind trainierbar, falls geeignete Trainingsdaten zur Verfügung stehen. Zur Klassifikation mit logistischen Klassifikatoren gibt es ein eigenes MaxEnt-Paket.[13]

LingPipe[14] ist eine Suite von Java-Modulen zur Verarbeitung von Text und enthält Methoden zur Satzzerlegung, Wortart-Erkennung, Erkennung der Sprache, Klassifikation (logistischer Klassifikator), Conditional Random Fields, Extraktion von wichtigen Schlagwörtern usw. Es ist recht ausgereift und bietet gut ausgearbeitete Tutorien zu den einzelnen Methoden. Es ist unter einer »royalty-free commercial license« verfügbar.

Der Stanford Parser und Part-of-Speech Tagger ist ein Java-Paket zur Syntaxanalyse.[15] Es enthält einen Satzzerleger, einen Tokenizer, einen Wortart-Erkenner mit deutschem Sprachmodell und einen statistischen Parser für die deutsche Sprache. Es ist unter einer GNU GPL Lizenz verfügbar.

Mallet ist eine Java Programmsammlung[16] zur Dokumentverarbeitung. Sie enthält Module zur Klassifikation (Bayes-Klassifikator, logistischer Klassifikator, Entscheidungsbaum), Conditional Random Fields, sowie zum Topic Modelling. Sie ist unter einer liberalen Common Public License verfügbar. Mallet ist recht ausgereift, aber nur wenig dokumentiert.

4.4.4.3 Diskussion

Die oben zusammengestellten Open Source Programmsammlungen zeigen, dass es möglich ist, mit geringem Aufwand eine Klassifikatorlösung zu erstellen. Den ausgereiftesten Eindruck macht dabei Scikit-Learn, welches in einer Demo zehn verschiedene Textklassifikatoren auf einen großen Korpus anwendet, die Ergebnisse vergleicht und sogar eine Merkmalsselektion anbietet. Erforderlich sind natürlich annotierte Daten, zum Beispiel für die Klasse »Unbedenklich« und die verschiedenen Klassen von »Jugendschutzrelevant«. Wie gut die Performanz der Klassifikatoren in unterschiedlichen Literaturstudien war, wird im nächsten Abschnitt beschrieben.

[9] siehe http://scikit-learn.org/stable/auto_examples/document_classification_20newsgroups.html, Stand 1.12.2012
[10] siehe http://code.google.com/p/crfpp, Stand 1.12.2012
[11] siehe http://radimrehurek.com/gensim, Stand 1.12.2012
[12] siehe http://opennlp.apache.org, Stand 1.12.2012
[13] siehe http://sourceforge.net/projects/maxent, Stand 1.12.2012
[14] siehe http://alias-i.com/lingpipe, Stand 1.12.2012
[15] siehe http://nlp.stanford.edu/software, Stand 1.12.2012
[16] siehe http://mallet.cs.umass.edu, Stand 1.12.2012

Tabelle 4.4
Vergleich von Textklassifikatoren auf zwei Dokumentensammlungen.

Klassifikationsverfahren	Abschnitt	F-Wert in %
Reuters Korpus		
Bayes-Klassifikator	4.4.2.1	80
Entscheidungsbaum	4.4.2.4	79
Neuronales Netz	4.4.2.3	84
k nächste Nachbarn	4.4.2.6	86
Support Vektor Maschine	4.4.2.5	87
AdaBoost	4.4.2.7	88
20 Newsgroups Korpus		
k nächste Nachbarn	4.4.2.6	72
Bayes-Klassifikator	4.4.2.1	83
logistischer Klassifikator	4.4.2.2	84
Support Vektor Maschine	4.4.2.5	86

4.4.5 Vergleich von Textklassifikatoren

Die Leistungsfähigkeit von Textklassifikatoren hängt immer auch von der jeweiligen Einsatzdomäne ab. Es ist daher üblich, die Leistungsfähigkeit an Hand vorgegebener Benchmark-Daten zu vergleichen.

Klassifikation von Agenturmeldungen und Newsgruppenbeiträgen Ein umfangreicher Vergleich wurde von [Sebastiani, 2002] durchgeführt. Er untersucht die Performanz einer größeren Anzahl von Klassifikatoren auf dem frei zugänglichen Reuters-Benchmarkdatensatz. Dieser enthält 12 902 englische Meldungen der Nachrichtenagentur Reuters, welche in 90 thematische Klassen eingeordnet wurden. Die Trainingsmenge besteht aus 9 603 Dokumenten und die Testmenge aus 3 299 Dokumenten. Der obere Teil von Tabelle 4.4 aus [Sebastiani, 2002, S. 38] gibt die besten Ergebnisse für jede Verfahrensgruppe wieder. Dabei verwendeten die Klassifikatoren die gleichen Eingabemerkmale x.

Die Tabelle zeigt einen deutlichen Unterschied zwischen den »einfachen Verfahren« Bayes-Klassifikator und Entscheidungsbaum und den komplexeren Verfahren Support Vektor Maschine und AdaBoost mit expliziter Regularisierung. Unabhängig von den gewählten Merkmalen ergibt sich bei Verwendung letzterer Methoden meist geringere Fehler.

Der untere Teil von Tabelle 4.4 gibt die Ergebnisse zur Klassifikation von Beiträgen aus 20 Newsgruppen wieder. Im Gegensatz zum Reuterskorpus enthalten die Beiträge viele Grammatik- und Rechtschreibfehler sowie Slangausdrücke. Die Klassifikation wurde mit dem Demoprogramm Scikit-learn durchgeführt (vgl. Abschnitt 4.4.4.1). Die Trainingsmenge enthielt 11 314 Dokumente, etwa 550 Dokumente pro Klasse. Der untere Teil von Tabelle 4.4 zeigt die Ergebnisse für einzelne Klassifikationsverfahren. Die lineare Support Vektor Maschine schneidet am besten ab. Abbildung 4.11 stellt die Performanz aller Klassifikatoren dar. Es zeigt sich, dass es eine Reihe von Klassifikatoren gibt mit etwas schlechterer Performanz als die SVM. Die Abbildung zeigt auch die relativen Trainings- und Testzeiten für die einzelnen Verfahren. Die lineare SVM benötigt zum Beispiel 7,126 Sekunden zum Training mit den 11 314 Trainingsdokumenten und 0,053 Sekunden für die Klassifikation der 7 532 Dokumente der Testmenge.

E-Mail-Klassifikation nach Spam mit unterschiedlichen Merkmalen In einer umfangreichen Studie [Bergholz et al., 2010] wurde die Performanz unterschiedlicher Arten von Eingabemerkmalen auf einem öffentlich verfügbaren Benchmarkdatensatz von etwa 7 900 E-Mails untersucht. Aus-

Tabelle 4.5
Klassifikationsergebnisse für einen öffentlich verfügbaren Spam Corpus bei unterschiedlichen Merkmalen.

Merkmale	F-Wert in %
Basismerkmale	95,9
Basismerkmale + Dynamische Markov-Ketten	97,9
Basismerkmale +Themenmodell	97,8
Basismerkmale + DMC + Topic	98,9
Merkmalsauswahl: Basismerkmale + DMC + Topic	99,5

gangspunkt war die Verwendung von 27 Basismerkmalen wie der Anzahl der unterschiedlichen HTML-Gruppen, die Anzahl und die Struktur der Links, das Vorhandensein von Java-Scripts sowie die Zugehörigkeit zu unterschiedlichen Wortlisten. Zusätzlich wurden die Ergebnisse unterstützender Modelle verwendet: dynamische Markov-Ketten (vgl. Abschnitt 4.4.3.2) und Klassen-Themenmodelle (vgl. Abschnitt 4.4.3.3).

Tabelle 4.5 zeigt die erzielten F-Werte unter Verwendung unterschiedlicher Merkmale bei 5-fach Kreuzvalidierung. Es zeigt sich, dass die verschiedenen Merkmale unterschiedliche Aspekte der E-Mails beschreiben und daher jeweils das Klassifikationsergebnis verbessern. Dabei ist hervorzuheben, dass bei einem F-Wert von 99,49 % nur noch wenige E-Mails falsch klassifiziert werden.

Klassifikation von Texten in Hinblick auf Internet-Mobbing Der Begriff Internet-Mobbing (auch Cyberbullying sowie Cyber-Stalking) bezeichnet verschiedene Formen der Diffamierung, Belästigung, Bedrängung und Nötigung anderer Menschen oder Firmen mit Hilfe elektronischer Kommunikationsmittel über das Internet, in Chatrooms, beim Instant Messaging und/oder auch mittels Mobiltelefonen.

Die Studie [Ptaszynski et al., 2010] untersucht einen Datensatz, der in japanischen Sekundarschulen gesammelt wurde. Sie verwendet als Merkmal die Zugehörigkeit eines Wortes zu einer Liste von 216 vulgären Wörtern. Mit Hilfe des Levenshtein-Abstandes wurde die Ähnlichkeit zwischen Wörtern berechnet, um Internetjargon und abgewandelte Schreibweisen zu berücksichtigen. Als Klassifikator wurde eine SVM verwendet (vgl. Abschnitt 4.4.2.5). Für einen Datensatz von 966 annotierten Fällen (750 Beispiele von Internet-Mobbing) ergaben sich bei einer 10-fach Kreuzvalidierung folgende Ergebnisse: 79,9 % Präzision, 98,3 % Trefferquote und 88,2 % F-Wert.

Klassifikation von Webseiten in Jugendschutzkategorien In einer anderen Studie wurden 3,1 Millionen manuell annotierte Internetseiten und 12 000 Merkmale verwendet [Erdmann et al., 2012]. Die Seiten wurden in 27 Kategorien und 103 Unterkategorien eingeteilt, zum Beispiel in die Kategorien Einkauf (Unterkategorien: Auktionen, allgemeines Einkaufen, Immobilienkauf, Kauf von Computerartikeln), Hobby (Unterkategorien: Musik, Prominente, Kochen, allgemeine Erholung), Illegal (Unterkategorien: Terror und Extremismus, Waffen, Beleidigung, Selbstmord, Durchbrennen), Erotik (Unterkategorien: Sex, Nacktheit, Prostitution, Pornosuchmaschine).

Aus diesen annotierten Daten wurden fünf unterschiedliche Trainingsmengen erzeugt, indem einzelne Kategorien als harmlos beziehungsweise unerlaubt eingeordnet wurden. So ergaben sich fünf Klassifikationsprobleme von restriktiv bis liberal. Es wurde eine SVM mit der radialen Basisfunktion als Kernfunktion verwendet (vgl. Abschnitt 4.4.2.5). Hierbei wurde eine Genauigkeit von 91,2 % auf der Testmenge erzielt. Diese Problemstellung entspricht direkt der Aufgabe der Klassifikation von Webseiten in jugendschutzrelevante Kategorien. Das Niveau der Ergebnisse sollte daher auf die Fragestellung dieser Studie übertragbar sein.

Klassifikation von Kinderseiten in einem Webverzeichnis [Eickhoff et al., 2011] führten eine umfangreiche Studie zur Klassifikation von Webseiten für Kinder durch. Als Beispiele für Kinderseiten wählten sie 900 Webseiten aus den »Kids and Teens« Webseiten des Open Directory Projects[17] und als Alternativklasse 900 Webseiten mit Inhalten für Erwachsene. Zudem wurde jede der 1 800 Seiten durch je fünf menschliche Annotatoren bewertet.

Neben den üblichen schon bisher diskutierten Merkmalen wurden in der Studie spezielle Merkmale verwendet, welche die »einfachen« Kinderwebseiten von den komplexeren »Erwachsenenwebseiten« abheben sollten:

- Eine Lesbarkeitsmaßzahl (engl.: readability score), welche die Komplexität des Satzbaus misst.
- Die Anzahl der Entitäten (Personen, Orte, usw.), die in einem Satz genannt werden.
- Ein Sprachmodell für Kinder, welches analysiert, ob die Abfolge der Wörter ähnlich zu der Abfolge bei anderen Kinderseiten ist (vgl. Abschnitt 4.4.3.1).
- Der Anteil der Begriffe, die nicht aus dem einfachen Kinderwortschatz kommen.
- Anteil der Kindersprache und lustigen Begriffe in den URLs.
- Anzahl und Art der Bilder sowie der Animationsskripte.

Als zusätzliche Merkmale berücksichtigten die Autoren die Klassifizierung der verlinkten Seiten: einerseits den Anteil der verlinkten Seiten, die als »Erwachsenenseite« klassifiziert wurden und andererseits den Anteil der Seiten, die als »Kinderseite« klassifiziert wurden.

In den Experimenten stellte sich heraus, dass Klassifizierung der Nachbarseiten den höchsten Informationsgehalt besaß, gefolgt von dem Sprachmodell für Kinder und den URL-Merkmalen. Der in der Studie entwickelte und trainierte Klassifikator erreichte einen $F_{1/2}$-Wert von 72 %. Bemerkenswert war, dass die menschlichen Annotatoren nur einen Wert von 75 % erreichten. Damit lag die Genauigkeit der maschinellen Lösung nahe an der menschlichen Klassifikationsgenauigkeit. Die Ergebnisse für die Klassifikation von Webseiten für junge Kinder war mit einem $F_{1/2}$-Wert von 79 % noch genauer.

4.4.6 Aktives Lernen

Täglich wird eine Vielzahl an neuen Webseiten ins Internet gestellt, welche für Jugendliche relevant sind. Dadurch ändert sich auch die Zusammensetzung der jugendschutzrelevanten Seiten und die Klassifikatoren müssen aktualisiert werden. Die periodische Erstellung von neuen Trainingsmengen ist dabei sehr aufwändig und ignoriert, dass ein Teil der Webseiten über lange Zeit aktuell bleibt. Als Alternative wurde das Aktive Lernen entwickelt, um sukzessive die Trainingsmengen und damit die Klassifikatoren anzupassen.

Das Aktive Lernen analysiert die neuen Webseiten kontinuierlich und wählt automatisch einen kleinen Bruchteil zur manuellen Annotation aus [Olsson, 2009]. Abbildung 4.12 skizziert den Informationsfluss beim Aktiven Lernen. Hierbei wählt das Verfahren Dokumente aus, deren Klassifikation mit dem aktuellen Klassifikator K_{t-1} besonders unsicher ist. Das sind einerseits Dokumente in der Nähe der Entscheidungsschwelle (zum Beispiel der Hyperebene bei der SVM, vgl. Abschnitt 4.4.2.5). Andererseits kann auch die Unsicherheit auf andere Art geschätzt werden, zum Beispiel mit Hilfe des Bagging-Verfahrens (vgl. Abschnitt 4.4.2.7). Aus diesen unsicheren Dokumenten wird eine Stichprobe zur manuellen Annotation ausgewählt.

[17] siehe http://www.dmoz.org/Kids_and_Teens/, Stand 1.12.2012

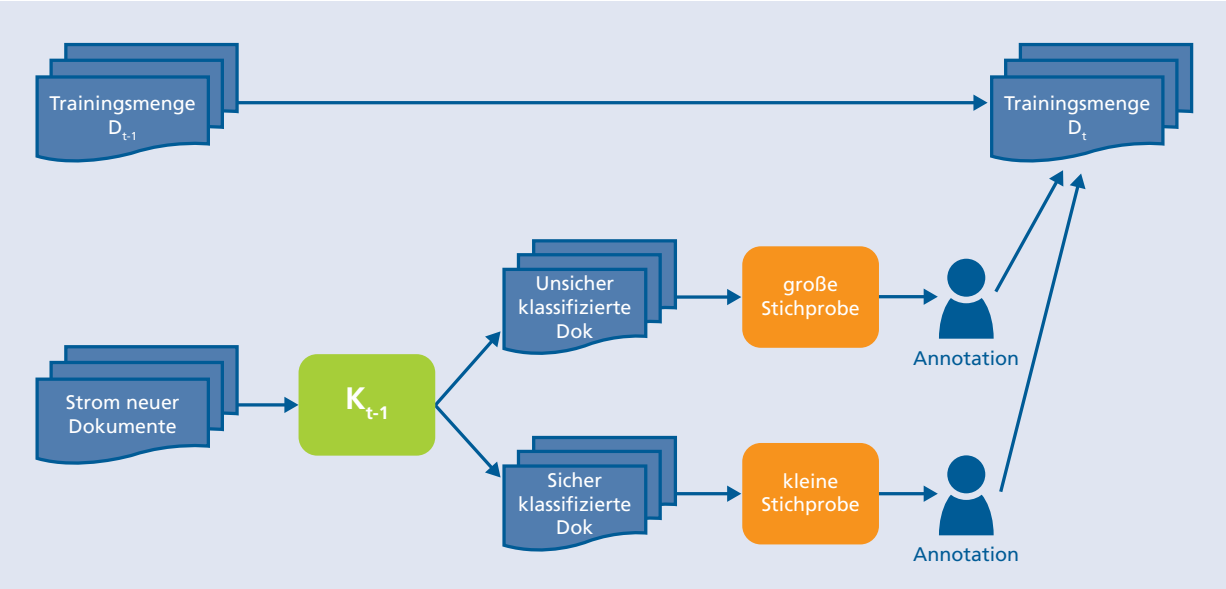


Abbildung 4.12
Informationsfluss beim Aktiven Lernen.

Treten komplett neue Arten von relevanten Dokumenten auf, so kann man diese nur erfassen, indem man von allen neuen Dokumenten eine noch kleinere Stichprobe zur Annotation selektiert. Die manuell annotierten Dokumente werden dann der alten Trainingsmenge D_{t-1} hinzugefügt und bilden die neue Trainingsmenge D_t. Mit dieser Trainingsmenge wird ein aktueller Klassifikator K_t bestimmt. Analog können auch »alte« Dokumente aus der Trainingsmenge entfernt werden, wenn sie beispielsweise keine Supportvektoren sind.

Dieses Verfahren wurde in [Bergholz, 2009] für die Klassifikation von Spam- und Phishing-E-Mails verwendet. Insgesamt wurden täglich 25 000 E-Mails gefiltert. Davon wurden täglich 500 (davon 100 zufällig) gemäß des Verfahrens von Abbildung 4.12 zur manuellen Annotation ausgewählt. In Abbildung 4.13 wird der Anteil der Falsch-Positiven und Falsch-Negativen E-Mails wiedergegeben. Die mittlere Rate der Falsch-Negativen sinkt von anfänglichen 26 % und schwankt danach von Tag zu Tag mit einem Mittelwert von 7,09 %. Die mittlere Rate der Falsch-Positiven ist mit 0,34 % deutlich geringer. Das Experiment mit insgesamt 500 000 untersuchten E-Mails zeigt, dass mit einem relativ geringen Aufwand von 500 annotierten E-Mails pro Tag der Fehler trotz Änderungen der E-Mail-Zusammensetzung unter 10 % gehalten werden kann. Die Rate der Falsch-Positiven ist dabei noch wesentlich geringer. Es ist zu erwarten, dass sich diese Zahlen in etwa auf die Aktive Lernen von Klassifikatoren von Webseiten für den Jugendschutz übertragen lassen.

[Sculley und Wachman, 2007] verwenden eine spezielle Version der Support Vektor Maschine, um kontinuierlich ankommende E-Mails zu klassifizieren. Mit diesem Ansatz erreichen sie ein sehr gutes Ergebnis bei dem Klassifikations-Vergleichstest Trec 2007 mit über 70 000 E-Mails.

4.5 Technische Verfahren für die Multimedia-Analyse

Das vorangegangene Kapitel und der letzte Abschnitt haben gezeigt, dass die deutlichsten Hinweise auf jugendschutzrelevante Inhalte im Netz durch textuelle Daten gegeben werden. Allerdings sind es meist die audiovisuellen Inhalte, die am verstörendsten sind.

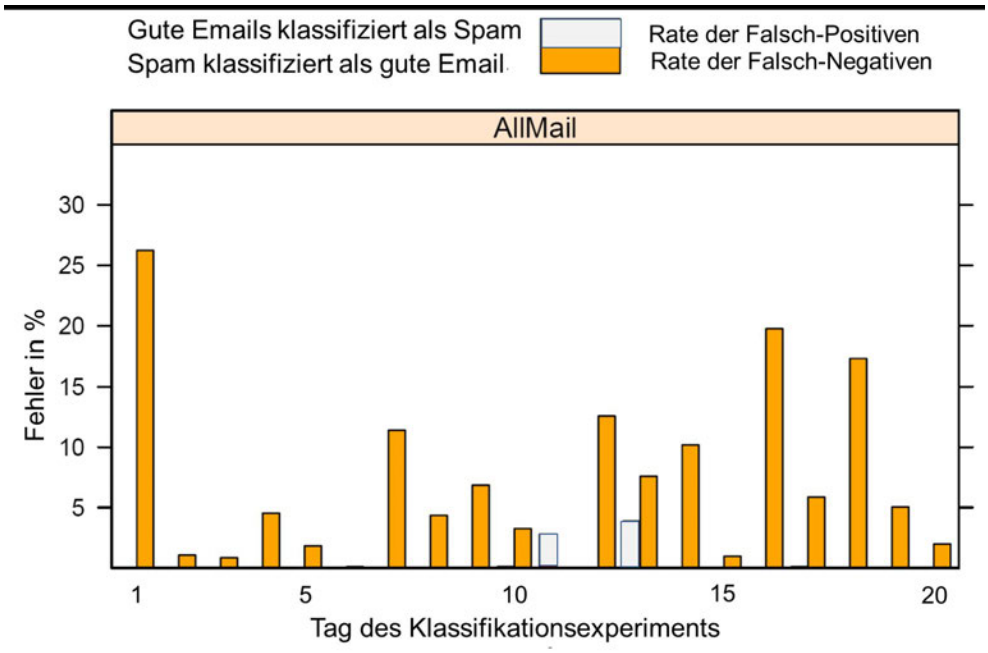


Abbildung 4.13
Klassifikation von Spam- und Phishing-E-Mails. Anteil der Falsch-Positiven und Falsch-Negativen E-Mails im Verlauf eines mehrtägigen Experiments.

Bei Inhalten, die hinreichend viel textuelle Beschreibung mitliefern, können auch die nicht-textuellen Inhalte auf Basis des Textes mit erkannt werden. Häufig finden sich jedoch audiovisuelle Inhalte, die mit wenig oder keiner textuellen Beschreibung ausgestattet sind, so zum Beispiel auf Videoplattformen wie YouTube. In diesem Abschnitt wird deshalb auf solche Verfahren eingegangen, die anhand des Inhaltes von Bildern, Videos und Tonaufnahmen Hinweise auf deren jugendschutzrelevanten Charakter geben können. Die entsprechenden Verfahren wurden durch intensive Analyse von Beispieldaten aus den Bereichen Gewalt (vgl. Abschnitt 4.1.2.2), Extremismus (vgl. Abschnitt 4.1.3.2) und Selbstgefährdung (vgl. Abschnitt 4.1.4.2) identifiziert. Die Tabellen 4.1, 4.2 und 4.3 geben eine Übersicht, welche Verfahren im jeweiligen Bereich geeignet erscheinen.

Diese lassen sich in der Anwendung in zwei Kategorien einteilen:

1. Verfahren, die bereits bekanntes Material aufgrund des Inhalts wiedererkennen, und
2. Verfahren, die durch eine Analyse des Inhalts Material als jugendschutzrelevant einstufen können.

Der ersten Kategorie sind vor allem Fingerprinting-Verfahren zuzuordnen, die in einem Index eine stark komprimierte Charakterisierung von Inhalten speichern und es erlauben, durch eine Suche auf diesem Index ähnliches Material wiederzuerkennen. Durch die starke Komprimierung erlauben diese Verfahren einerseits ein sehr schnelles Durchsuchen großer Mengen von Inhalten und andererseits eine gewisse Toleranz gegenüber Modifikationen von Inhalten und reduzieren gleichzeitig den Speicherbedarf erheblich.

Der zweiten Kategorie sind Verfahren zuzuordnen, die Informationen aus audiovisuellen Inhalten ableiten, das heißt inhaltliche Elemente erkennen. Durch ein technisches Verstehen der Inhalte kann dann entschieden werden, wie hoch die Wahrscheinlichkeit ist, dass die analysierten Inhalte jugendschutzrelevant sind. Solche Verfahren sind einerseits durch die automatische Spracherkennung und die optische Texterkennung gegeben. Diese helfen, gesprochene oder geschriebene Information aus Bild- oder Tondokumenten in digitalen Text umzuwandeln, der dann vom Computer analysiert wer-

den kann. Andererseits erlauben es Objekterkennung und die Erkennung semantischer Konzepte, Geschehen und Objekte in der visuellen Darstellung zu verstehen und ihrerseits analysierbar zu machen.

In den folgenden Abschnitten beschreiben wir die genannten Verfahren, stellen den Stand der Technik dar und untersuchen, wie diese Verfahren eventuell modifiziert oder erweitert werden müssen, um dem Jugendschutz dienen zu können.

4.5.1 Fingerprinting zur Wiedererkennung bekannter Inhalte

Das folgende Szenario tritt immer wieder auf: Bild- oder Video-Material wird als jugendschutzrelevant erkannt und eine Webseite, die das Material verbreitet, wird gesperrt oder der Inhalt indiziert. Allerdings werden immer wieder Kopien des Materials in teilweise modifizierter Form verbreitet. Häufig wird indiziertes Bild- oder Tonmaterial mit neuen Inhalten kombiniert. So ist das Material über viele Wege weiterhin für Jugendliche zugänglich. Eine technische Möglichkeit, dieses Problem zu behandeln, besteht darin, eine Datenbank mit bereits als jugendschutzrelevant bekanntem Material zu erstellen und alle Inhalte vor Auslieferung an Jugendliche zunächst mit diesem Material abzugleichen. Dies kann auch in Kombination mit der Erstellung von Blacklists geschehen, so dass einerseits Inhalte nicht bei jedem Zugriff erneut abgeglichen werden müssen und andererseits Varianten von Material, das bereits gelistet ist, ebenfalls automatisch erkannt werden können. Die Entscheidung zur Aufnahme in diese Datenbank kann sowohl durch menschliche Bewerter als auch durch algorithmische Verfahren oder durch die Kombination beider Ansätze erfolgen.

Genau an diesem Punkt setzten Fingerprinting-Verfahren an. Die technische Aufgabenstellung an diese Verfahren ist folgende: Es ist eine große Datenbank bekannter multimedialer Dokumente vorhanden. Für einen – eventuell modifizierten – Ausschnitt aus einem solchen in der Datenbank vorhandenen Dokument sollen diejenigen Dokumente der Datenbank gefunden werden, in denen der Ausschnitt ebenfalls vorkommt. Mögliche Modifikationen solcher Ausschnitte stellen zum Beispiel die wiederholte Codierung mit einem verlustbehafteten Komprimierungsverfahren (zum Beispiel JPEG, mp3) dar. Aber auch ein Verrauschen des Materials oder die Überlagerung mit anderem Material (zum Beispiel Textüberlagerungen in Videos, Überlagerung der Tonspur mit gesprochenem Text) kommen häufig vor. Verfahren, die dieses Problemfeld effizient behandeln, werden seit über einem Jahrzehnt intensiv erforscht und haben inzwischen den Übergang von der Forschung in die industrielle Anwendung gefunden.

Je nach vorliegendem Datenmaterial kommen unterschiedliche Techniken zum Einsatz, das heißt es existieren spezialisierte Verfahren für Bild- beziehungsweise Tonmaterial, die für die Erkennung von Videos geeignet kombiniert werden können. In den folgenden Abschnitten beschreiben wir für Verfahren zum Audio-Fingerprinting und zum Video-Fingerprinting jeweils den Stand der Technik und untersuchen dann, inwiefern dieser den Einsatz im Jugendschutz realistisch erscheinen lässt.

4.5.1.1 Audio-Fingerprinting

Das klassische Szenario für die Anwendung von Audio-Fingerprinting-Techniken ist folgendes: Ein Anwender sitzt im Auto oder in einer Bar und hört ein Lied, das ihm gut gefällt. Leider kennt er weder den Titel noch den Interpreten. Deshalb aktiviert er einen Dienst auf seinem Handy, der das Lied für ihn erkennt. Das Handy nimmt wenige Sekunden des Liedes auf und schickt es an einen Server. Dort wird das Lied analysiert und schnell mit einer sehr großen Datenbank von Liedern abgeglichen. Innerhalb von Sekundenbruchteilen wird das Lied erkannt und der Nutzer erhält alle vorhandenen

Informationen zu diesem Titel auf seinem Handy und möglicherweise auch direkt die Möglichkeit, das Lied zu erwerben.

Solche Anwendungen sind seit längerer Zeit kommerziell umgesetzt und entsprechende Dienste für viele mobile Endgeräte verfügbar. Die eingesetzten Verfahren sind so robust, dass sie schon anhand kurzer Ausschnitte ein Lied identifizieren können, selbst dann, wenn Hintergrundgeräusche, die Aufnahme mit dem Handy-Mikrofon und die verlustbehaftete Kodierung bei der Signalübertragung zum Server die Aufgabe deutlich erschweren.

Dieses Szenario zeigt, dass es möglich ist, bekanntes Material auch unter erschwerten Bedingungen sicher wiederzuerkennen. Dies ist genau die Eigenschaft, die nötig ist, um solche Verfahren auch für den Jugendschutz zu nutzen. Ist eine Menge als jugendschutzrelevant bekanntes Material, wie zum Beispiel Lieder rechtsradikaler Bands oder die Tonspur indizierter Filme, vorhanden, können immer wieder auftauchende Kopien dieses Materials erkannt werden. Dies ist auch dann möglich, wenn sie nur in Ausschnitten oder leicht modifiziert verwendet werden. Wie wir in diesem Abschnitt sehen werden, ist es dazu nicht nötig, das eigentliche Material zu speichern. Es genügt eine Datenbank von Merkmalen vorzuhalten, die das Material zwar hinreichend charakterisieren, aber erheblich weniger Speicherbedarf haben. Es ist hierbei nicht nötig, die eigentlichen jugendschutzrelevanten Daten in der Datenbank zu speichern. Stattdessen werden ausschließlich die genannten abstrakten Merkmale gespeichert, aus denen sich die ursprünglichen Daten nicht rekonstruieren lassen.

Um solche Anwendungen zu ermöglichen, basieren Audio-Fingerprinting-Verfahren auf folgendem allgemeinen Vorgehen: Zunächst müssen die Audiodaten, die später erkannt werden sollen, vorliegen. Sie werden dann analysiert und auf eine stark komprimierte Darstellung, den sogenannten Fingerprint oder Fingerabdruck, reduziert. Die Fingerprints werden dann in einer Datenbank gespeichert und stehen für den späteren Abgleich mit unbekanntem Material bereit. Weiteres Audiomaterial kann bei diesen Verfahren auch nachträglich hinzugefügt werden. Zur Identifikation eines Audiostücks wird für dieses ebenfalls ein Fingerprint berechnet und mit den Fingerprints in der Datenbank verglichen. Ein Fingerprinting-Verfahren wird dadurch charakterisiert, wie die Fingerprints bestimmt werden und wie deren Abgleich gelöst wird. Die Fingerprints sollten sich dabei, wenn sie aus dem gleichen Material oder einer leicht gestörten Version dieses Materials (s. u.) gewonnen werden, wenig unterscheiden. Außerdem sollten sie kompakt sein, damit wenig Speicherplatz nötig ist und sie schnell verglichen werden können. Darüber hinaus müssen die Fingerprints abschnittsweise erstellt werden, damit auch Ausschnitte identifiziert werden können. Der Abgleich der Fingerprints muss schließlich so realisiert sein, dass er sehr schnell durchgeführt wird, da in der Anwendung für gewöhnlich große Mengen von Material durchsucht werden müssen.

Die Berechnung der Fingerprints beruht auf Verfahren zur Merkmalsextraktion aus Audiodaten. Dazu werden Eigenschaften des Audiomaterials meist aus dem Zeit-Frequenz-Verlauf bestimmt. Hier kommen zum Beispiel das zeitliche Auftreten besonders hervorstechender Frequenzanteile oder die Verteilung der Signalenergie über das Spektrum zum Einsatz.

Verfahren zum Matching vergleichen schnell und fehlertolerant eine kurze Merkmalssequenz (den Fingerprint) mit einer großen Menge solcher Sequenzen. Als Beispiele geeigneter Techniken hierzu seien Hidden Markov Modelle, spezielle Varianten des Hashing und Weiterentwicklungen der invertierten Listen aus dem Textretrieval genannt.

In Abschnitt 4.5.1.2 stellen wir exemplarisch ein Fingerprinting-Verfahren vor, anhand dessen die Anwendbarkeit im Jugendschutz-Szenario in Abschnitt 5.3.2 des folgenden Kapitels untersucht wird.

Besonders relevant für die Anwendung von Fingerprinting-Verfahren ist die Frage nach ihrer Robustheit. Diese stand schon von Beginn der Entwicklung solcher Verfahren im Mittelpunkt. Im Folgenden

geben wir einen Überblick über die Art von Störungen, die toleriert werden sollten, und in welchem Maße dies von aktuellen Verfahren umgesetzt wird.

Ausschnittbildung und zeitliche Verschiebung: Das Ziel der meisten Anwendungen ist es, Ausschnitte von Musikstücken oder anderen Tonaufnahmen zu identifizieren. Daher sind praktisch alle Verfahren zum Audio-Fingerprinting darauf ausgelegt, gegenüber Ausschnittbildung und zeitlichen Verschiebungen robust zu sein. In der Praxis stellt sich hier die Frage, wie lang ein Anfrageausschnitt sein muss, damit er mit hoher Wahrscheinlichkeit identifiziert werden kann. Dabei gilt typischerweise, dass sich ein Ausschnitt umso sicherer identifizieren lässt, je länger er ist. Für ungestörtes Material reichen dabei in der Regel Anfragen von ein bis drei Sekunden Länge aus. Bei Material, dass durch Effekte gestört ist, wie sie in den nächsten Punkten beschrieben werden, lassen sich mit drei bis fünf Sekunden Material gute Ergebnisse erreichen.

Equalizer und Lautstärkeänderung: Zur Beeinflussung der Klangeigenschaften einer Tonaufnahme wird häufig die Lautstärke modifiziert. Dies geschieht für gewöhnlich in unterschiedlichen Frequenzbereichen unterschiedlich stark. Als Beispiel sei eine relative Verstärkung der Bässe genannt. Solche Modifikationen sind sehr häufig und sollten keinen merkbaren Einfluss auf die Erkennung von Tonaufnahmen haben. Dennoch stellt diese Art von Modifikation für manche Audio-Fingerprinting-Verfahren ein Problem dar, das allerdings durch geeignete Vorverarbeitung leicht in den Griff zu bekommen ist.

Audiokodierung und Bandbeschränkung: Audiomaterial wird über das Internet zumeist in komprimierter Form verbreitet, da so der Speicher- und Bandbreitenbedarf drastisch reduziert wird. Meist kommen dabei verlustbehaftete Verfahren zum Einsatz, bei denen nicht die volle Information, die in einer Aufnahme steckt, gespeichert wird, sondern nur der für den Menschen hörbare Teil. Dies verändert die Charakteristik des Materials, so dass die Identifikation erschwert wird. Häufig kommt es zudem zu Umkodierungen, wodurch zusätzliche Artefakte in die Aufnahme eingebracht werden. Dabei wird unter Umständen auch die Abtastrate verändert, was zu einer Beschränkung des Frequenzbereiches führt. Dies sind ebenfalls Effekte, die für die typischen Anwendungsszenarien des Fingerprinting charakteristisch sind und deshalb schon beim Entwurf der Verfahren mit berücksichtigt werden. Die Erkennungsraten beim Audio-Fingerprinting sinken erst bei starker Reduktion der Bitrate unter 64 kbit/s merkbar.

Rauschen und Hintergrundgeräusche: Je nach Aufnahmesituation werden in die zu erkennenden Aufnahmen Hintergrundrauschen oder Störgeräusche eingeführt. Diese verändern die Frequenzcharakteristik des Materials und führen ebenfalls dazu, dass die Erkennung erschwert wird. In welchem Maße dies der Fall ist, liegt vor allem an der relativen Lautstärke der Störungen im Vergleich zum Ausgangssignal, dem sogenannten Signal-zu-Rausch-Verhältnis. Problematisch wird das Wiedererkennen von Tonaufnahmen, in denen die Energie des Störsignals etwa gleich groß oder größer ist als die des eigentlichen Signals.

Übertragung von Lautsprecher zu Mikrofon: Ein beliebtes Einsatzszenario für das Fingerprinting sieht die Aufnahme der Anfrage über ein Handymikrofon vor. Dies bedeutet, dass eine Aufnahme über einen Lautsprecher wiedergegeben wird und dann mit dem Mikrofon aufgenommen wird. Hierbei kommt es zu Effekten, die das Ursprungssignal stark verändern. Allerdings wurde einerseits bei der Entwicklung moderner Fingerprinting-Verfahren ein besonderes Augenmerk auf diese Art von Störung gelegt, so dass hier auch recht starke Störungen toleriert werden können. Andererseits spielt dieses Szenario für den Jugendmedienschutz kaum eine Rolle.

Diese Liste macht deutlich, dass Audio-Fingerprinting-Verfahren gegenüber einer weiten Reihe von Störungen, die im täglichen Einsatz zu erwarten sind, robust sind. In der Forschung wird das Pro-

Tabelle 4.6
Audio-Fingerprinting-Lösungen und ihre Verfügbarkeit

Produkt	Lizenz	Quelle
AudibleMagic	kommerziell	www.audiblemagic.com
AudioID	kommerziell	www.mufin.com
Shazam	kommerziell	www.shazam.com
Civolution	kommerziell	www.civolution.com
AcoustID	Open Source	acoustid.org
Echoprint	Open Source	echoprint.me
pHash	Open Source	www.phash.org

blem des Audio-Fingerprintings daher weitgehend als gelöst betrachtet. Hier sind nun neue Aufgabenstellungen von Interesse, die zum Beispiel darauf abzielen, auch verschiedene Versionen des gleichen Musikstücks zu identifizieren. So geht es heutzutage beispielsweise darum, Cover- oder Live-Versionen eines Musikstücks beim Vergleich mit der Studioversion zu erkennen. In diesem Gebiet ist die Entwicklung noch recht jung. Dennoch untersuchen wir in Abschnitt 5.4.5 des folgenden Kapitels erste Anwendungsmöglichkeiten für das Jugendmedienschutz-Szenario.

In den letzten Jahren sind zunehmend kommerzielle Anwendungen des Fingerprintings entstanden. Entsprechend ist diese Technologie sowohl über kommerzielle Anbieter als auch durch Open-Source-Projekte verfügbar. Tabelle 4.6 gibt einen Überblick über eine Auswahl verfügbarer Lösungen.

Das Audio-Fingerprinting bringt alle Voraussetzungen mit, um ein zentrales Hilfsmittel im Jugendmedienschutz zu werden. Zum Einen löst es eine Aufgabe, die im Alltag immer wieder auftritt, nämlich die Identifikation bekannten jugendschutzrelevanten Materials. Mit den dabei häufig auftretenden leichten Variationen des Materials können die aktuellen Verfahren sehr gut umgehen. Zum Anderen ist die vorhandene Technologie hinreichend weit entwickelt und kommerziell beziehungsweise Open Source verfügbar und kann so von Entwicklern von Jugendschutz-Software leicht lizenziert und eingesetzt werden. Insbesondere eignet sich die Technik, um sehr große Datenmengen schnell zu überprüfen.

4.5.1.2 Spectral-Geometry-Coding-Verfahren

In diesem Abschnitt wollen wir exemplarisch an einem konkreten Verfahren zum Audio-Fingerprinting deren Vorgehensweise darstellen. Wir haben dazu das Spectral-Geometry-Coding-Verfahren [Bardeli et al., 2012] ausgewählt, da es einerseits typisch für diese Klasse von Verfahren ist und andererseits in der Evaluation (siehe Abschnitt 5.3.2) zum Einsatz kommt.

Wie im vorangegangenen Abschnitt beschrieben, bestehen Fingerprinting-Verfahren aus zwei zentralen Komponenten. Einerseits gilt es, geeignete kompakte Darstellungen für die Audiodaten zu finden, die Fingerprints. Andererseits müssen diese schnell abgeglichen werden. Dabei gilt ein wichtiges Augenmerk der Robustheit des Verfahrens gegenüber verschiedenen Störungen oder Modifikationen. Diese Robustheit kann in beiden Komponenten implementiert werden, wie wir an diesem Beispiel sehen werden.

Betrachten wir zuerst die Konstruktion der Fingerprints. Ein wichtiger Aspekt beim Entwurf von Fingerprints ist die Frage nach lokalen Fingerprints, das heißt nach solchen, die jeweils kurze Abschnitte einer Audiodatei beschreiben. Entwirft man die Fingerprints so, dass sie auch im Frequenzbereich

lokal sind, ermöglicht man schon in diesem Schritt Robustheit gegenüber vielen frequenzabhängigen Störungen, wie zum Beispiel Bandbeschränkung. Mit Rücksicht auf diese Aspekte erstellen wir Fingerprints, die so lokal wie möglich sind, das heißt sowohl anhand von Informationen aus einem möglichst kurzen Zeitabschnitt als auch aus einem möglichst engen Frequenzbereich gewonnen werden. Konkret bedeutet dies, dass wir eine Tonaufnahme durch Datenelemente beschreiben, die jeweils zu einem gegebenen Zeitpunkt und Frequenzband lokal das Audiosignal beschreiben. Die Gesamtheit dieser Datenelemente, bestehend aus dem Tupel (Zeit, Frequenzband, lokale Beschreibung), ergibt dann den Fingerprint. Sie werden aus der Zeit-Frequenz-Darstellung, das heißt dem Spektrum des Audiosignals gewonnen, indem dieses zunächst in Frequenzbänder eingeteilt wird. Dies erfolgt mit einer annähernd logarithmischen Einteilung, motiviert durch die menschliche Frequenzwahrnehmung. Die Einteilung in Frequenzbänder stellt eine erste Vergröberung der Information dar und führt somit sowohl zu einer komprimierteren Darstellung als auch zu einer höheren Robustheit. Für jeden Zeitpunkt und jedes Frequenzband wird dann eine Merkmalsklasse bestimmt. Dabei wird zunächst eine Folge von Bits, das heißt von Nullen und Einsen, anhand der Nachbarschaft im Spektrum bestimmt. In Abbildung 4.14 ist dies schematisch dargestellt. Die benachbarten Positionen werden der Reihe nach betrachtet und wann immer die Energie an der benachbarten Position im Spektrum signifikant höher ist als an der aktuellen Position wird der Bitfolge eine Eins angehängt, sonst eine Null. Dies ergibt eine Folge von 16 Bits, die zunächst einmal als Merkmalsklasse verwendet werden kann. Sie ist bereits dadurch robust, dass ein Bit sich nur dann ändert, wenn sich der Unterschied zwischen den verglichenen Zeit-Frequenz-Elementen stark ändert.

Zusätzlich wird jedoch noch eine Idee aus der robusten Signalübertragung verwendet. Hier werden bei der Übertragung die Bitfolgen zunächst auf längere Folgen abgebildet, die auf geschickte Art und Weise Redundanz einführen, so dass eine gewisse Zahl an fehlerhaft übertragenen Bits erkannt und korrigiert werden kann. Wir drehen diesen Prozess um und gehen davon aus, dass die oben beschriebene Bitfolge bereits durch einen solchen Codierungsprozess entstanden ist und dekodieren sie mit einem linearen Code. Dies führt dazu, dass auch bei Änderung einer kleinen Zahl der Bits in der Merkmalsklasse immer noch auf das gleiche Codewort dekodiert wird und führt somit einen weiteren Mechanismus ein, der die Robustheit der Fingerprints erhöht. Wir haben nun jeden Punkt mit Information versehen und entscheiden schließlich noch anhand der Merkmalsklasse, ob er indexiert wird oder nicht. Dabei schließen wir das Codewort, das vollständig aus Nullen besteht, aus, da es am ehesten an Stellen im Signal auftritt, die wenig Informationsgehalt haben.

Nun, da die Fingerprints konstruiert sind, können wir einerseits für eine große Menge von Tonaufnahmen die Fingerprints in einer Datenbank speichern. Andererseits können wir die gleichen Fingerprints für Anfragen bestimmen, um sie dann mit denen in der Datenbank abzugleichen. Zur Aufnahme in die Datenbank wählen wir zu jedem Zeitpunkt zufällig eine feste Anzahl von Positionen aus, die dann indexiert werden. Dies führt zu einer weiteren Reduktion der Datenrate. Bei diesem Verfahren werden Reduktionen auf circa 4 % der Originalgröße erreicht, das heißt, um eine Menge von einem Gigabyte unkomprimierten Audiodaten (ca. 3 Stunden Material) wiedererkennen zu können, muss ein Index von etwa 40 Megabyte vorgehalten werden.

Für die Erkennung von Audiodaten fehlt nun noch ein Verfahren, die Fingerprints einer Anfrage schnell mit der gesamten Datenbank abzugleichen. Dazu verwenden wir ein Konzept aus der Volltextsuche. In der Datenbank werden die Fingerprints nach ihrer Merkmalsklasse sortiert gespeichert. Zu einer Anfrage können dann zunächst alle Fingerprints der Datenbank, die zu einer der Merkmalsklassen in der Anfrage gehören, bestimmt werden. Diese werden dann daraufhin untersucht, ob sie in der gleichen zeitlichen Abfolge wie in der Anfrage auftreten. Dabei wird eine gewisse Fehlertoleranz eingeführt, die leichte Variationen im zeitlichen Ablauf erlaubt. Sobald hinreichend viele Merkmale der Anfrage in der richtigen zeitlichen Reihenfolge in der Datenbank aufgefunden worden sind, wird ein Treffer gemeldet und die Anfrage kann identifiziert werden.

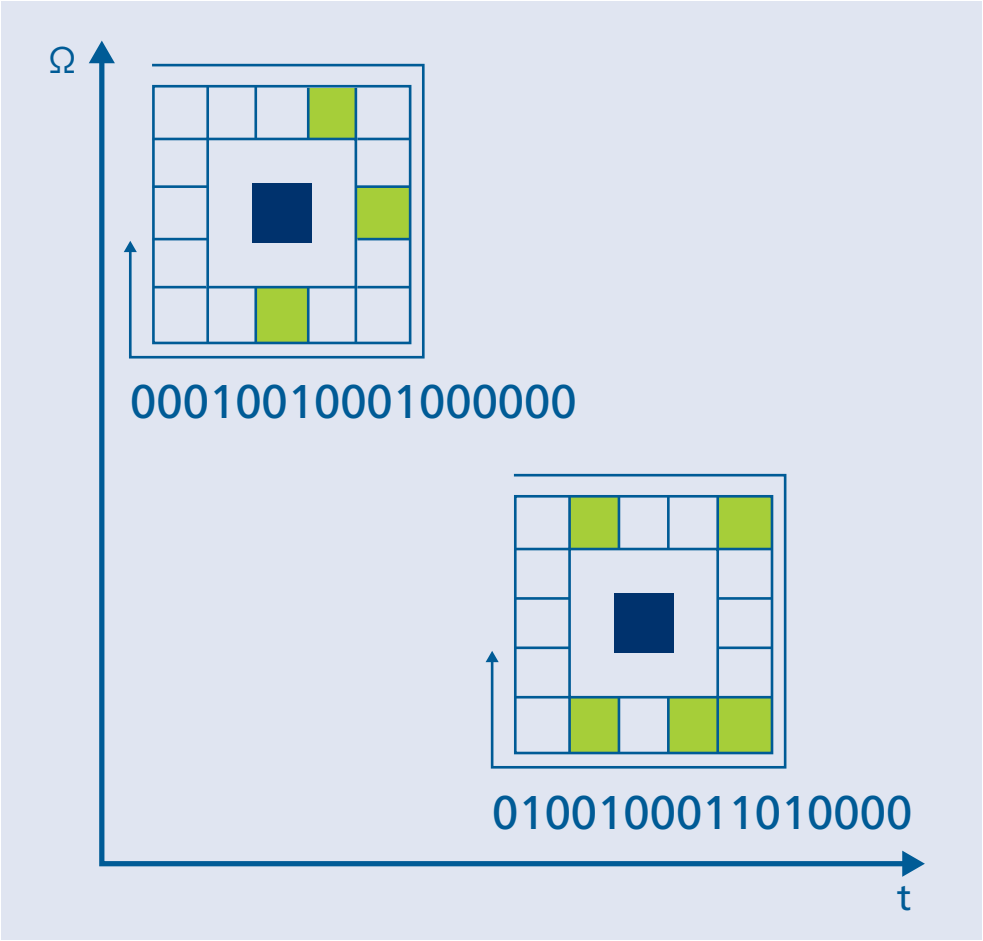


Abbildung 4.14
Fingerprint-Extraktion im Frequenzbereich. Zu jedem Zeitpunkt t und jedem Frequenzband Ω wird eine Merkmalsklasse anhand eines Vergleichs zwischen dem Spektrum an dieser Position (blau) und einer Reihe benachbarter Positionen (grün) bestimmt.

Anhand dieses Verfahrens konnten wir wichtige Aspekte von Audio-Fingerprinting-Verfahren darstellen. Zunächst sollten Fingerprints lokale Informationen über eine Tonaufnahme liefern. Zusammen mit weiteren Maßnahmen muss schon beim Entwurf die Robustheit der Fingerprints gewährleistet werden. Schnelle Matching-Verfahren sind notwendig, um mit großen Datenmengen umgehen zu können. Diese werden bereits durch den Entwurf der Fingerprints ermöglicht, indem diese einerseits bereits eine möglichst starke Komprimierung der Information vornehmen und andererseits die Daten so strukturieren, dass sie schnell anhand einer Anfrage ausgesiebt werden können. Schließlich kann auch in den Matching-Verfahren die Robustheit weiter erhöht werden.

Strukturell gelten diese Beobachtungen nicht nur für das Audio-Fingerprinting, sondern auch für das Video-Fingerprinting, das wir im nächsten Abschnitt betrachten wollen und bei dem dann Bildinhalte die Rolle der Toninhalte übernehmen und diese ergänzen.

4.5.1.3 Video-Fingerprinting

Während sich die vorangegangenen Abschnitte mit dem Wiedererkennen von Toninhalten auseinandergesetzt hat, wollen wir uns in diesem Abschnitt den Bildinhalten zuwenden. Dabei betrachten wir sowohl statische Bilder als auch Videos. Wir wollen den Stand der Technik zunächst an einem Beispiel illustrieren. Wer regelmäßig Googles Bildsuchfunktion verwendet, wird festgestellt haben, dass neben dem Eingabefeld für die textuelle Suche nun auch ein kleines Icon in Form eines Fotoapparats hinzugekommen ist. Mit dieser Option kann man ein Beispielbild auswählen und erhält als Suchergebnisse ähnliche Bilder und zusätzliche Informationen, die mit diesen Bildern in Zusammenhang stehen. So ist es zum Beispiel möglich, ein bekanntes Gebäude auf einem Foto zu identifizieren, indem ein ähnliches Bild auf der Wikipedia-Seite zu diesem Gebäude gefunden wird und die Informationen dieser Webseite mit dem entsprechenden Foto verknüpft werden. Entsprechend lässt sich auch im Jugendschutz vorgehen. Wird eine Datenbank mit Bildinhalten, die als jugendschutzrelevant bekannt sind, vorgehalten, so kann unbekanntes Material schnell mit diesen Bildern abgeglichen und, falls es bereits in der Datenbank vorhanden ist, wiedererkannt werden. Solche Techniken finden sich in sehr ähnlichen Szenarien bereits im Einsatz. Microsoft bietet mit PhotoDNA[18] ein Bild-Fingerprinting-Verfahren, dass helfen soll, Kinderpornografie im Internet aufzuspüren und zu beseitigen. Dieses wurde im Jahr 2012 auch vom Bundeskriminalamt zu Testzwecken beschafft.[19] Ähnliches gilt auch für Videomaterial. Hier lassen sich alle Informationsebenen eines Videos nutzen, um bereits untersuchte Videos wiederzuerkennen. Dabei können sowohl Verfahren zum Audio- und Bild-Fingerprinting verwendet werden als auch Methoden, die speziell die Bewegtinhalte des Videos berücksichtigen, zum Beispiel Schnittfolgen.

Typische Video-Fingerprinting-Verfahren bestehen aus vier zentralen Elementen. Als Erstes werden Keyframes extrahiert, also eine Folge von Einzelbildern, in der jede Einstellung im Video durch ein Bild repräsentiert wird. Darauf folgt die Merkmalsextraktion, bei der aus jedem Keyframe Eigenschaften wie Farb- und Formverteilung bestimmt werden. Aus diesen Merkmalen werden dann die Fingerprints erstellt und in einer Datenbank indexiert. Schließlich erfolgt dann der Abgleich von Fingerprints eines Anfragevideos mit denjenigen in der Datenbank.

Typische Merkmale sind hierbei zunächst blockbasierte Merkmale, bei denen das Bild in eine feste Anzahl von Teilen zerlegt wird, zu denen dann Merkmale wie etwa die Farbverteilung bestimmt werden. Weiterhin wurden zur Charakterisierung der Formen im Bild Histogramme von Gradientenrichtungen verwendet. Hierbei werden zunächst die Kanten im Bild detektiert und ihre Ausrichtung bestimmt. Die möglichen Richtungen werden dann in eine feste Menge an Richtungsbereichen (In-

[18] siehe http://www.microsoft.com/en-us/news/presskits/photodna/, Stand 5.11.2012

[19] BT-Drucksache 17/11299: Antwort zu einer Kleinen Anfrage zum Thema Beteiligung deutscher Polizeien am EU-Projekt »iCOP« zur Entwicklung von Mustererkennung und Filtertechnologie für Tauschbörsen (BT-Drucksache 17/11068).

tervalle von Winkeln) zusammengefasst und die Anzahl der Kanten, die in die jeweiligen Bereiche fallen, gezählt. Schon frühzeitig kamen zudem die in der Bilderkennung sehr beliebten SIFT-Merkmale zum Einsatz [Lowe, 2004]. Diese bestimmen aus dem Bild eine Reihe von prominenten Punkten (zum Beispiel Ecken), deren Extraktion unter typischen Bildtransformationen, zum Beispiel Drehungen und Änderungen des Kamerawinkels, aber vor allem unter Skalierungen robust bleibt. Diese Punkte werden dann mit lokalen Bildinformationen angereichert.

In den letzten Jahren wurden die bestehenden Verfahren zunehmend optimiert, um eine höhere Robustheit gegenüber verschiedenen Störungen und Modifikationen des Videomaterials (s. u.) zu erreichen. Neben einem besonderen Augenmerk auf die Kombination von Audio- und Bildmerkmalen wurden dabei mehr und neue Bildmerkmale verwendet. Unter den Bildmerkmalen sind hier vor allem die sogenannten *Bags of Visual Words* zu nennen. Bei diesem Ansatz wird versucht, Bilder ähnlich zu beschreiben, wie dies bei Text-Dokumenten im sogenannten Information Retrieval geschieht [Sivic und Zisserman, 2009]. Wird in letzterem Szenario Text durch die in ihnen vorkommenden Wörter charakterisiert, werden beim Bags of Visual Words Ansatz zunächst aus einer großen Menge kleiner Bildausschnitte Häufungspunkte ausgewählt, die dann als Äquivalente zu Wörtern agieren. Es werden dabei kleine Bildausschnitte aus dem zu beschreibenden Bild betrachtet und dem ähnlichsten *Wort* zugeordnet. So lassen sich Bilder als Listen von in ihnen vorhandenen visuellen Wörtern auffassen. Schließlich werden spezielle Verfahren für den Umgang mit einzelnen Videotransformationen entwickelt, etwa um Bild-in-Bild-Inhalte zu erkennen.

Vom National Institute of Standards and Technology (NIST) werden im Rahmen der TRECVID-Kampagne [Over et al., 2011] regelmäßig Evaluationen im Bereich Video-Retrieval durchgeführt. Das Video-Fingerprinting wird dabei seit 2008 vor allem mit Hilfe der Evaluationsaufgabe *Content-based copy detection* untersucht. Dabei sollen Kopien von Videos in einer großen Datenbank wiedergefunden werden. Die Robustheit der Verfahren wird dabei überprüft, indem getestet wird, ob Kopien trotz folgender Transformationen gefunden werden:

Simuliertes Abfilmen: Hier werden die Videos künstlich so modifiziert, dass Effekte auftreten, wie sie zum Beispiel beim Abfilmen einer Kinovorführung mit einer Handkamera entstehen.

Bild-in-Bild: Das Originalvideo wird in einem kleineren Bild vor einem anderen Hintergrundvideo wiedergegeben.

Einfügen eines Musters: Hier wird ein Muster über einen gewissen Zeitraum in das Video eingebettet. Diese Muster entsprechen etwa Bauchbinden oder Informationseinblendungen, die zum Beispiel aus einer Kombination von Text und Logos bestehen.

Kompression: Ähnlich wie bei Audio-Fingerprinting-Verfahren werden die Videos einer verlustbehafteten Kompression mit niedrigen Bitraten unterzogen.

Helligkeitsänderung: Der Helligkeitsverlauf in den Bildern des Videos wird mit Hilfe von Gammakorrektur modifiziert. Dieses Verfahren ist dazu gedacht, die in digitalen Bildern eher lineare Wiedergabe von Intensitäten an die eher logarithmische Wahrnehmung von Intensitäten beim menschlichen Betrachter anzupassen.

Reduktion der Qualität: Hier werden zufällig drei der folgenden Transformationen auf das Video angewendet: Unschärfe einführen, Helligkeit ändern, Einzelbilder auslassen, Kontraständerung, Kompression, Änderung des Seitenverhältnisses, Bildrauschen.

Nachbearbeitung mit Studiotechnik: Wiederum werden drei zufällige Transformationen angewandt, die diesmal allerdings aus folgender Liste von Nachbearbeitungstechniken stammt, wie sie in der Filmproduktion üblich sind: Bilder zuschneiden, Bilder verschieben, Kontrast optimieren, Untertitel einfügen, Bild spiegeln, Muster einfügen, Bild-in-Bild mit dem Orgininalvideo.

Tabelle 4.7
Auswahl von Video-Fingerprinting-Lösungen und ihre Verfügbarkeit

Produkt	Lizenz	Quelle
iPharro	kommerziell	http://www.ipharro.com
Civolution	kommerziell	http://www.civolution.com
pHash	Open Source	http://www.phash.org

Zusätzlich werden im Rahmen der TRECVID-Kampagne auch Kombinationen von Video- und Audio-Fingerprinting evaluiert und Transformationen der Tonspur betrachtet, wie sie in Abschnitt 4.5.1.1 beschrieben sind.

Anhand der Evaluationsergebnisse aus dem Jahr 2011 lässt sich ein guter Eindruck über den aktuellen Stand des Video-Fingerprintings mit Hinsicht auf die Robustheit der Verfahren gewinnen. Abbildung 4.15 gibt einen Überblick über die Güte der evaluierten Verfahren von 22 teilnehmenden Forschungsgruppen. Dabei wird der F-Wert (siehe Abschnitt 4.2.1) der Verfahren einzeln für jede von 8 Video- und 7 Audio-Transformationen, die durch eine zufällige Kombination der oben genannten Transformationen erzeugt werden, dargestellt. Anhand der möglichen Kombinationen und der Berücksichtigung unterschiedlicher Szenarien wurden dann insgesamt 73 Aufgaben zusammengestellt. Zu jeder Aufgabe wurden circa 200 Anfragen erstellt, die dann in einer Datenbank von über 12 000 Videos mit einer Gesamtlaufzeit von circa 4 000 Stunden wiedergefunden werden mussten. Die in der Abbildung dargestellten Ergebnisse beziehen sich auf ein Szenario, in dem die Systeme so eingestellt wurden, dass sie keine Fehlalarme liefern, das heißt, jede gemeldete Fundstelle muss tatsächlich dem gesuchten Videoausschnitt entsprechen. Die Einflüsse der meisten Transformationen, vor allem im Bildbereich, können sehr gut behandelt werden. Am schwierigsten stellen sich in dieser Evaluation Transformationen im Audiobereich dar, bei denen zusätzliche Sprache in den Originalton eingemischt wird. F-Werte im Bereich von 80–90 % bedeuten in diesem Zusammenhang, dass immer noch ein erheblicher Anteil der modifizierten Anfragen identifiziert werden kann, selbst wenn die Verfahren so eingestellt werden, dass sie ausschließlich richtige Treffer liefern. Es lässt sich somit ablesen, dass heutige Video-Fingerprinting-Verfahren bereits sehr robust einen hohen Anteil an Kopien von Videos wiedererkennen können, selbst wenn diese recht stark modifiziert werden. Dies ist eine sehr gute Ausgangssituation für den Einsatz im Jugendmedienschutz.

Videomaterial hat im Vergleich zu den anderen hier betrachteten Datenformaten die höchsten Datenraten. Daher sind Verfahren, die mit dem Inhalt von Videos arbeiten, sehr rechenaufwändig. Schon das Dekodieren der komprimiert gespeicherten Informationen in Bildfolgen benötigt dabei nicht unerheblichen Aufwand. In letzter Zeit wird deshalb verstärkt an Verfahren gearbeitet, die die zu analysierenden Videos nicht dekodieren müssen, um Aussagen über ihren Inhalt machen zu können. Dabei wird ausgenutzt, dass Informationen über Bewegungen von Bildelementen von einem Bild zum nächsten zur Kompression verwendet werden. Diese Informationen lassen sich direkt aus dem kodierten Videostrom auslesen und können dann nach geeigneter Vorverarbeitung Informationen über die Bewegungsinhalte im Video unabhängig vom Kodierverfahren liefern.

Das bildbasierte Video-Fingerprinting ist etwas jünger als das Audio-Fingerprinting. Entsprechend ist die industrielle Verfügbarkeit hier etwas weniger ausgeprägt. Dennoch bieten viele Anbieter von Audio-Fingerprinting-Verfahren auch bildbasierte Systeme an. Aufgrund der höheren Anforderungen bei der Implementierung von Video-Fingerprinting-Verfahren ist die Auswahl an Open-Source-Lösungen hier geringer. Tabelle 4.7 gibt einen Überblick über eine Auswahl verfügbarer Lösungen.

Für das Video-Fingerprinting gilt in Hinblick auf die Anwendbarkeit im Jugendschutz-Szenario Ähnliches wie für das Audio-Fingerprinting. Allerdings sind die Datenmengen im Videobereich erheblich

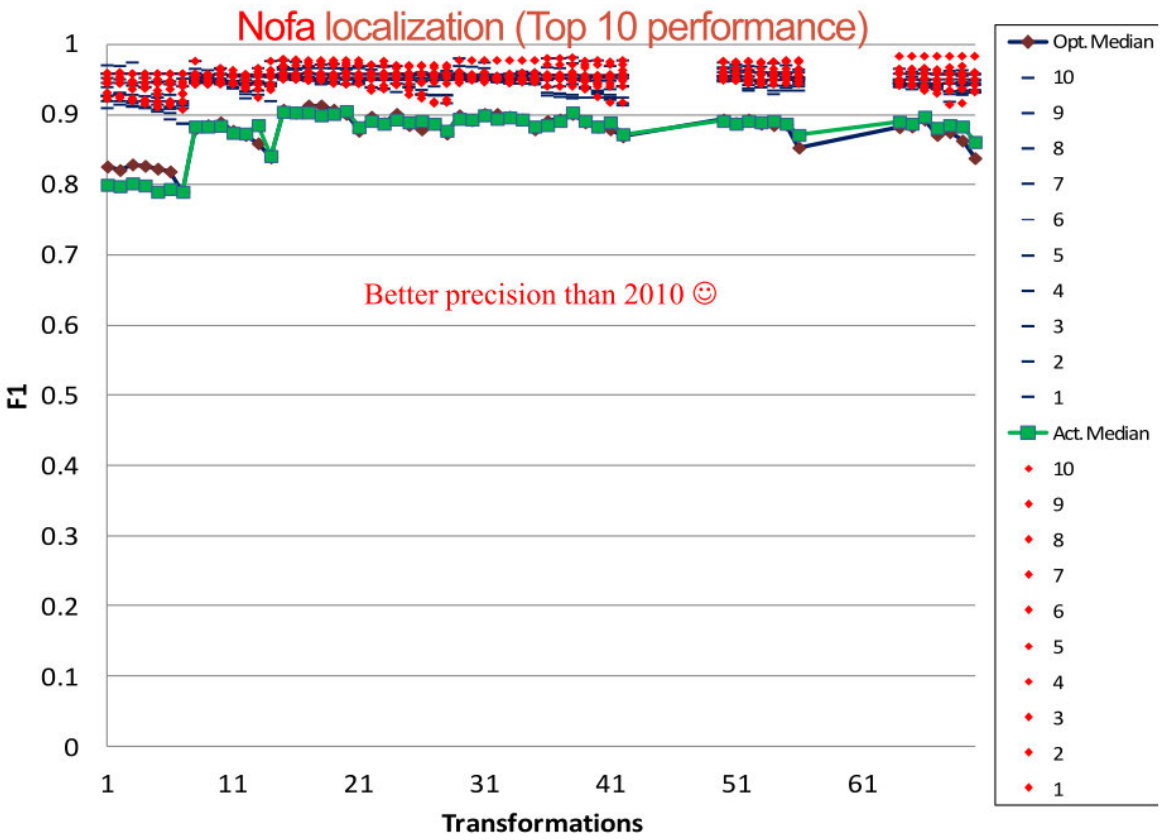


Abbildung 4.15
TRECVID-Evaluationsergebnisse aus dem Jahr 2011 zur Detektion von Videokopien. Zu verschiedenen Transformationen einer Menge von Anfragevideos wird der F-Wert anhand der Suche in einer großen Menge von Videos angegeben. Das dargestellte Evaluationsszenario fordert, dass keine Fehlalarme vom System erzeugt werden (*NO False Alarms* (NOFA)).
Bildquelle: [Kraaij und Awad, 2011]

größer und die möglichen Variationen des Materials vielfältiger. Deshalb sind besonders Verfahren, die ohne ein Dekodieren des Videos auskommen, interessant.

4.5.2 Automatische Spracherkennung

Die automatische Spracherkennung hat zum Ziel, Tonaufnahmen gesprochener Sprache automatisch in geschriebenen Text umzusetzen, der dann von Computern weiter verarbeitet werden kann. Entwicklungen an dieser Aufgabe finden seit über vierzig Jahren statt und die Forschung auf diesem Gebiet hat vergleichsweise starke Förderung erhalten. Entsprechend hat die Technik sich hier von der Erkennung weniger Einzelwörter hin zur Erkennung frei gesprochener Texte stark weiter entwickelt. Die Qualität der Ergebnisse hängt dabei heutzutage vor allem von zwei Faktoren ab: Zunächst ist entscheidend, ob ein Spracherkennungssystem speziell auf einen bestimmten Sprecher trainiert ist. Danach ist vor allem die akustische Qualität wichtig, die durch Hall, Störgeräusche und ähnliche Einflüsse auf das Sprachsignal beeinträchtigt werden kann. Unter optimalen Bedingungen, das heißt wenn das System auf den Sprecher trainiert ist und über ein Nahsprechmikrofon nahezu ausschließlich die Sprache des Sprechers aufgenommen wird, werden hier durchaus stabil Worterkennungsraten von 99 % erreicht. Dabei ist die Worterkennungsrate eines der zentralen Maße der Güte eines Spracherkennungssystems. Sie gibt an, wie viel Prozent der Wörter korrekt erkannt worden sind. Häufig wird auch alternativ die Wortfehlerrate angegeben, die beschreibt wie viel Prozent der Wörter nicht korrekt erkannt worden sind.

Vor allem durch die starke Verbreitung mobiler Endgeräte und intelligente Applikationen wie Siri, die auf automatische Spracherkennung aufbauen, ist auch die Akzeptanz beim Endnutzer in letzter Zeit gestiegen. Mit Blick auf den Jugendschutz sind jedoch andere Anwendungsszenarien von Interesse. Für den Jugendschutz relevante Anwendungsszenarien benötigen dabei vergleichsweise geringe Erkennungsraten, um sinnvoll einsetzbar zu sein. Ein relevantes Szenario ist die Sprachsuche. Dabei geht es darum, große Archive von Tonaufnahmen inhaltlich durchsuchbar zu machen. Während die schnelle Suche in riesigen Mengen von Texten heutzutage unproblematisch ist, liegen die Informationen in Tonaufnahmen zunächst für den Computer unzugänglich vor. Wenn nicht zusätzliche textuelle Metadaten die Dokumente eines Tonarchivs näher beschreiben, ist die Suche in solchen Archiven zunächst nur durch tatsächliches Abhören der Dokumente möglich. Mit Hilfe der automatischen Spracherkennung kann dieser Aufwand erheblich reduziert werden. Selbst wenn die Worterkennungsraten nur 60–80 % betragen, können Tonaufnahmen bereits komfortabel durchsucht werden oder durch geeignete Verfahren Zusammenfassungen erstellt werden. Solche Verfahren befinden sich bereits im stabilen kommerziellen Einsatz. Ein Beispiel hierzu stellt die ARD-Mediathek dar.[20] In diesem Online-Archiv des öffentlich-rechtlichen Fernsehsenders kann bereits in einem Volltext, der durch automatische Spracherkennung erzeugt wurde, innerhalb ausgewählter Formate, wie zum Beispiel der Tagesschau, gesucht werden.

Dieses Szenario ist der Problematik des technischen Jugendmedienschutzes schon recht nahe. Wenn durch automatische Spracherkennung der sprachliche Inhalt von jugendschutzrelevantem Material für die automatische Analyse erschlossen werden kann, können Verfahren, wie sie in Abschnitt 4.4 beschrieben wurden, zum Einsatz kommen. Entscheidend ist dabei, ob die akustische Qualität des zu erkennenden Materials hinreichend gut ist. Dies dürfte zum Beispiel bei Reportagen mit politisch extremistischem Inhalt der Fall sein, während zum Beispiel Material, das stark mit Musik hinterlegt ist, schon schwieriger zu bearbeiten sein dürfte. Eine exemplarische Untersuchung zu den Einsatzmöglichkeiten wird in Abschnitt 5.3.4 gegeben.

[20] siehe http://www.ardmediathek.de/, Stand November 2012

Im Folgenden geben wir zunächst einen Überblick über den Stand der Technik im Bereich der automatischen Spracherkennung und betrachten dann noch einmal genauer die besonderen Herausforderungen, die der Jugendschutz an diese Technologie stellt.

4.5.2.1 Stand der Technik

Erste Systeme zur Spracherkennung wurden bereits ab den 1950er Jahren entwickelt und konzentrierten sich zunächst darauf, einfache Spracheinheiten isoliert zu erkennen. Solche Einheiten sind zum Beispiel Vokale, Ziffern und Phoneme. Vokale lassen sich anhand der relativen Verteilung der Signalenergie über verschiedene Frequenzbereiche auseinanderhalten. Diese Eigenschaft wurde in frühen Systemen ausgenutzt und zum Beispiel auch für die Erkennung von Ziffern eingesetzt. Die Ziffernerkennung stellt eine der einfachsten Aufgaben der Spracherkennung dar, die sich bereits seit längerer Zeit in kommerziellen Anwendungen einsetzen lässt. Phoneme sind die atomaren Einheiten der Sprache, das heißt die kürzesten Lauteinheiten, die sich noch unterscheiden lassen. Daher ist die Erkennung von Phonemen ein wichtiger Schritt für die Entwicklung von Spracherkennungssystemen. Auf diesem Gebiet wurde zunächst die Forschung intensiviert und mit Verfahren wie Template Matching zum direkten Vergleich von Zeit-Frequenz-Mustern (Templates) versucht, die Erkennung aller für eine Sprache benötigten Phoneme zu ermöglichen. Dabei wurden bereits Verfahren wie das Dynamic Time Warping entwickelt, um der zeitlichen Variabilität in der Aussprache Herr zu werden.

Schließlich haben stochastische Verfahren in der Informatik, insbesondere in der Mustererkennung, zu erheblichen Fortschritten geführt. Während traditionelle regelbasierte Verfahren mit typischen Phänomenen realer Daten, wie zum Beispiel der großen Variabilität natürlicher Sprache, nicht umgehen konnten, lassen sich solche Probleme durch den Einsatz stochastischer Verfahren umgehen. Dabei werden die statistischen Eigenschaften von zu erkennenden Phänomenen modelliert und Daten anhand derjenigen Hypothese klassifiziert, unter der sie am wahrscheinlichsten sind. Stochastische Verfahren werden heute insbesondere auch in der Spracherkennung eingesetzt. Seit Einführung des nun dominierenden statistischen Paradigmas der Hidden Markov Modelle (HMM) sind scheinbar unlösbare Probleme überwunden und die Basis dafür geschaffen worden, dass zur Verfügung stehende große Mengen von Beispieldaten und hinreichend schnelle Computer zur Verarbeitung dieser Daten konstruktiv genutzt werden können.

Da die Spracherkennung mit Hidden Markov Modellen praktisch allen modernen Systemen zugrunde liegt, wollen wir nun einen kurzen Überblick über diesen Ansatz geben, bevor wir die eigentlichen Modelle beschreiben. Dabei wird Sprache zunächst auf drei Ebenen modelliert: Auf der ersten Ebene wird der Klang der Phoneme modelliert. Dazu wird ein HMM für jedes Phonem der zu erkennenden Sprache trainiert. Auf der zweiten Ebene müssen Phoneme zu Wörtern zusammengesetzt werden. Dies geschieht für gewöhnlich mit Hilfe von Aussprachewörterbüchern, das heißt von Listen, die zu jedem Wort eine oder mehrere Aussprachevarianten in Form von Phonemfolgen angeben. Schließlich wird auf der dritten Ebene Sprache statistisch modelliert, das heißt es wird ermittelt, welche Wortfolgen in einer gegeben Sprache wie wahrscheinlich sind (siehe dazu auch Abschnitt 4.4.3.1). Diese Informationen werden im Prinzip zu einem großen hierarchischen HMM kombiniert.

Hidden Markov Modelle beschreiben jeweils für kurze Zeiträume die statistische Verteilung geeigneter Merkmale eines Audiosignals (zum Beispiel die im folgenden Absatz beschriebenen Mel Frequency Cepstral Coefficients). Dazu repräsentieren verschiedene Zustände des Modells jeweils mehr oder weniger stationäre Phasen. Dies sind typischerweise Anfang, Mitte und Ende eines Phonems, beziehungsweise Übergänge zwischen Phonemen. Der Übergang zwischen diesen Zuständen erfolgt mit Hilfe einer Wahrscheinlichkeitsverteilung, die zu jedem Zustand die Wahrscheinlichkeit des Übergangs in einen der anderen Zustände im nächsten Zeitschritt angibt. Zu einem Audiosignal kann dann diejenige Zustandsfolge berechnet werden, die die Wahrscheinlichkeit maximiert, dass die Merkmale

des Audiosignals gemäß der Wahrscheinlichkeitsverteilungen dieser Zustandsfolge verteilt sind. Die Zustandsfolge enthält dann die Information, welche Phonemfolge am wahrscheinlichsten vorliegt. Mit Hilfe des hierarchischen Modells lässt sich daraus auch die wahrscheinlichste Wortfolge ableiten.

Bei der Spracherkennung wird zunächst nicht mit dem rohen Audiosignal gearbeitet. Vielmehr werden geeignete Merkmale im zeitlichen Verlauf aus dem Signal extrahiert, die möglichst gut diejenige Information im Signal wiedergeben, die auf den gesprochenen Text zurück schließen lässt. Unwesentliche Information, wie zum Beispiel die Stimmlage des Sprechers oder die spezifische Intonation, versucht man dabei herauszufiltern. Für den breiten Gebrauch stellen hierbei die sogenannten Mel Frequency Cepstral Coefficients (MFCCs) die besten Merkmale dar. Der Entwurf besserer Merkmale ist ein aktives Forschungsgebiet. Allerdings zeichnet sich bisher keine Alternative zu den MFCCs ab, die diesen in einer breiten Palette von Anwendungen überlegen wäre.

Zuletzt hat die Kombination von Neuronalen Netzen (vgl. Abschnitt 4.4.2.3) und Hidden Markov Modellen zu einer klaren Verbesserung geführt. Der Ansatz, Neuronale Netze als Schritt in der Merkmalsextraktion zu verwenden, stammt bereits aus den 1990er Jahren. Allerdings fehlte es zunächst an effizienten Lernverfahren für sogenannte Tiefe Neuronale Netze. Erst in den letzten Jahren konnte mit dem Einzug neuer Verfahren auf diesem Gebiet die Kombination der beiden Verfahren überzeugende Ergebnisse liefern.

4.5.2.2 Herausforderungen bei zu erwartendem Material aus dem Jugendmedienschutz

In der Anwendung von Spracherkennungsverfahren zählt vor allem die Worterkennungsrate auf Material aus der entsprechenden Anwendungsdomäne. Diese hängt auch heute noch stark von der Komplexität des zu erkennenden Tonmaterials ab. Seit 1987 werden vom NIST regelmäßig Evaluationswettbewerbe durchgeführt, die den Stand der Technik dokumentieren. Abbildung 4.16 gibt einen guten Überblick über die Historie dieser Evaluationen und damit auch über die Entwicklung der Spracherkennungstechnologie. Zum Vergleich ist zunächst die menschliche Erkennungsleistung angegeben. Ein von Hand erstelltes Transkript einer Tonaufnahme enthält im Mittel 2–4 % falsch erkannte Wörter. Die Interpretation der Wortfehlerrate hängt von der gewünschten Anwendung ab. Während für Diktieranwendungen erst Fehlerraten unter 5 % zu einer sinnvollen Anwendung führen, können in anderen Anwendungen erheblich höhere Fehlerraten toleriert werden. Sollen zum Beispiel nur Schlüsselwörter erkannt werden, Textinhalte zusammengefasst werden oder Toninhalte durchsucht werden, können schon ab Fehlerraten unter 50 % nützliche Anwendungen realisiert werden. Dies ist insbesondere deshalb der Fall, da längere Wörter zuverlässiger erkannt werden als kürzere Wörter und gleichzeitig mit höherer Wahrscheinlichkeit bedeutungstragend sind. Fehlerraten unter 20 % auf Texten mit großem Vokabular sind für die zuletzt genannten Anwendungen bereits eine sehr stabile Basis.

Für unterschiedliche Szenarien zeigen sich hier ähnliche Trends. Zunächst lässt sich gelesene Sprache ohne Störgeräusche je nach Vokabulargröße schon seit einigen Jahren auch automatisch mit Fehlerraten im Bereich der Fähigkeiten menschlicher Hörer erkennen. Schwierig bleiben in diesem Szenario große Vokabulargrößen ab 20 000 Wörtern und Störeinflüsse wie Hintergrundgeräusche oder wechselnde Mikrofontypen, die zu Fehlerraten führen, die eher im Bereich von 20 % liegen.

Eine Anwendung, die größere Herausforderungen an Spracherkenner stellt, sind Dialog-Systeme. In diesem Szenario sollen Benutzer mit einem Computersystem interagieren, um zum Beispiel Informationen zu erfragen (grüne Kurve). Dabei ist die Behandlung spontaner, nicht geplanter Sprache notwendig. In Szenarien mit einem stark begrenzten Umfang möglicher Interaktionen, zum Beispiel bei Informationssystemen zur Flugbuchung, lässt sich die Fehlerrate besonders nahe an die menschliche Leistung bringen.

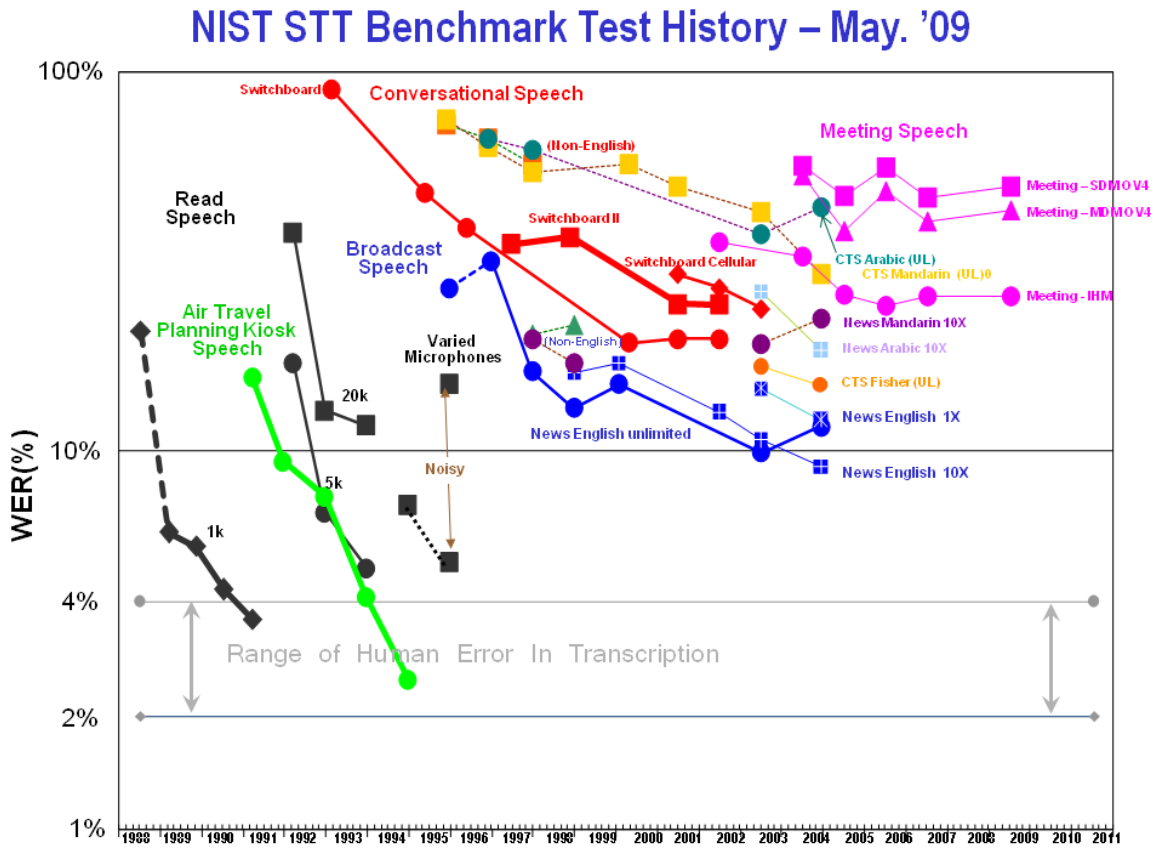


Abbildung 4.16
Zusammenfassung der NIST-Evaluationsergebnisse zur automatischen Spracherkennung seit 1987. Der Fortschritt lässt sich in der Reduktion der Wortfehlerrate (vertikale Achse) über die Zeit (horizontale Achse) gut ablesen. Es sind Ergebnisse auf verschiedenen Testkorpora dargestellt, die über die Zeit von zunehmender Komplexität und Schwierigkeit sind.
Bildquelle: [NIST, 2009]

Das Datenmaterial im Jugendschutzszenario lässt sich, wie einleitend beschrieben, am ehesten mit Material aus Rundfunk und Fernsehen vergleichen (blaue Kurven), wenn man bedenkt, dass hier auch Spielfilme, Diskussionssendungen und Ähnliches einzuordnen sind. Eine typische Anwendung besteht in diesem Kontext darin, Nachrichtensendungen durch automatische Spracherkennung in Text zu übersetzen und so durchsuchbar zu machen. Eine besondere Herausforderung stellt dabei das große, sich ständig ändernde Vokabular dar. Dennoch lassen sich hier inzwischen Wortfehlerraten von knapp 20 bis hinab zu 10 Prozent erreichen.

Eine große Herausforderung stellt weiterhin die automatische Erkennung von Sprache in natürlichen Dialogen zwischen Menschen (rote Kurven) dar. Die Variabilität dieser Gespräche ist sehr stark und das Problem noch lange nicht als gelöst zu betrachten.

Noch schwieriger ist die Spracherkennung in Meeting-Situationen (rosa Kurven), in denen nicht jeder Teilnehmer über ein eigenes Mikrofon verfügt und in denen die Sprache sehr spontan und häufig unvollständig ist.

Durch die politischen Entwicklungen der letzten Jahre sind schließlich eine Reihe von Sprachen in den Mittelpunkt des Interesses getreten, die bisher weniger stark durch die Spracherkennung abgedeckt wurden, und die zum Teil besondere Herausforderungen mit sich bringen (zum Beispiel tonale Sprachen wie das Chinesische).

4.5.2.3 Marktsituation

Auf dem Markt sind seit Jahren zahlreiche kommerzielle Produkte zur automatischen Spracherkennung erhältlich. Allerdings hat sich der Markt in den letzten Jahren stark konsolidiert und die Anzahl der Anbieter ist gesunken. Eine Übersicht über kommerzielle Systeme wird in Tabelle 4.8 gegeben. Insbesondere existieren auch sehr spezialisierte Produkte für bestimmte Anwendergruppen mit besonderen Anforderungen, die zum Beispiel häufig viel Text mit fachspezifischem Vokabular diktieren. Hierzu zählen unter anderem Ärzte, Anwälte oder Menschen mit Behinderung.

Es sind zudem eine Reihe von Open-Source-Lösungen für Spracherkennungssysteme verfügbar. Eine Übersicht dazu findet sich ebenfalls in Tabelle 4.8. Obwohl damit die grundsätzliche Technologie zur Verfügung steht, benötigt die Erstellung eines einsatzfähigen Systems noch erheblichen Aufwand. Dieser besteht insbesondere darin, geeignete Modelle für die gewünschte Einsatzdomäne zu trainieren. Dazu sind unter anderem große Mengen transkribierter Tonaufnahmen notwendig, die einerseits den Lautumfang der zu erkennenden Sprache und andererseits – und hier liegt die besondere Herausforderung – die Variabilität der Sprachdomäne gut abdecken müssen. Konkurrenzfähige Systeme setzen typischerweise einen Datenumfang von wenigstens 100 bis 200 Stunden voraus. Die Transkripte dieser Tonaufnahmen müssen zudem hoch genau sein und es ist auf eine Reihe von Normalisierungen zu achten, was zum Beispiel die Schreibweise von Zahlen angeht. Es ist gerade dieser Schritt, der die Erstellung eines Spracherkennungssystems sehr teuer macht. Für das Sprachmodell sind zudem umfangreiche Textdaten in der Zielsprache und möglichst aus der Zieldomäne von Nöten. Diese sind aufgrund der großen Mengen an Text, die über das Internet verfügbar sind, heutzutage allerdings verhältnismäßig leicht zugänglich.

Ein aktueller Trend, der den Einsatz von Spracherkennung für Drittanbieter erheblich vereinfacht, besteht in Hosted-Speech-Applications. Dabei handelt es sich um Server-Lösungen, die von einigen Herstellern von Spracherkennungssystemen angeboten werden. Über ein Webinterface können Drittanbieter so Spracherkennungslösungen auf Basis sehr robuster kommerzieller Systeme entwickeln, ohne sich um deren Einrichtung oder Wartung kümmern zu müssen. Außerdem ist es so möglich, automatisch von Verbesserungen eines solchen Systems profitieren zu können.

Tabelle 4.8
Spracherkennungslösungen und ihre Verfügbarkeit.

Produkt	Lizenz	Quelle
Dragon	kommerziell	http://www.nuance.com
Microsoft Tellme	kommerziell	http://www.microsoft.com/en-us/tellme
Websphere Voice	kommerziell	http://www-01.ibm.com/software/voice
Sphinx	Open Source	http://cmusphinx.sourceforge.net
Julius	Open Source	http://julius.sourceforge.jp
HTK	offen für Forschungszwecke	http://htk.eng.cam.ac.uk

Es gibt im Kontext der Studie eine Reihe von Anwendungsbereichen, in denen Spracherkennung sinnvoll erscheint, da die Erkennungsleistung moderner Systeme in einigen der im Jugendschutz relevanten Domänen durchaus hoch genug ist, um anschließend mit Methoden aus der Textanalyse Erkenntnisse über den Inhalt des Materials zu gewinnen. Diese werden wir in Kapitel 5 an ausgewählten Szenarien noch eingehender untersuchen.

Für Entwickler von Jugendschutz-Lösungen, die selbst keine Expertise im Bereich Spracherkennung haben, bietet es sich an, auf eine gehostete Spracherkennungslösung zurückzugreifen. Auch wenn es Open-Source-Lösungen gibt, die durchaus mit kommerziellen Produkten mithalten können, ist es doch die Entwicklung und Wartung der notwendigen Modelle, die die Güte eines Spracherkennungssystems ausmachen und erhebliche Anforderungen an Expertise und Ressourcen erfordern.

4.5.3 Optische Texterkennung

Die Aufgabe der optischen Texterkennung ist vergleichbar mit der automatischen Spracherkennung. Sie besteht in diesem Fall wiederum darin, geschriebenen Text zu erzeugen, der von Computern verarbeitet werden kann. Allerdings besteht die Eingabe aus Bildern, die Text enthalten. Ein typisches Beispiel stellen hier gescannte Buch- oder Zeitungsseiten dar, die zwar für den menschlichen Benutzer lesbar sind, deren Inhalt aber, da der Text nur implizit vorliegt, zunächst nicht ohne Weiteres von Computern durchsucht oder weiterverarbeitet werden kann. Im Falle von Zeitungsarchiven liegt somit die Situation vor, dass große Archive zurückliegender Ausgaben auch dann, wenn sie bereits digitalisiert sind, nur auf dem klassischen Wege durchsucht werden können, das heißt durch tatsächliches Lesen der einzelnen Überschriften und Artikel. Wünschenswert wären aber Volltextsuchfunktionen, wie sie für digitale Texte Standard sind. Die optische Texterkennung macht dies zusammen mit weiteren Verfahren möglich. Gescannte Seiten werden zunächst geeignet vorverarbeitet, das heißt die Seiten werden entzerrt und aufgerichtet, um Einflüsse aus dem Scanprozess auszugleichen. Zusätzlich werden Farbe, Kontrast und andere Bildeigenschaften für die Texterkennung optimiert, zum Beispiel bei vergilbten oder ausgeblichenen alten Zeitungen. Zudem gilt es, in den Text eingefügte Bilder vom eigentlichen Text zu unterscheiden. Die automatische Texterkennung übersetzt dann diese aufbereiteten Bilder in eine Folge von zunächst Buchstaben und dann Wörtern, die bereits durchsucht werden können. Darauf können dann weitere Analyseverfahren aufsetzen, die zum Beispiel die Struktur einer Zeitungsausgabe rekonstruieren, das heißt Fließtext von Überschriften unterscheiden, den Text in einzelne Artikel segmentieren und den Textfluss in mehrspaltigen Layouts untersuchen. Für all diese Verfahren gibt es bereits industriell verfügbare Lösungen, die auch für den Jugendmedienschutz von Bedeutung sind. Sie ermöglichen es auch, Textinformation, die in Form von Einblendungen in Bildern oder in gescannten Texten vorkommen, der Analyse zugänglich zu machen. Wir geben in diesem Kapitel zunächst einen Überblick über verfügbare Verfahren und ihre

Eigenschaften. In Abschnitt 5.5 werden wir dann am Beispiel des Themas Selbstgefährdung genauer untersuchen, welchen Beitrag diese Verfahren liefern können.

Die optische Texterkennung hat eine ähnlich lange Geschichte wie die automatische Spracherkennung. Eine treibende Kraft war dabei anfangs das Problem der Postsortierung. Deshalb beschäftigen sich viele frühe Verfahren mit der Ziffernerkennung, die dann zur Erkennung von Postleitzahlen dient. Hinzu kam ebenfalls früh die Frage, wie gedruckte Texte computerlesbar gemacht werden können. Dabei hat sich eine allgemeine Strategie für Texterkennungsverfahren entwickelt, deren Schritte von den meisten Verfahren verwendet werden. Unterschiedliche Verfahren unterscheiden sich dann vor allem in der Umsetzung der einzelnen Schritte. Die Texterkennung folgt im Allgemeinen folgenden Schritten:

Vorverarbeitung: Dieser Schritt umfasst alle Techniken, die helfen, die nachfolgenden Aufgaben einfacher zu gestalten. Insbesondere finden hier Schritte wie das Aufrichten (Drehen) der Seite, die Entzerrung von gebogenen Seiten, die Verbesserung des Kontrasts und meist auch eine Binarisierung, das heißt die Reduktion auf die Töne Schwarz und Weiß, statt.

Segmentierung: Ziel dieses Schritts ist das Auffinden der einzelnen Buchstaben auf einer Textseite. Dies wird häufig realisiert, in dem zunächst die Textzeilen anhand der Leerräume zwischen ihnen gefunden werden, und dann die einzelnen Zeilen weiterverarbeitet werden. Bestimmte Buchstabenkombinationen, insbesondere Ligaturen, lassen sich nicht oder nur schwer separieren. In diesem Fall kann auf eine Segmentierung auf Buchstabenebene verzichtet werden, indem etwa alle Ligaturen als zusätzliche Buchstaben aufgefasst werden.

Zeichenerkennung: In diesem Schritt werden kleine Textausschnitte, die möglichst nur noch Bildpunkte enthalten, die zu einem einzelnen Buchstaben gehören, mit Hilfe statistischer Klassifikatoren (vgl. Abschnitt 4.4.2) klassifiziert und so der Buchstabe erkannt. Für das Training dieser Klassifikatoren wird eine möglichst große Trainingsmenge von entsprechenden Bildausschnitten benötigt, für die der Buchstabe bereits bekannt ist.

Integration linguistischer Information: Die Erkennung der einzelnen Zeichen ist mit einer Unsicherheit verbunden. Insbesondere durch Fehler bei der Segmentierung und durch Artefakte vom Scan- oder Binarisierungsprozess kann es zu Fehlklassifizierungen kommen. Auf diese Weise falsch erkannte Buchstaben können aber häufig durch linguistische Zusatzinformationen korrigiert werden. Im einfachsten Fall wird zum Beispiel in einem Wort ein Buchstabe falsch erkannt, so dass das Ergebnis kein gültiges Wort ergibt, sich jedoch eindeutig auf das richtige Wort korrigieren lässt. Hier spielen ähnliche Verfahren wie in Abschnitt 4.4.3.1 beschrieben eine Rolle.

Bei der optischen Texterkennung treten eine Reihe unterschiedlicher Szenarien von unterschiedlicher Komplexität und Schwierigkeit auf:

Drucktext auf weißem Hintergrund: Dies ist sicherlich das am häufigsten betrachtete Szenario, das gerade im Kontext großer Buchdigitalisierungsprojekte von besonderer Bedeutung ist. In Abhängigkeit von der Scan-Qualität und der verwendeten Schriftart ist dies für gewöhnlich das einfachste Szenario.

Komplexe Schriftarten: Schwieriger ist die Behandlung einiger komplexer Schriftarten, zum Beispiel Fraktur oder Textur, die aufgrund vieler Ligaturen und stark ausgeprägter Serifen und anderer Verzierungen bereits bei der Segmentierung größere Schwierigkeiten bereiten.

Komplexer Hintergrund: In Bilder eingebettete oder anderweitig vor einen komplexen Hintergrund gesetzte Schrift erschwert alle Phasen der Texterkennung. Bei der Segmentierung

Tabelle 4.9
Lösungen zur optischen Texterkennung und ihre Verfügbarkeit

Produkt	Lizenz	Quelle
Abbyy Fine Reader	kommerziell	http://finereader.abbyy.de
OmniScan	kommerziell	http://www.omniprisecrm.com/omniscan
Tesseract	Open Source	http://code.google.com/p/tesseract-ocr
NIST Form-Based Handprint Recognition System	Open Source	http://www.nist.gov
GOCR	Open Source	http://jocr.sourceforge.net

kommt es dazu, dass Buchstaben schlecht separiert sind und bei der Zeichenerkennung muss eine größere Variabilität toleriert werden. In manchen Fällen lässt sich dieses Problem allerdings in der Vorverarbeitung lösen, indem der Text vom Hintergrund getrennt wird, zum Beispiel wenn Text und Hintergrund in unterschiedlichen Farben vorliegen.

Verschiedene Schriftsysteme: Neben der lateinischen Schrift sind eine Reihe weiterer Schriftsysteme für die Texterkennung von Interesse und stellen zum Teil besondere Herausforderungen dar. Exemplarisch seien hier die arabische Schrift, die in vielen indischen Sprachen verwendete Devanagri-Schrift, sowie chinesische beziehungsweise japanische Schriftzeichen genannt. Im vorliegenden Kontext werden diese nicht weiter betrachtet.

Handschrift: Die Erkennung von Handschrift ist eine der größten Herausforderungen für die Texterkennung. Hier ist vor allem die große Variabilität von Handschrift entscheidend. Zudem wird bei Verwendung von Schreibschrift auch die Segmentierung erheblich erschwert. Wird die Handschrift direkt über geeignete Eingabegeräte am Computer eingegeben, wird die Erkennung erheblich vereinfacht, da auch Informationen über den Schreibprozess mit in die Erkennung einbezogen werden können.

Die Evaluation von Texterkennungssystemen findet in weniger organisierter Form statt als es für die anderen Verfahren in diesem Kapitel der Fall ist. Daher lässt sich nicht ohne Weiteres ein vergleichbarer Überblick über die Robustheit solcher Verfahren über ein breites Spektrum von Szenarien gewinnen. Hier wird deshalb exemplarisch eine Studie im Bereich der Erkennung von Adressen auf Postsendungen im industriellen Einsatz dargestellt. Abbildung 4.17 zeigt Erkennungsergebnisse in den drei Szenarien Massensendungen, gedruckte Adressen und handgeschriebene Adressen. Hier lässt sich vor allem der klare Unterschied zwischen handgeschriebenem und gedrucktem Text erkennen. Es zeigt sich aber auch, dass sich im industriellen Einsatz insbesondere für gedruckten Text sehr hohe Erkennungsraten erreichen lassen. Dies ist auch für die Anwendung im Jugendschutz relevant, da hier einerseits, wie im industriellen Kontext, große Mengen von Daten verarbeitet werden müssen. Andererseits liegt hier ebenfalls häufig maschinell gesetzter Text vor, der in einer Reihe unterschiedlicher Schriftarten auftreten kann.

Die industrielle Verfügbarkeit von Texterkennungssystemen ist ausgezeichnet. Auch hier ist die Situation vergleichbar mit derjenigen in der automatischen Spracherkennung. Es sind sowohl sehr ausgereifte kommerzielle Systeme verfügbar als auch komfortable Open-Source-Lösungen. Eine Übersicht in Tabelle 4.9 stellt eine Auswahl solcher Systeme zusammen.

Im Kontext des Jugendmedienschutzes treten immer wieder Daten auf, die Text in Bildform enthalten. Diese Texte scheinen durchaus geeignet, diese Daten zu charakterisieren. Die Datenlage passt dabei gut mit einigen der oben beschriebenen Szenarien zusammen. Insbesondere der Umgang mit komplexen Schriftarten und Text vor komplexem Hintergrund ist notwendig, um hier ein überzeugendes Werkzeug zur Verfügung zu stellen.

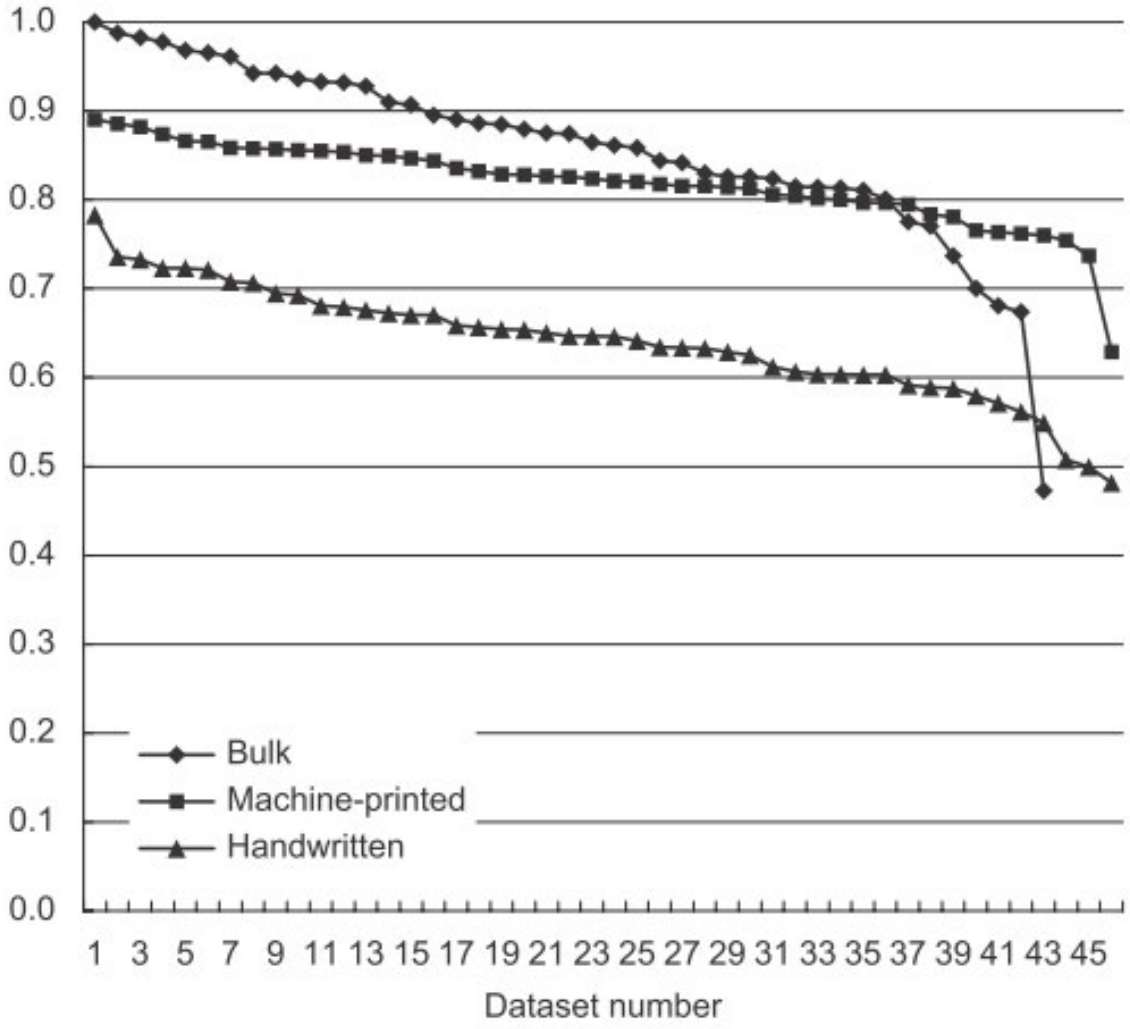


Abbildung 4.17
Ergebnisse einer größeren Evaluation zur automatischen Texterkennung von Adressen auf Postsendungen. Dargestellt sind Erkennungsraten für zunehmend schwierigere Datensätze für drei Arten von Briefen: Massensendungen (Rauten), gedruckte Adressen (Quadrate) und handgeschriebene Adressen (Dreiecke).
Bildquelle: [Fujisawa, 2008]

4.5.4 Logodetektion

Bei der Verfolgung ihrer Kommunikationsziele setzen Firmen und andere Organisationen Logos ein, um mit einfachen Mitteln eine hohe Wiedererkennbarkeit zu erreichen. Dabei besteht der Wunsch, die Erreichung der Kommunikationsziele zu überprüfen und zu dokumentieren. Dazu bedient man sich typischerweise der Medienbeobachtung, um herauszufinden, wo, wie häufig und in welchem Kontext ein bestimmtes Logo gezeigt wird. Dies war bis vor kurzem eine Aufgabe, die ausschließlich von menschlichen Betrachtern durchgeführt werden konnte. Zunehmend steht aber Technologie zur Verfügung, die es ermöglicht, Logos automatisch in Bild- und Videomaterial zu erkennen. Dies ist auch für den Jugendmedienschutz relevant, da jugendschutzrelevantes Material häufig die Logos der Webseiten, durch die sie verteilt werden, tragen. Zudem treten immer wieder Symbole (etwa Hakenkreuze) auf, die für solches Material charakteristisch sind. Die automatische Erkennung solcher Logos in Videos und Bildern kann daher leicht einen Hinweis auf gefährdendes Material geben. In diesem Abschnitt beschreiben wir zunächst die Logodetektion in ihrem klassischen Kontext. In Abschnitt 5.4.7 beleuchten wir dann am Beispiel von Symbolen aus dem rechtsradikalen Bereich, wie Logodetektion konkret in diesem Umfeld eingesetzt werden kann.

In der Logodetektion werden folgende Schritte kombiniert, um effiziente Systeme zu entwickeln:

Extraktion und Repräsentation von Formen: Logos kommen in digitaler Form meist als Pixelbilder vor. Um die eigentliche Form und Struktur der Logos für die Detektion verwenden zu können, müssen sie deshalb zunächst in eine geeignete Form überführt werden. Dazu gibt es zwei konzeptionell unterschiedliche Ansätze. Einerseits werden lokale Merkmale verwendet, um die Logos zu beschreiben. Diese leiten sich zumeist aus auffälligen Punkten, zum Beispiel Ecken, innerhalb der Logos ab und beschreiben die Form des Bildes in einer Umgebung solcher Punkte. Andererseits wird versucht, die Form des Logos durch Polygonzüge wiederzugeben.

Merkmale zur Bild und Formerkennung: Unabhängig davon, ob lokale Merkmale oder Polygonzüge verwendet werden, um Logos zu repräsentieren, werden meist komplexere Merkmale abgeleitet, die dann für den Vergleich herangezogen werden. Für Polygonzüge können dies entweder die Polygone selbst sein oder Merkmale wie Winkelfolgen, die bereits bessere Invarianzeigenschaften aufweisen. Winkelfolgen sind zum Beispiel im Wesentlichen invariant unter Verschiebungen, Drehungen und Skalierungen eines Polygonzugs. Lokale Merkmale werden meist zu Histogrammen zusammengefasst, zum Beispiel in den bereits in Abschnitt 4.5.1.3 erwähnten Histogrammen von Gradientenrichtungen. Außerdem eignet sich auch hier die Verwendung von Visual Words (vgl. Abschnitt 4.5.1.3).

Indexieren und Vergleichen von Formen: Logodetektion wird meist verwendet, um eine kleine Menge von Logos in einer großen Datenbank von Bildern wiederzufinden. Dazu werden die entsprechenden Merkmale aus allen Bildern extrahiert und dann in einer für den schnellen Vergleich entworfenen Datenstruktur gespeichert. Im Polygon-Ansatz werden dazu Abstandsbegriffe für Polygonzüge benötigt. Hierzu eignen sich zum Beispiel die Hausdorff- und die Frechét-Distanz. Zudem können Merkmale wie Winkelfolgen als Zeichenketten aufgefasst werden und so mit Verfahren aus der Textsuche verglichen werden. Für den Vergleich von Formen auf Basis lokaler Merkmale können zum Beispiel der gemeinsame Informationsgehalt (Mutual Information) von Teilbildern in der Datenbank und dem Anfragelogo bestimmt werden. Alternativ kann für jeden Bildpunkt in der Datenbank bewertet werden, wie ähnlich die Visual Words, in denen er vorkommt, zu denjenigen im Anfragelogo sind.

Die besondere Herausforderung bei der Logo-Erkennung besteht darin, Logos auch dann wiedererkennen zu können, wenn sie verzerrt oder teilweise verdeckt sind. Tabelle 4.10 gibt einen Überblick

Tabelle 4.10
Robustheit von Verfahren zur Logo-Erkennung, erweitert nach [Chen et al., 2011]. Die betrachteten Störungen, gegenüber denen die Verfahren robust sein sollten, sind: SC: Skalierung, RO: Rotation, BL: zufällig unterbrochene Linien, AS: zufällig eingefügte Bildstreifen, OC: Verdeckung, GN: Gaußsches Rauschen, SK: Scherungen, CE: komponentenweise Änderung, PD: Foto-Deformationen.

Verfahren	SC	RO	BL	AS	OC	GN	SK	CE	PD
MLHD [Chen et al., 2011]	✔	✔	✔	✔	✔	✔	✔		
Geometrische Invarianten [Doermann et al., 1996]	✔	✔							
Formmerkmale [Soffer und Samet, 1998]	✔						✔	✔	
Zernike Momente [Kim und Kim, 1998]	✔	✔	✔			✔			✔
String-Matching [Ting, 1998]	✔	✔					✔		
Template-Matching [Jain und Vailaya, 1998]	✔	✔				✔			
Formkontext [Zhu und Doermann, 2009]	✔	✔				✔			
Merkmalskombinationen [Mehtre et al., 1997]	✔	✔							
Globale Formstruktur [Wei et al., 2009]	✔	✔					✔	✔	
Interaktives Feedback [Ciocca und Schettini, 2001]	✔	✔							
Neuronale Netze [Cesarini et al., 1995]	✔	✔				✔			
Autoassoziator Netze [Cesarini et al., 1997]	✔	✔	✔	✔	✔				
Edge Backpropagation [Gori et al., 2003]	✔	✔	✔	✔	✔	✔	✔		
Visual Phrases [Jiang et al., 2012]	✔	✔	✔	✔	✔	✔	✔	✔	✔

über Verfahren zur Logo-Erkennung sowie deren Fähigkeiten, typische Veränderungen, die eine Erkennung der Logos nicht beeinflussen sollten, zu tolerieren. Solche Veränderungen entstehen zum Beispiel bei der Aufnahme von Logos in natürlichen Bildszenen und umfassen Effekte wie Skalierung (Größenänderung), Rotation, verschiedene Formen von Verdeckungen, Bildrauschen, Scherungen (etwa durch perspektivische Verzerrungen) oder Effekte wie durch Bildbearbeitung am Computer. Die in der Tabelle aufgeführten Verfahren sind mannigfaltig, folgen aber alle dem oben skizzierten Grundaufbau. Hinzu kommen noch auf Neuronalen Netzen basierende Verfahren, die zunächst Formmerkmale hierarchisch anhand der Schachtelung von Konturen extrahieren, um dann Logos anhand von Trainingsbildern mit einem hierarchischen Neuronalen Netz als Klassifikator zu lernen (ähnlich wie in Abschnitt 4.4.2.3 beschrieben).

Im Bereich der Logodetektion gibt es keine zentralen Evaluierungskampagnen. Allerdings steht mit BelgaLogos [Joly und Buisson, 2009] ein großes Vergleichskorpus mit realistischen Detektionsaufgaben zur Verfügung. Abbildung 4.18 zeigt Beispiele zur Logodetektion in diesem Korpus. Man sieht, dass es sich dabei um recht schwierige Aufgabenstellungen handelt, in denen die Logos praktisch allen denkbaren Transformationen unterzogen werden.

Um ein Gefühl für den Stand der Forschung in diesem Bereich zu bekommen, fassen wir die Ergebnisse von Verfahren zusammen, die in den Jahren 2009 bis 2012 veröffentlicht wurden und diesen Vergleichskorpus zur Evaluation genutzt haben. Dabei ist zu beachten, dass die Erkennungsaufgabe auf diesem Korpus sehr schwer ist, da die Logos häufig stark verzerrt oder verdeckt sind. Dies ist im Jugendschutzkontext weniger häufig der Fall. Die Ergebnisse hängen dabei unabhängig vom Verfahren sehr stark vom Logo ab. Je nach Logo liegt die Präzision der Verfahren zwischen 10 % und 100 %. Abbildung 4.19 zeigt typische Ergebnisse für eine Reihe von Logos anhand eines aktuellen Verfahrens und eines Vergleichsverfahrens, das zusammen mit dem BelgaLogos-Korpus eingeführt wurde. Die Präzision der Verfahren wird dabei meist mit Hilfe der Mean Average Precision gemessen.

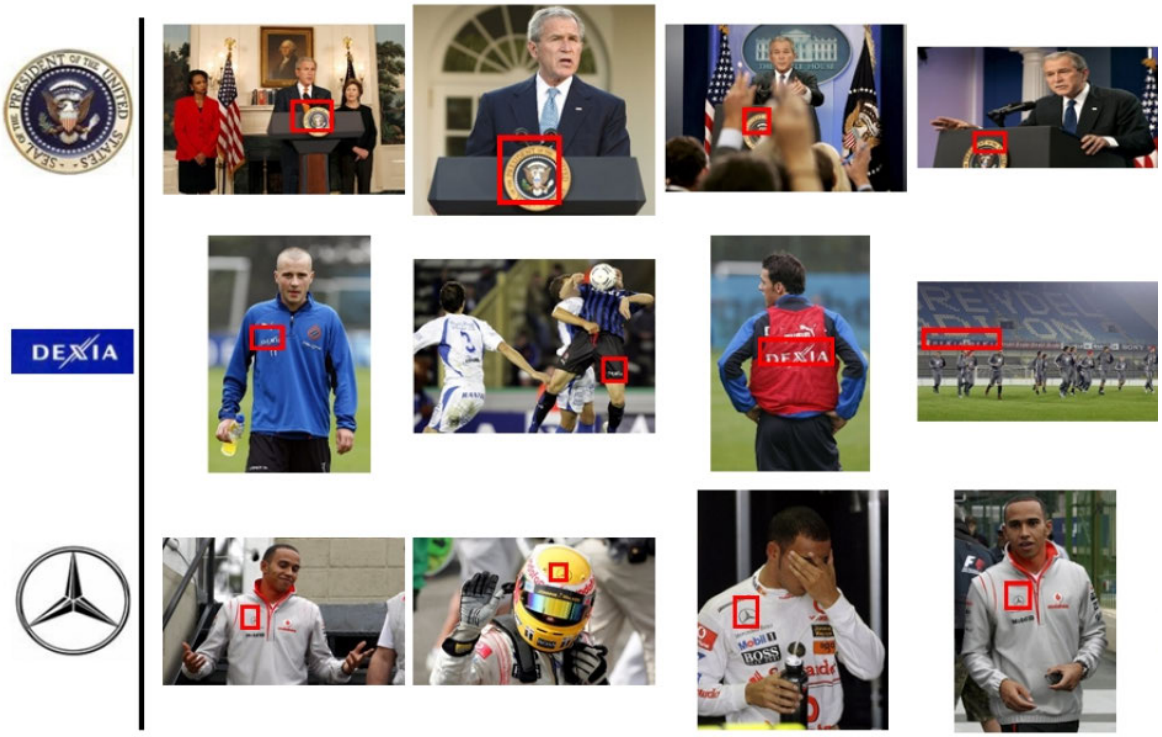


Abbildung 4.18
Beispiel zur Logodetektion anhand des BelgaLogos-Korpus. Die Abbildung zeigt für drei Beispiele von Logos als Anfrage die jeweils vier besten Treffer aus der Datenbank. Dabei sieht man, dass die Logos trotz einer Reihe von Deformationen, zum Beispiel starke Skalierung, Scherung und Verzerrung auf Kleidung und teilweise Verdeckung, erkannt werden.
Bildquelle: [Jiang et al., 2011]

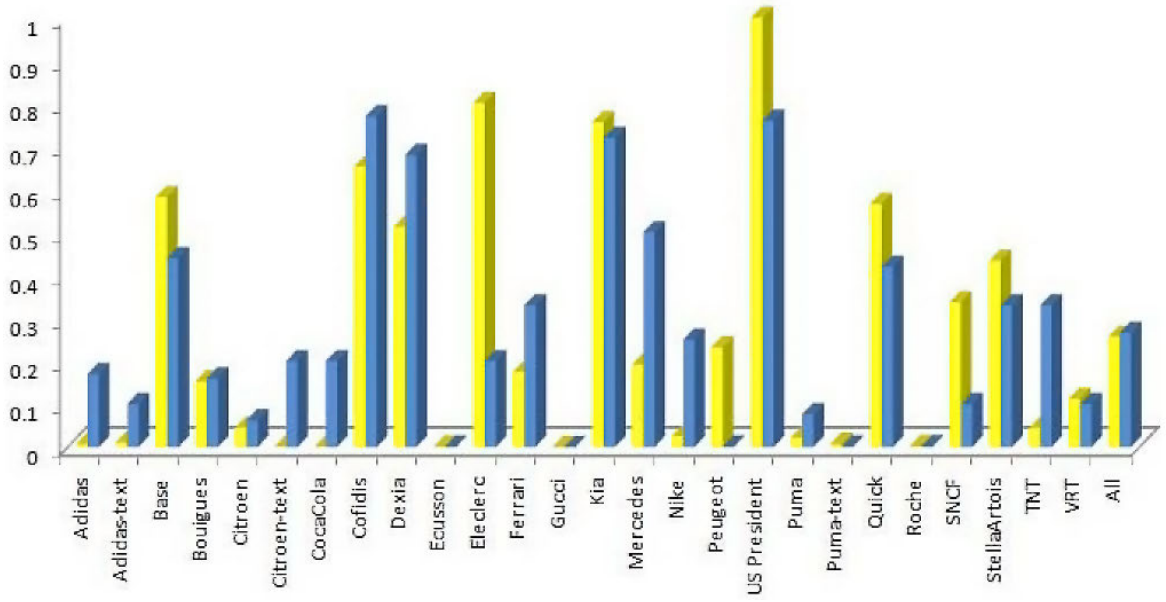


Abbildung 4.19
Vergleich zweier Verfahren zur Logodetektion auf dem BelgaLogos-Korpus. Die Balken geben die Mean Average Precision an. Dabei sind die Ergebnisse für das Verfahren [Joly und Buisson, 2009], das zusammen mit dem BelgaLogos-Korpus eingeführt worden ist, in gelb und die für ein aktuelles Verfahren [Fu et al., 2010] in blau dargestellt.
Bildquelle: [Fu et al., 2010]

Diese bestimmt zunächst für jede Anfrage in einer nach Ähnlichkeitswert sortierten Ergebnisliste die durchschnittliche Präzision. Diese Ergebnisse werden dann über alle Anfragen gemittelt.

Die industrielle Verfügbarkeit von dedizierter Software zur Logodetektion ist noch recht beschränkt. Häufig ist die Technologie in Produkten zur Medienbeobachtung integriert, die nur als Gesamtpaket vertrieben werden. Orpix Logo Recognition[21] scheint derzeit das einzige kommerzielle Produkt speziell zur Logodetektion zu sein. Auch spezielle Open-Source-Projekte zu diesem Thema scheinen nicht verfügbar. Allerdings sind viele der grundlegenden Bausteine, die zur Logodetektion notwendig sind, in Computer Vision Bibliotheken wie zum Beispiel OpenCV[22] enthalten.

Jugendschutzrelevantes Material wird immer wieder auf dedizierten Webseiten gesammelt und über diese verteilt. Häufig werden dabei Logos der entsprechenden Seite in Bilder und Videos eingefügt. Auch extremistische Gruppierungen identifizieren sich häufig durch Logos und Symbole. Vor diesem Hintergrund erscheint der Einsatz von Logoerkennung im Jugendmedienschutz sehr sinnvoll. Die Datenlage zeigt sich den im wissenschaftlichen Kontext untersuchten Situationen sehr ähnlich, so dass die vorhandenen Verfahren ohne Weiteres anwendbar sein sollten, wenn das Problem der Softwareverfügbarkeit gelöst ist.

4.5.5 Konzepterkennung

Eine Reihe von Gewalt-Ereignissen, insbesondere solche, die in den Bereich des Terrorismus fallen, hat dazu geführt, dass in den letzten Jahren verstärkt das Bedürfnis besteht, die Sicherheit im öffentlichen Raum zu erhöhen. In diesem Zusammenhang wurde in vielen Ländern die Videoüberwachung,

[21] siehe http://www.orpix-inc.com/#!products/vstc11=logo-recognition, Stand 15.10.2012
[22] siehe http://opencv.org/, Stand 15.10.2012

zum Beispiel auf öffentlichen Plätzen oder in öffentlichen Verkehrsmitteln, erheblich ausgebaut. Dies führt dazu, dass vielfach erheblich mehr Material von Überwachungsvideos zur Verfügung steht als tatsächlich überwacht werden kann und dieses Material deshalb in den meisten Fällen erst nach einem Ereignis im Rahmen der Aufklärung eingesetzt wird. Automatische Verfahren zur Detektion relevanter Ereignisse können hier helfen, die Aufmerksamkeit auf diese Ereignisse zu richten und so eine frühere Reaktion zu ermöglichen und gleichzeitig weniger unverdächtiges Verhalten sichten zu müssen. Da Ereignisse wie Gewalt sich meist durch eine hohe Variabilität in der Realisation auszeichnen, fasst man sie in Klassen zusammen, den sogenannten Konzepten. Ausgehend von solchen und ähnlichen Szenarien haben sich Forscher dem Thema der Konzepterkennung zugewandt, in der verschiedene Konzeptklassen, zum Beispiel so diverse Konzepte wie *Geburtstagsfeier* oder *Zubereiten eines Sandwiches*, erkannt und auseinandergehalten werden sollen. Im Kontext des Jugendmedienschutzes sind eine Reihe von Konzepten, etwa bestimmte Gewalthandlungen oder Konzepte aus dem rechtsradikalen Milieu, wie der Hitlergruß, denkbar, deren automatische Erkennung klare Hinweise auf jugendschutzrelevantes Material liefern könnte. Wir wollen in diesem Abschnitt zunächst wieder den Stand der Technik auf diesem Gebiet zusammenfassen. In Abschnitt 5.5.4 wird anhand des Konzepts *Ritzbild* untersucht, wie diese Methodik im Zusammenhang selbstgefährdenden Verhaltens eingesetzt werden kann.

Das Thema Konzepterkennung umfasst eine Reihe unterschiedlicher Aufgabenstellungen, und Anwendungen in diesem Bereich bedienen sich entsprechend unterschiedlicher Methoden. Das Hauptaugenmerk in diesem Zusammenhang liegt in der Erkennung verschiedener Arten menschlicher Handlungen aus Videoaufnahmen. Allerdings können auch, wie im vorangegangenen Absatz beschrieben, Konzepte von Interesse sein, die aus Einzelbildern erkennbar sind. Letztere können mit geeigneten Merkmalen, wie den zuvor beschriebenen Visual Words (vgl. Abschnitt 4.5.1.3) und geeigneten Klassifikationsverfahren wie in Abschnitt 4.4.2 beschrieben, behandelt werden. In diesem Kapitel wird verstärkt die Handlungserkennung aus Videos betrachtet.

Verfahren zur Handlungserkennung bestehen im Kern aus zwei Schritten:

1. **Merkmalsextraktion:** Zunächst werden aus dem Videostrom Merkmale extrahiert, die für die Handlungsrepräsentation geeignet sind. Dazu gibt es drei Herangehensweisen:
   - Merkmale, die aus dem optischen Fluss abgeleitet werden,
   - Merkmale, die speziell für die Handlungserkennung entworfen worden sind,
   - Sequenzen von Posen, die durch menschliche Silhouetten beschrieben werden.
2. **Klassifikatoren:** Anhand von Trainingsdaten für die zu erkennenden Handlungsklassen wird mit Hilfe der Merkmale ein statistisches Klassifikationsverfahren wie in Abschnitt 4.4.2 beschrieben trainiert und so ein Klassifikator für eine Menge von Handlungsklassen gelernt.

Der optische Fluss beschreibt, wie sich die Bildpunkte in Videos von einem Bild zum nächsten bewegen und eignet sich deshalb als Ausgangsbasis, um Bewegungen zu charakterisieren. Speziell für die Handlungserkennung entworfene Merkmale modellieren zum Beispiel Objekte als Konstellationen flexibler Elemente, die mit lokalen Bildausschnitten in Übereinstimmung gebracht werden. Diese werden durch geeignete Verfahren, die interessante Regionen in Bildern identifizieren, bestimmt. Solche Merkmale erlauben zwar gute Ergebnisse, sind aber in der Berechnung sehr zeitaufwändig. Neurobiologische Studien [Giese und Poggio, 2003] zeigen, dass Sequenzen von Posen für den menschlichen Betrachter hinreichend sind, um Handlungen zu erkennen.

Aufgrund der hohen Variabilität in der Ausführung von Handlungen über verschiedene Personen hinweg konzentriert sich die Forschung derzeit zum Teil auf die Entwicklung von Verfahren, die auf eine einzelne Person adaptiert werden. Ein weiterer Ansatz, der von der Einzelperson abstrahiert, beschäftigt sich mit dynamischen Klassifikationsschemata. Dabei werden Systemparameter entweder dynamisch oder mit Hilfe von Trainingsdaten an die Anwendungsdomäne angepasst. Ein Klassifika-

tionsverfahren, das in der Handlungserkennung gute Ergebnisse erzielt, ist durch Neuronale Netze gegeben. Allerdings sind die verwendeten Trainingsverfahren sehr langsam und benötigen ein gewisses Maß an manueller Interaktion durch den Benutzer. Daher wurde als Alternative das Verfahren der Extreme Learning Machine vorgeschlagen. Diese kann als ein Neuronales Netz verstanden werden, dass zunächst eine nicht-lineare Abbildung in einen hochdimensionalen Merkmalsraum vornimmt und dort eine lineare Klassifikation durchführt. Dieses Verfahren bringt Ideen aus den Bereichen Neuronale Netze und Support Vektor Maschinen zusammen und eignet sich gut für die Handlungserkennung.

Einen guten Überblick über den Stand der Technik gibt auch in diesem Fall die TRECVID-Evaluation. Im Jahr 2011 wurden hierbei Verfahren zur Multimedia-Event-Erkennung anhand der folgenden Event-Klassen getestet:

- Tierpflege (Grooming Animal)
- Näharbeit (Sewing Project)
- Ein Sandwich zubereiten (Sandwich)
- Umzug (Parade)
- Reifenwechsel (Changing Tire)
- Geburtstagsfeier (Birthday Party)
- Feststeckendes Fahrzeug befreien (Vehicle Unstuck)
- Parkour (Parkour)
- Reparieren eines Apparats (Repairing Appliance)
- Flash-Mob (Flash Mob)

Zu jeder dieser Klassen lagen gut 100 Videos als Trainingsmaterial vor. Die Klassifikatoren wurden dann auf circa 100 unbekannten Testvideos pro Klasse getestet. Abbildung 4.20 zeigt jeweils das beste Ergebnis zu jeder Event-Klasse bei insgesamt 19 teilnehmenden Forschergruppen.

Diese Ergebnisse zeigen, dass diese Aufgabenstellung noch sehr schwierig ist. Im Falle der besten Klasse werden bei circa 1 % Fehlalarmen gut 14 % Fehlklassifikationen vorgenommen. Für die meisten Klassen liegen diese Raten jedoch deutlich höher, im schlechtesten Fall liegen bei knapp 3 % Fehlalarmen gut 34 % Fehlklassifikationen vor.

In früheren Auflagen der TRECVID-Evaluation wurde unter der Aufgabenstellung *Feature Extraction* auch die Kategorie *Physical Violence*, also körperliche Gewalt untersucht. Dies ist für die Anwendung im Jugendschutz hoch relevant. Die Erkennungsraten waren jedoch sehr gering. Im Rahmen der Evaluationskampagne MediaEval wurde in den Jahren 2011 und 2012 eine Aufgabenstellung zur automatischen Detektion von Gewaltszenen in Spielfilmen bearbeitet. Die Ergebnisse waren hier im Jahr 2011 noch weit von der für einen sinnvollen Einsatz nötigen Qualität entfernt. Je nach Material und Verfahren wurden meist F-Werte im Bereich von 10–20 % erreicht, in seltenen Fällen um die 30 %. Im Jahr 2012 ließen sich aber schon deutliche Fortschritte erkennen. In einer Bewertung der jeweils 100 am wahrscheinlichsten als Gewalt eingestuften Szenen pro Film wurde von den besten Systemen eine Präzision von etwa 60 % erreicht. In der Praxis bedeutet dies, dass diese Systeme einen menschlichen Bewerter auf Szenen hinweisen können, die vermutlich Gewalt enthalten und wenigstens jede zweite vorgeschlagene Szene auch tatsächlich Gewalt enthält. Dies deutet darauf hin, dass diese Systeme in naher Zukunft geeignet sein dürften, menschliche Bewerter zu unterstützen, in dem diese sich nur kurze Ausschnitte aus Filmmaterial anschauen müssen.

Die Konzepterkennung ist noch ein sehr junges Gebiet und entsprechend ist industrielle Verfügbarkeit von Systemen zur Konzepterkennung gering. Es sind praktisch weder kommerzielle noch Open-Source-Systeme verfügbar. Allerdings ist die zugrunde liegende Technologie, vor allem im Bereich der Klassifikatoren, vielfach als Open-Source-Lösung verfügbar.

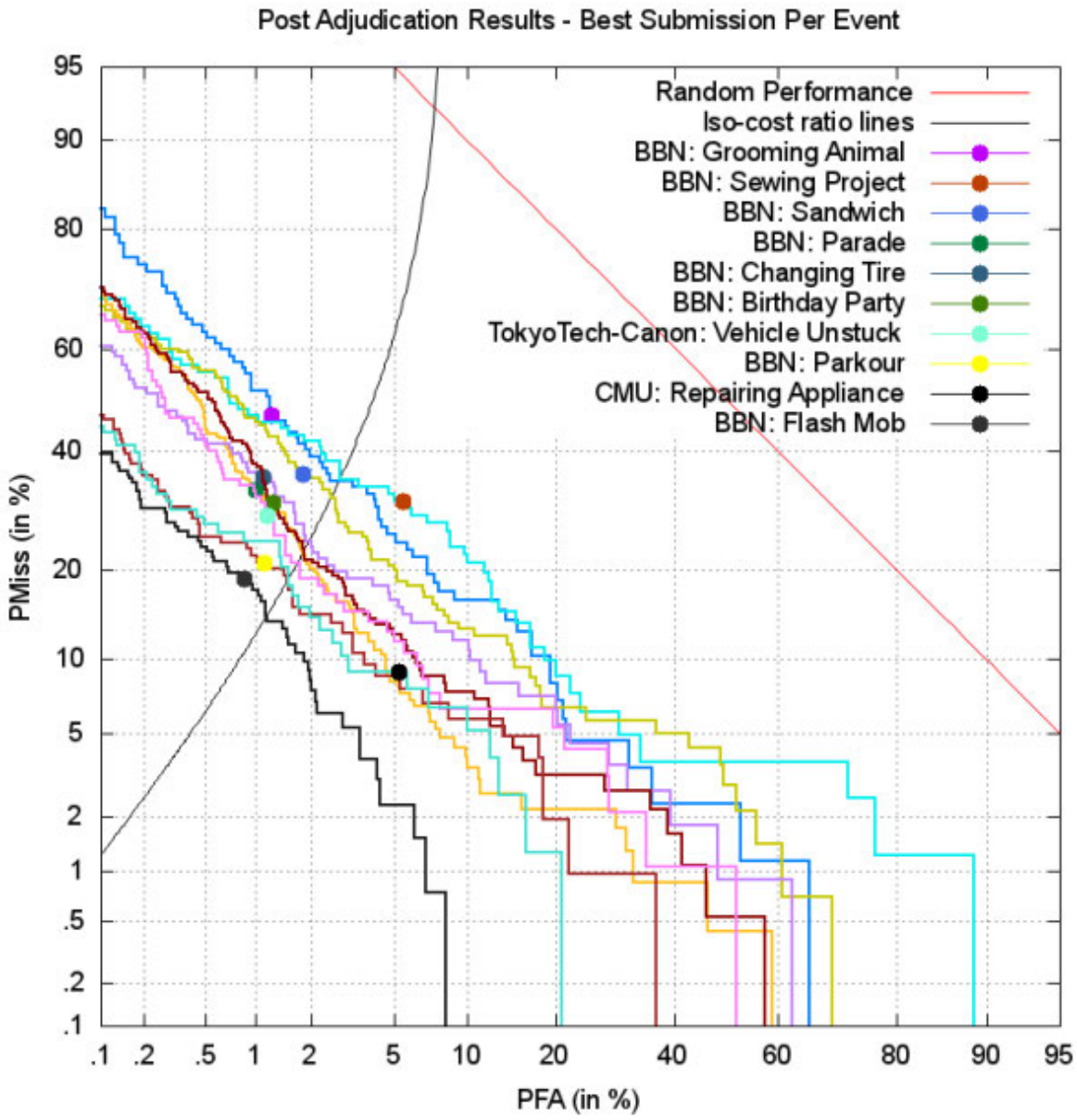


Abbildung 4.20
TRECVID-Evaluationsergebnisse aus dem Jahr 2011 zur Multimedia-Event-Erkennung. Aufgetragen ist die Wahrscheinlichkeit, ein Konzept zu erkennen, obwohl es nicht vorliegt (PFA) gegenüber der Wahrscheinlichkeit, ein Konzept nicht zu erkennen, obwohl es vorliegt (PMiss).
Bildquelle: [Antoniskek et al., 2011]

Die Konzepterkennung steckt noch weitgehend in den Kinderschuhen. Dies ist bereits an der Auswahl der Konzepte in den entsprechenden Evaluierungskampagnen ablesbar. Die dort verwendeten Konzepte wie *einen Kuchen backen* oder *Reparieren eines Apparats* sind eher anhand der verfügbaren Daten als an tatsächlichen Anwendungen motiviert. Dennoch sind Verfahren zur Konzepterkennung meist generalisierungsfähig und lassen sich für neue Konzeptklassen trainieren, wenn entsprechende Trainingsdaten vorhanden sind. Dies wird in Abschnitt 5.5.4 am Beispiel eines Konzepts aus dem Bereich Selbstgefährdung untersucht.

4.6 Umfeldanalyse: Klassifikation einer Webseite im Kontext ihres Umfelds

Neben dem Inhalt einer einzelnen Webseite bieten sich zusätzlich weitere Merkmale zur Einstufung des Inhalts an. Über die Analyse von Links kann das Umfeld, also der Kontext, einer Webseite erschlossen werden.

Die Lage einer Webseite im Webgraph, also in der aus Hyperlinks gebildeten Graphstruktur des WWW, liefert auch unabhängig vom eigenen Inhalt der Webseite wichtige Informationen über ihre inhaltliche Orientierung. Abbildung 4.21 zeigt die Beziehung zwischen einer Webseite und ihren durch Links verbundenen Nachbarn. Verweisen also beispielsweise mehrere nach ihrem Inhalt als rechtsextrem klassifizierte Webseiten (Eltern) auf die aktuell betrachtete Seite, oder führen die ausgehenden Links einer Webseite auf entsprechende Inhalte (Kinder), so steigt die Wahrscheinlichkeit, dass auch die aktuelle Webseite in diesem Bereich anzusiedeln ist.

Zur Verwendung dieser Merkmale ist allerdings ein vollständiger Webgraph die notwendige Voraussetzung. Diesen zu beschaffen und zu speichern ist aufwändig und teuer. Alternativ lassen sich auch Suchmaschinen dazu nutzen, die Liste der eingehenden Hyperlinks aufzubauen, indem spezielle Anfragen[23] gestellt werden.

Mit Hilfe von Linkanalysealgorithmen können Rückschlüsse darüber gewonnen werden, ob eine bestimmte Webseite eine mehr oder weniger prominente Rolle im Webgraphen einnimmt. Solchen Algorithmen liegt die Annahme zugrunde, dass das Surfen im Web einem Random Walk folgt. Das bedeutet, dass angenommen wird, dass an einem zufälligen Knoten des Graphs begonnen wird und in jedem Schritt ein zufälliger Link zur Fortsetzung des Surfens ausgewählt wird. Ursprünglich hat die Linkanalyse ihre Wurzeln in der Analyse von Zitationen in Fachzeitschriften (zum Beispiel dem Impact Factor nach Garfield [Garfield, 1972]). Der erste prominente Vertreter der Linkanalyse ist der HITS-Algorithmus nach Kleinberg [Kleinberg, 1999]. Dabei unterscheidet Kleinberg Webseiten, die von vielen anderen verlinkt werden (sogenannte »authorities«) und solche, die selbst zu vielen prominenten Seiten zeigen (sogenannte »hubs«). Diese können iterativ aus dem Webgraphen berechnet werden. Der bekannteste Vertreter der Linkanalyse ist aber sicher der von Google entwickelte PageRank- Algorithmus, eine vereinfachte Variante von HITS [Brin und Page, 1998]. Hier wird die Unterscheidung zwischen verschiedenen Typen von Seiten aufgehoben, nur ein Wert, der sogenannte PageRank, wird für alle Webseiten berechnet. Der PageRank ist ein Indikator für die Bedeutung einer Webseite, jeder Link zu der Webseite zählt als Stimmabgabe. Der PageRank ist rekursiv definiert und wird iterativ berechnet. Eine Webseite, die von vielen Seiten mit hohem PageRank verlinkt wird, hat dann selbst einen hohen PageRank. Damit bildet der PageRank-Algorithmus einen zufällig durch das Netz surfenden Benutzer nach, also den Random Walk des Websurfens.

Neben dem direkten Rückschluss auf die Kategorie einer verlinkten Webseite können aus dem Umfeld des Links weitere Merkmale gewonnen werden. Ein Überblick findet sich in [Qi und Davison, 2009]. Primär bietet sich der Linktext selbst an, aber auch Informationen aus

[23] bei Google zum Beispiel link:http://xyz.de/ oder bei http://www.opensiteexplorer.org/, Stand 19.12.2012.

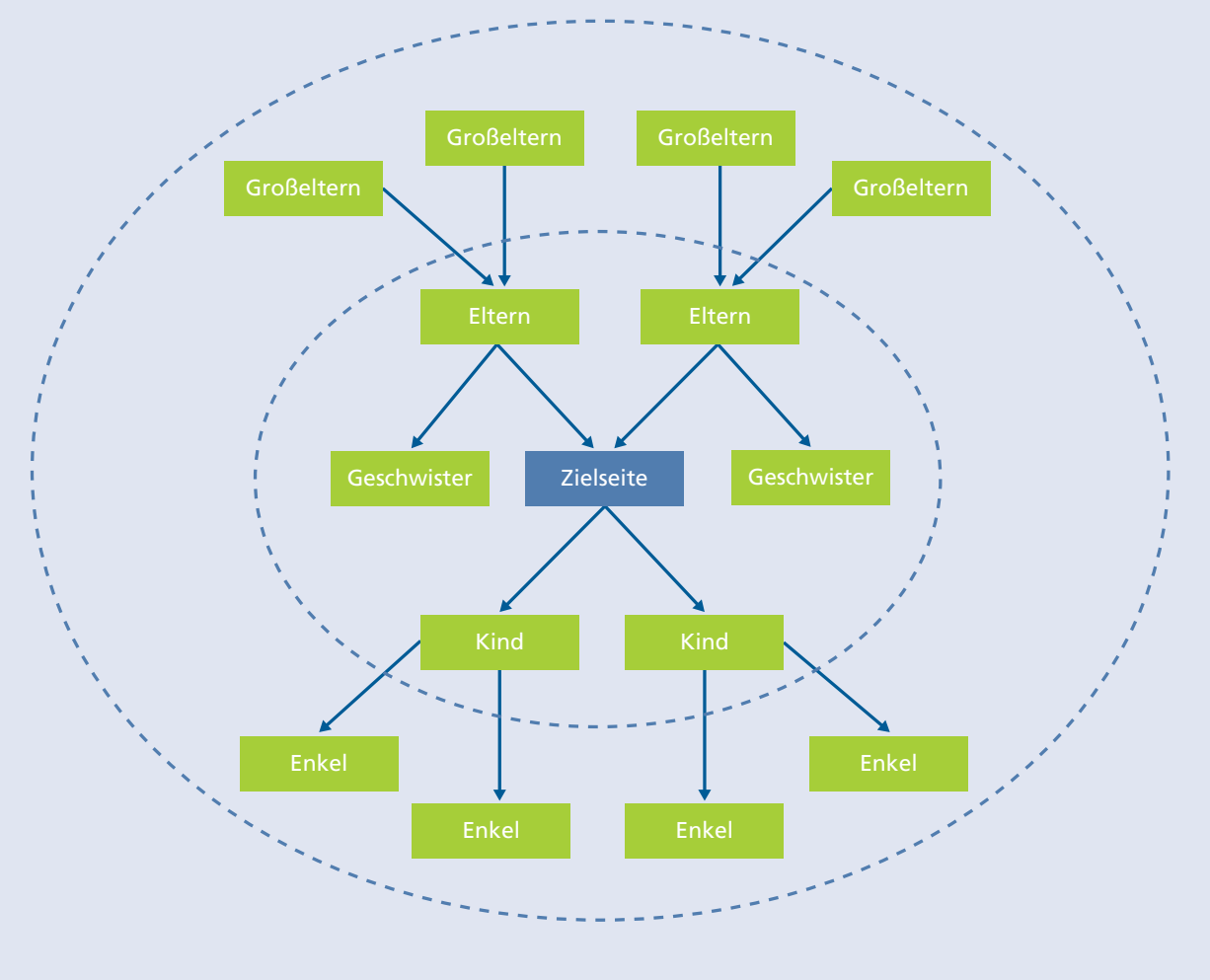


Abbildung 4.21
Die Nachbarseiten einer Zielseite gemäß der Struktur der gerichteten Weblinks, welche durch Pfeile symbolisiert sind. Die von der Zielseite ausgehenden Links verweisen auf die Kinder. Von dort ausgehende Links gehen zu den Enkeln. Umgekehrt gibt es Links von den Eltern auf die Zielseite und von den Großeltern auf die Eltern.

der textuellen Umgebung des Links lassen sich sinnvoll einsetzen. So können zum Beispiel auf Aufklärungsseiten auch jugendschutzrelevante Inhalte verlinkt sein: »Die für rechtsextreme Hetze bekannte Webseite …meldete gestern…«. In diesem Fall ist ein direkter Rückschluss von verlinkender auf verlinkte Webseite nicht möglich. Eine Analyse des Linktextes und des Bereiches um den Link herum kann allerdings die richtige Kategorisierung der verlinkten Webseite ermöglichen. [Qi und Davison, 2009] empfehlen zudem, den Ankertext in den Links der Eltern zu berücksichtigen und auch die Klassifikation von Geschwisterseiten stärker einzubeziehen.

[Eickhoff et al., 2011] führten eine umfangreiche Studie zur Klassifikation von Webseiten für Kinder durch. Als Merkmale berücksichtigten sie die Klassifikation der verlinkten Seiten (Kinder): einerseits den Anteil der verlinkten Seiten, die als »Erwachsenenseite« klassifiziert wurden und andererseits den Anteil der Seiten, die als »Kinderseite« klassifiziert wurden. Es zeigte sich, dass diese beiden Merkmale mit weitem Abstand den höchsten Informationsgehalt aller Merkmale besaßen. Experimente ergaben, dass eine Analyse der Enkelseiten keinen zusätzlichen Nutzen hatte. Die positiven Ergebnisse der Studie werden in Abschnitt 4.4.5 diskutiert.

4.7 Fusion von Klassifikationsergebnissen

In den vorangegangenen Abschnitten wurde eine Reihe von ganz unterschiedlichen Verfahren vorgestellt, die Hinweise liefern können, ob die Inhalte einer Webseite als jugendschutzrelevant einzustufen sind oder nicht. Daraus leitet sich die Frage ab, wie all diese Teilergebnisse in einem Gesamtsystem zusammengeführt werden können. Es gibt zunächst eine breite Literatur zu Verfahren, die sich damit beschäftigt, die Ausgaben verschiedener Klassifikatoren für die gleiche Klassifikationsaufgabe so zu kombinieren, dass das Gesamtergebnis besser ist als das Ergebnis des besten einzelnen Klassifikators (siehe auch Abschnitt 4.4.2.7). Typische Verfahren in diesem Bereich sind die in Abschnitt 4.4.2.7 beschriebenen Boosting-Verfahren. Sie basieren auf Klassifikatoren, die zwar zu viel Material in die zu erkennende Klasse einsortieren, aber zur Klasse gehörendes Material praktisch nie zurückweisen. Durch geschickte Kombination einer ganzen Reihe solcher Klassifikatoren, die jeweils unterschiedliches Material korrekt zurückweisen, erreicht man ein siebartiges Konzept. Die einzelnen Siebe (Klassifikatoren) sind dann für gewöhnlich so eingestellt, dass sie alle erwünschten Daten passieren lassen, aber in jeder Siebstufe unterschiedliche, nicht erwünschte Daten aussortieren. Im Folgenden soll jedoch nicht näher auf diese Verfahren eingegangen werden, sondern an einem konkreten Beispiel gezeigt werden, wie ein Gesamtsystem auf Basis von Einzelklassifikatoren aussehen kann.

Der INDECT Multimedia Crawler [Grega et al., 2011] ist ein System, das viele der in diesem Kapitel beschriebenen Aspekte aufgreift und sich mit der Integration von Klassifikationsverfahren zur Bekämpfung von Online-Kriminalität in einem Webcrawler beschäftigt. Das Einsatzszenario ist ähnlich wie das dieser Studie. Es sollen automatisch Webseiten aufgefunden werden, die mit hoher Wahrscheinlichkeit strafrechtlich relevantes Material beinhalten. Das System soll allerdings keine Filterliste erstellen, sondern direkt von den Strafverfolgungsbehörden eingesetzt werden. Es besteht zunächst aus einem Webcrawler, der ausgehend von initialen Webseiten anhand von Links alle von dort aus erreichbaren Webseiten untersucht. Über ein Plugin-Konzept können Multimedia-Analyseverfahren eingebunden werden, die dann die gefundenen Webseiten weiter untersuchen.

Die derzeit verwendeten beziehungsweise in der Entwicklung befindlichen Komponenten dieses Systems behandeln die folgenden Aspekte:

- In einem Modul zur Detektion von Webseiten-Elementen werden zunächst solche Bilder, die zum Design der Webseite gehören, von denjenigen, die tatsächlich Inhalt tragen, unterschieden.
- Mit Hilfe von Gesichtsdetektion wird festgestellt, ob und wo in einem Bild ein Gesicht vorhanden ist.
- Mittels Symbolerkennung werden bestimmte verbotene Symbole, zum Beispiel Hakenkreuze, in Bildern erkannt.
- Eine Methode zur Formerkennung soll die Symbolerkennung um die Möglichkeit erweitern, Symbole anhand von Beispielen zu erkennen.
- Die Erkennung von Kinderpornografie ist ein zentrales Anliegen des Projekts und befindet sich in der Entwicklung.
- Anhand von automatischer Spracherkennung sollen Phrasen aus Tonaufnahmen und Videos extrahiert werden, die für die strafrechtliche Bewertung relevant sind.
- Optische Texterkennung soll das Gleiche für Schrift in Bildern und Videos leisten.
- Basierend auf der Gesichtsdetektion soll die Gesichtserkennung es ermöglichen, gezielt anhand von Beispielen nach Bildern, die bestimmte Personen darstellen, zu suchen.

Bemerkenswert ist, dass viele dieser Module auf Basis von Open-Source-Komponenten entwickelt werden.

Das System ist so ausgelegt, dass ein Teil der Komponenten für eine Vorverarbeitung der Webseiten sorgt und dann die restlichen Komponenten das Material klassifizieren. Bisher werden für die Kombination lediglich die einzelnen Ergebnisse zusammengeführt, das heißt jede Webseite, die von einem der Verfahren gemeldet wird, erscheint in einer grafischen Benutzeroberfläche. Dennoch zeigt dieses System schon recht eindrucksvoll, wie ein Gesamtsystem aus Text- und Multimedia-Analyseverfahren angegangen werden kann. In Kapitel 6 wird detaillierter untersucht, wie die Integration von Mechanismen zum Jugendmedienschutz in ein umfassendes System erfolgen sollte.

4.8 Zusammenfassung

In diesem Kapitel haben wir ausgewählte Techniken aus dem Bereich der Text- und Multimedia-Analyse identifiziert und vorgestellt, die zunächst aufgrund der Datenlage für den Einsatz im Jugendmedienschutz sinnvoll erscheinen. Ein Vergleich des aktuellen Stands der Technik mit den Anforderungen in diesem Kontext zeigt, dass die meisten dieser Verfahren einen Entwicklungsstand erreicht haben, der einen Einsatz, zumindest in unterstützender Funktion, auch realistisch erscheinen lässt. Zudem ist für die meisten Verfahren die industrielle Verfügbarkeit hoch, so dass sie in Softwaresysteme[24] für den Jugendmedienschutz integrierbar wären. Konkret ergibt sich für die einzelnen Verfahren folgendes Bild:

- In der Textanalyse sind in den letzten Jahren starke Verfahren für die Informationsextraktion und die Textklassifikation entwickelt worden. Sie zeigen in Szenarien, die direkt mit dem Jugendmedienschutz in Verbindung stehen, sehr gute Ergebnisse und sind daher äußerst vielversprechend für die Weiterentwicklung entsprechender Softwarelösungen. Es sind ausgereifte Open-Source-Programmsammlungen verfügbar.
  - Themenmodelle erzeugen eine automatische Gruppierung von thematisch zusammengehörigen Wörtern. Sie benötigen keine Trainingsdaten, ordnen selbsttätig neue Begriffe in Themen ein und können mit vorliegenden Kategorieinformationen kombiniert werden.

[24] Die Frage, an welcher Stelle eines Gesamtsystems für den Jugendmedienschutz diese Verfahren integriert werden sollten, wird in Kapitel 6 erörtert.

- Klassifikatoren ordnen Dokumente in vorbestimmte Kategorien ein, welche durch Beispieldaten definiert sind. Sind genügend umfangreiche Trainingsdaten verfügbar, so lassen sich sehr oft hohe Klassifikationsgenauigkeiten erreichen. Die Genauigkeit der maschinellen Lösung liegt teilweise nahe an der erreichbaren Klassifikationsgenauigkeit menschlicher Annotatoren.
- Zu einem Dokument kann oft ein Indikator für die Zuverlässigkeit der Klassifikation berechnet werden. Durch Verschieben der Klassifikationsschwelle für diesen Indikator kann man entweder die Präzision oder die Trefferrate der Klassifikation betonen. Schließlich kann man auch im Vergleich zu diesen Schwellen »unsichere« Dokumente manuell zuordnen und damit das Fehlerrisiko verringern.
- Durch Aktives Lernen lassen sich die Klassifikatoren mit relativ geringem manuellem Aufwand an neue Webseiten anpassen.

- Multimedia-Analyse: Für die Bewertung der Inhalte multimedialer Daten ergeben sich je nach Daten- und Inhaltstyp unterschiedliche Perspektiven:
  - Das Fingerprinting bringt alle Voraussetzungen mit, um ein zentrales Hilfsmittel im Jugendmedienschutz zu werden. Es löst eine Aufgabe, die im Alltag immer wieder auftritt, nämlich die Identifikation bekannten, jugendschutzrelevanten Materials. Die Technologie ist hinreichend weit entwickelt und kommerziell verfügbar.
  - In einer Reihe von Anwendungsbereichen erscheint der Einsatz automatischer Spracherkennung sinnvoll, da die Erkennungsleistung moderner Systeme hoch genug ist, um anschließend mit Methoden aus der Textanalyse Erkenntnisse über den Inhalt des Materials gewinnen zu können. Für Entwickler von Jugendschutz-Lösungen, die selbst keine Expertise im Bereich Spracherkennung haben, bietet es sich an, auf eine gehostete Spracherkennungslösung zurückzugreifen.
  - Daten, die Text in Bildform enthalten, sind für den Jugendmedienschutz oft relevant. Die im wissenschaftlich-industriellen Kontext untersuchten Daten entsprechen zumeist weitgehend den im Jugendmedienschutz relevanten Daten. Der Einsatz optischer Texterkennung lässt sich bei erfolgreichem Umgang mit komplexen Schriftarten und Text vor komplexem Hintergrund als überzeugendes Werkzeug gut vorstellen.
  - In jugendschutzrelevantem Material sind häufig Logos einschlägiger Webseiten eingebettet. Entsprechend erscheint es sehr sinnvoll, Logoerkennung im Jugendmedienschutz einzusetzen, insbesondere da sich die Datenlage ähnlich und zum Teil sogar einfacher darstellt, als im wissenschaftlichen Kontext untersucht.
  - Die Erkennung von Handlungskonzepten aus Videos ist noch nicht hinreichend fortgeschritten, um einen Einsatz für den Jugendmedienschutz realistisch erscheinen zu lassen. Verfahren zur Erkennung von Konzepten in Einzelbildern erscheinen hier zunächst vielversprechender.
- Umfeldanalyse: Die Analyse der Linkstruktur, in die eine Webseite eingebettet ist, birgt möglicherweise wertvolle Zusatzinformationen für die Klassifikation jugendschutzrelevanten Materials. Allerdings sind entsprechende Technologien sehr aufwändig, etwa vergleichbar mit dem Betrieb einer Suchmaschine.

Insgesamt ist somit also eine Reihe nicht nur inhaltlich, sondern auch technisch vielversprechender Verfahren identifiziert. Das folgende Kapitel 5 wendet sich nun der Bewertung ausgewählter Verfahren anhand von Anwendungsszenarien und Daten aus dem Jugendmedienschutz zu.

KAPITEL 5

BEWERTUNG AUSGEWÄHLTER, VERFÜGBARER VERFAHREN

Im vorherigen Kapitel wurden technische Verfahren zur Klassifikation von jugendschutzrelevantem Material aus exemplarischen Datensammlungen hergeleitet und beschrieben. Nun wird es darum gehen, ausgewählte technische Verfahren zu bewerten.

Zunächst werden wir die technischen Verfahren zusammenstellen, die wir in diesem Kapitel durchleuchten wollen, und eine Reihe von kategorieübergreifenden Fragen an diese Verfahren aus Sicht eines Medienwächters stellen, die wir im Zuge der Experimente in diesem Kapitel beantworten wollen (Abschnitt 5.1). Im Anschluss werden wir uns, weiterhin kategorieübergreifend, mit Methoden der Textklassifikation auseinandersetzen und überprüfen, inwieweit diese die redaktionelle Klassifikation unterstützen können (Abschnitt 5.2). Aufbauend auf diesen Ergebnissen untersuchen wir dann speziell angepasste Text- und Multimediamethoden zur Klassifikation von Web-Inhalten für die Unterkategorien Gewalt (Abschnitt 5.3), Extremismus (Abschnitt 5.4) und Selbstgefährdung (Abschnitt 5.5).

Bei der Betrachtung dieser Unterkategorien nehmen wir sowohl qualitative als auch quantitative Bewertungen vor. In den qualitativen Untersuchungen wird an einzelnen Beispielen betrachtet, welche Verfahren sich besonders gut für einzelne Aufgaben eignen. Die vielversprechendsten Techniken werden dann in quantitativen Experimenten auf ihre Effektivität überprüft, das heißt, es werden anerkannte Fehlermaße an manuell vorbewerteten Testdaten gemessen.

In Abschnitt 5.6 geben wir eine abschließende Bewertung der betrachteten Verfahren ab. Zu den am Eingang des Kapitels gestellten Fragen leiten wir eine mögliche Alltagstauglichkeit für den Jugendmedienschutz ab, indem wir insbesondere Faktoren wie Effizienz, Skalierbarkeit, Echtzeitfähigkeit und Wartung mit einbeziehen. Die wichtigsten Erkenntnisse fassen wir dann in im letzten Abschnitt zusammen (Abschnitt 5.7).

Integrationsaspekte, die bei einem Einsatz im Alltag zu beachten sind (etwa notwendige Server-Infrastruktur etc.), werden dann weiter in Kapitel 6 betrachtet.

5.1 Zielsetzung und allgemeines Vorgehen

Um eine Interpretierbarkeit der Ergebnisse zu erleichtern, haben wir eine Liste von Fragen zusammengestellt, zu denen uns die Experimente in diesem Kapitel eine Bewertung ermöglichen sollen. Diese lauten für Textinhalte:

1. Inwiefern kann man durch Methoden der Textklassifikation die manuelle Erstellung und Wartung von Keyword-Listen (Schlüsselwörter) unterstützen oder ersetzen?

2. Ist eine automatische Klassifikation von jugendschutzrelevanten Inhalten auch detailliert, also mit konkreter Altersklassifikation, möglich? Wie verhält es sich konkret bei den Unterkategorien?
3. Können automatisch bestimmte sprachliche Merkmale für Unterkategorien von jugendschutzrelevanten Inhalten gefunden werden? Kann man anhand dieser nahe beieinander liegende Inhalte, zum Beispiel Geschichtsrevisionismus und historische Online-Enzyklopädien, unterscheiden?
4. Kann automatisch ein typisches Themenspektrum in einer Menge verdächtiger Webseiten generiert werden?
5. Kann man anhand des Textes erkennen, inwieweit die dargestellten Bilder jugendschutzrelevant sind?

Zudem wollen wir folgenden Fragen für Multimedia-Objekte (Bild, Ton, Video) nachgehen:

1. Kann man ein neu hochgeladenes, aber bereits bekanntes Video erkennen, auch dann, wenn es leicht verändert wurde?
2. Kann man neu hochgeladene, jugendschutzrelevante Musikstücke automatisch erkennen, auch dann, wenn sie etwa nachgesungen werden?
3. Wie kann man die redaktionelle Klassifikation bei unbekannten Liedern möglichst effizient gestalten?
4. Kann man bereits bekannte jugendschutzrelevante Bilder wiederfinden, auch dann, wenn sie verändert worden sind, zum Beispiel in Collagen?
5. Kann man unbekannte jugendschutzrelevante Bilder aufgrund ihres Inhaltes identifizieren?
6. Kann man bei Spielfilmen mit mehreren, altersdifferenzierten Fassungen eine Zuordnung vornehmen, auch wenn nur einzelne Szenen gezeigt werden?
7. Reicht die Qualität einer automatischen Spracherkennung auf Videos aus, um auf deren Inhalt schließen zu können?
8. Reicht die Qualität einer automatischen Texterkennung auf Videos oder Bildern aus, um auf deren Inhalt schließen zu können?

Wie in Kapitel 4 gesehen, stehen dabei eine ganze Reihe von Techniken zur Verfügung, um sich diesen Herausforderungen zu stellen. Bei den meisten Verfahren lässt sich im Vorfeld aber noch keine allgemeine Bewertung vornehmen – viele Aspekte wie etwa Effektivität oder Laufzeit hängen sehr stark von der Natur der Daten ab. Wir haben daher in diesem Kapitel eine möglichst große Anzahl der Techniken an konkreten jugendschutzrelevanten Daten hinsichtlich ihrer Effektivität überprüft.

Zu folgenden Techniken auf Textebene geben wir eine Bewertung ab:

Keyword Extraktion: Hier werden Keywords (Schlüsselwörter) automatisch aus einem Text extrahiert. Wir nehmen einen Vergleich von manuell und automatisch angelegten Keyword-Listen vor und setzen diesen in Kontext zur Textklassifikation mittels Support Vector Maschinen (Abschnitt 5.2.1).

Unsupervised Topic Clustering: Diese Technik bewirkt die automatische Gruppierung von Wörtern nach Zugehörigkeit und gemeinsamem Auftreten und wird benutzt zur Sortierung und Einordnung von unbekannten Textinhalten (Abschnitt 5.2.2).

Textklassifikation: Dieses Verfahren bewirkt die automatische Einordnung von Dokumenten in eine vordefinierte Hierarchie. In dieser Evaluation führen wir binäre Klassifikation durch, das heißt wir stellen fest, ob ein Dokument zu einem Konzept (z. B. »Gewalt ab 18«) gehört oder nicht (Abschnitte 5.3.6, 5.4.3 und 5.5.6).

Tabelle 5.1
Übersicht der experimentell ausgewerteten Techniken, nach Themen

Verfahren	Allgemein	Gewalt	Extremismus	Selbstgefährdung
Textverfahren				
Topic Clustering	✔			
Keyword Generierung	✔			
Textklassifikation	✔	✔	✔	✔
Multimediaverfahren				
Fingerprinting (Bild)		✔		✔
Fingerprinting (Audio)		✔	✔	✔
Logoerkennung			✔	
Spracherkennung		✔	✔	
Texterkennung				✔
Konzepterkennung (Bild)				✔

Auf Multimedia bewerten wir die folgenden Techniken:

Fingerprinting (Bild): Das Wiedererkennen von bereits klassifizierten Bildern und die Robustheit gegenüber von Nutzern veränderten Daten überprüfen wir anhand von Zusammenschnitten eines Horrorfilms (Abschnitt 5.3.2.3) und anhand eines typischen Anorexie-Bildes (Abschnitt 5.20).

Fingerprinting (Audio): Das Wiedererkennen von bereits klassifizierten Audiosegmenten in Horrorfilmen (Abschnitt 5.3.2.2) sowie das Zuordnen von rechtsradikalen Liedern (Abschnitt 5.4.5), beziehungsweise die automatische Gruppierung von Liedern bei selbstgefährdendem Inhalt (Abschnitt 5.5.2) wird hier untersucht.

Logoerkennung: Das Wiedererkennen von strafrechtlichen Symbolen wie etwa dem Banner der Blood & Honour Vereinigung sowie des Hakenkreuzes untersuchen wir in Abschnitt 5.4.7.

Spracherkennung: Unter Spracherkennung wird hier die Verschriftlichung von gesprochenem Wort verstanden. Wir untersuchen sie in »Let's Play« Videos von Gewaltspielen (Abschnitt 5.3.4) und in Videos von Holocaust-Leugnern (Abschnitt 5.4.8).

Texterkennung: Die Texterkennung lokalisiert und identifiziert Wörter, die isoliert auf neutralem Hintergrund sowie über Bilder gemalt auftreten können. Wir untersuchen die Effektivität in Videos mit Selbstverletzungen (Abschnitt 5.5.3).

Konzepterkennung (Bild): Die Konzepterkennung teilt ein Bild aufgrund des Dargestellten in zwei Klassen. Wir untersuchen, inwieweit eine Unterscheidung von Selbstverletzungen und Tattoos möglich ist (Abschnitt 5.5.4).

Eine Übersicht aller untersuchten Techniken mit den jeweils betrachteten thematischen Bereichen des Jugendschutzes findet sich in Tabelle 5.1.

Insbesondere für Multimediatechniken sind keine kategorieübergreifenden Experimente vorgesehen, da es kaum entsprechende Merkmale gibt, in der eine gesamtheitliche Betrachtung sinnvoll erscheint. Viele in den Unterkategorien gewonnenen Erkenntnisse lassen sich aber ohne Weiteres auch auf die anderen Bereiche übertragen; so funktioniert ein Fingerprinting auf Liedern mit rechtsradikalem Inhalt auch auf Liedern, die die Selbstverletzung verherrlichen. Andere Techniken wie etwa die Konzepterkennung auf Ritzbildern hingegen sind zu speziell, um daraus generalisierende Aussagen ableiten zu können. Wir werden diesbezüglich in den jeweiligen Unterabschnitten die mögliche Übertragbarkeit auf andere Bereiche erörtern. Einige Experimente besitzen aufgrund der

eingeschränkten Datenlage eher exemplarischen Charakter, da nur wenige Daten aus den entsprechenden Kategorien zur Verfügung standen. Diese Experimente sind eher als Nachweis prinzipieller Machbarkeit anzusehen. Die Größe der verwendeten Datensätze ist für jedes Experiment angegeben.

5.2 Kategorieübergreifende Textanalyse-Verfahren

In diesem Abschnitt betrachten wir Textanalyse-Verfahren auf als jugendschutzrelevant markierten Textseiten. Dazu wurde auf bestehende Listen mit Webseiten, die jugendschutzrelevanten Text beinhalten, zurückgegriffen und die Inhalte heruntergeladen und entsprechend vorbereitet (das heißt, der dargestellte Text von anderen Inhalten separiert und getrennt betrachtet).

Ausgehend von diesem Material wollen wir nun überprüfen,

- inwieweit sich eine manuell generierte Keyword-Liste durch automatische Verfahren ergänzen lässt (Abschnitt 5.2.1)
- und ob man mit automatischen Verfahren zur Sortierung und Einordnung Themenschwerpunkte finden kann, die die redaktionelle Klassifikation erleichtern (Abschnitt 5.2.2).

Der Abschnitt endet mit einer Zusammenfassung.

5.2.1 Keyword-Extraktion

Eine weit verbreitete Methode in der Dokumentklassifikation ist das Verwenden von manuell erstellten und gewarteten Keyword-Listen. Diese gibt es in Form von Blacklists (unerwünschte Keywords) oder Whitelists (erwünschte Keywords), beziehungsweise als gewichtete Liste (Plus- bzw. Minuspunkte für bestimmte Keywords). Liegt eine solche gewichtete Liste vor, so kann man einem Text entsprechend der in ihm vorkommenden Wörter eine Gesamtbewertung (*Dokument-Score*) geben mit dem Ziel, dass jugendschutzrelevante Inhalte einen möglichst hohen Wert erhalten und harmlose Inhalte einen möglichst niedrigen. Je besser die Keyword-Liste, desto besser ist damit eine saubere Trennung der Inhalte mittels eines Schwellwertes möglich.

Wir wollen in diesem Unterabschnitt untersuchen, ob man durch automatisch erzeugte Keywords jugendschutzrelevante Dokumente charakterisieren und identifizieren kann. Dazu vergleichen wir eine manuell angelegte Liste mit automatischen Methoden. Konkret vergleichen wir drei Szenarios:

1. *keywords-manual*: Mit einer manuell angelegten Keyword-Liste werden die Dokument-Scores erzeugt. Diese Liste besteht aus ein paar tausend händisch ausgewählten und gewichteten Begriffen. So bekommt beispielsweise das Wort »Selbsthilfegruppe« ein negatives Gewicht, da es eher gegen Jugendschutzrelevanz spricht, wobei »adults only« ein positives Gewicht bekommt. Diese Gewichte werden bei jedem Treffer des dazugehörigen Begriffs im Text entsprechend aufsummiert. Daraus wird dann der Score des gesamten Dokuments ermittelt. Diese Scores sind in Abbildung 5.1(a) dargestellt.
2. *keywords-linearsvm*: In diesem Szenario wird eine gewichtete Keyword-Liste automatisch durch eine Support Vektor Maschine (vgl. Abschnitt 4.4.2.5) erstellt.[1] Analog zur ersten Variante (*keywords-manual*) werden diese Begriffe in den zu testenden Dokumenten gesucht

[1] Eine Support Vektor Maschine (SVM) lässt sich mit einer linearen Kernfunktion zur automatischen Bestimmung sinnvoller Keywords für eine Zielkategorie verwenden. Hierzu wird zunächst ein Klassifikationsmodell erstellt. Multipliziert mit dem Gewicht der Supportvektoren werden die Werte der dazugehörigen Merkmale aufsummiert. So erhält jedes Merk-

und ihre Gewichte aufsummiert, um den Score des jeweiligen Dokuments zu ermitteln. Die Ergebnisse sind in Abbildung 5.1(b) dargestellt.

3. *rbfsvm*: Im Unterschied zu den beiden ersten Varianten wird in diesem Szenario nicht mit einer Liste von gewichteten Keywords gearbeitet, sondern ein nicht-lineares Modell, die Support Vektor Maschine mit radialer Basisfunktion (vgl. Abschnitt 4.4.2.5), angewendet. Das heißt, dass nicht mehr nur das Auftreten oder Fehlen eines bestimmten Wortes bewertet werden kann, sondern auch komplexere Regeln und Kombinationen möglich sind, zum Beispiel eine Wertung nur dann erfolgt, wenn zwei Wörter in Kombination auftreten oder ein Wort nur dann ein Gewicht erhält, wenn ein bestimmtes anderes *nicht* auftritt.[2] Somit stellt es eine Erweiterung des klassischen Keyword-Ansatzes dar (Abbildung 5.2).

Die uns vorliegende, manuell erstellte Keyword-Liste beinhaltet zu einem sehr großen Teil pornographisches Material (etwa 95% aller »ab 18« markierten Inhalte). Da die Identifikation dieses Themenbereiches nicht Teil der Studie ist, wurden diese Inhalte soweit möglich nicht mitgeladen. Für das folgende Modell haben wir 3 087 Webseiten, die mit »ab 6« markiert waren, gegen 1 154 Webseiten »ab 18« gehalten. Die Dokument-Scores werden einmal für Dokumente »ab 6« und einmal für Dokumente »ab 18« erzeugt und dargestellt.

In den Abbildungen 5.1 und 5.2 werden die Häufigkeitsverteilungen der Testwebseiten über die berechneten Scores dargestellt. Zur Verdeutlichung dieser Abbildung ist es hilfreich, sich eine vertikale Trennlinie zu denken. Alle Dokumente auf der linken Seite dieser Linie werden als nicht jugendschutzrelevant erkannt, alle Dokumente auf der rechten Seite entsprechend als relevant. Der Bereich unter der blauen Kurve auf der rechten Seite der gedachten Linie stellt die Menge der fälschlicherweise als relevant ermittelten Dokumente dar. Ist dieser Bereich groß, gibt es also ein hohes Overblocking. Analog dazu steht der Bereich unter der roten Kurve auf der linken Seite dieser Linie für das Underblocking.

Es zeigt sich deutlich, dass die manuelle Keyword-Liste eine klare Abgrenzung für besonders eindeutige Werte (ab einem Score von $>= 300$) bietet, da dort viele der einschlägigen Keywords enthalten sind (Abbildung 5.1(a)). Allerdings sind die Klassen in einem Großteil des sonstigen Spektrums kaum zu unterscheiden. Auffallend ist auch, dass sich in der Gegend um den Score null viele Dokumente beider Klassen befinden, da sie einfach kein einziges der auf den Listen befindlichen Keywords enthalten.

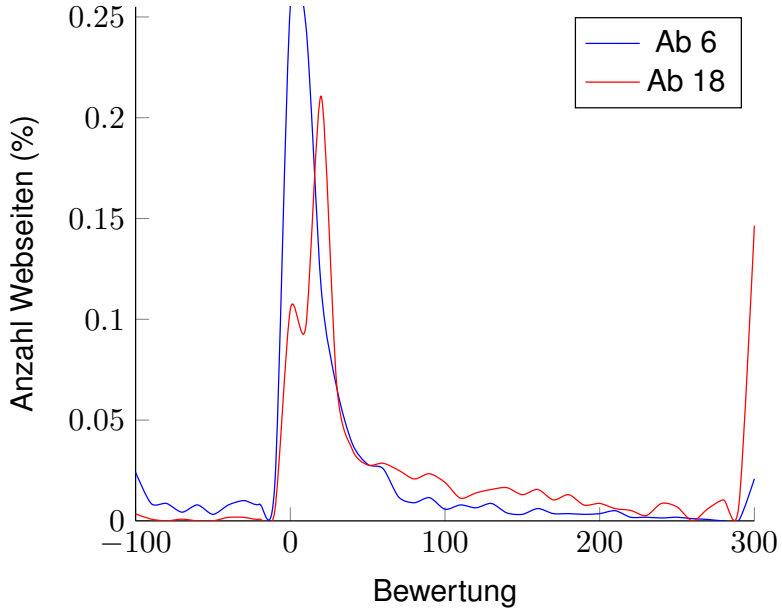
Extrem falsch klassifiziert, also mit besonders hoher Score trotz Eignung »ab 6«, werden hier exemplarisch

- Bibelsammlungen, die besonders drastische Verse etwa aus dem alten Testament zitieren,
- englischsprachige Nachrichtenseiten, in denen vermeintlich eindeutige Wörter in einem anderen Kontext stehen (etwa: »look at his *facial* expression« im US-Wahlkampf) und
- Beratungsseiten etwa bei Essstörungen, die schon deshalb viele relevante Wörter enthalten, um von Suchmaschinen gefunden zu werden.

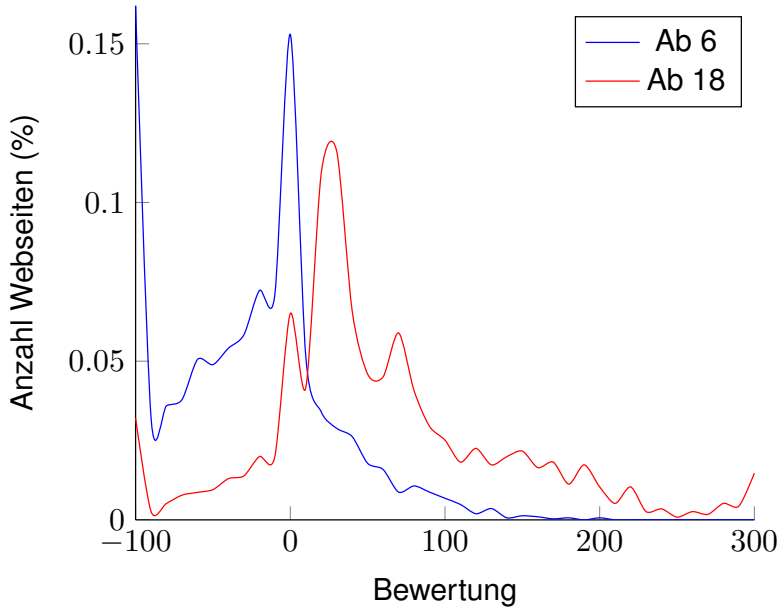
Umgekehrt werden Webseiten »ab 18« zum Beispiel dann falsch klassifiziert, wenn

mal, in diesem Fall einfacherweise die im Text enthaltenen Wörter, automatisch ein positives oder negatives Gewicht. Um eine Vermischung von Trainings- und Testdokumenten zu vermeiden, wurde eine Kreuzvalidierung (siehe Abschnitt 4.2.3) durchgeführt. Weiterhin wurde darauf geachtet, dass die Webseiten derselben Webpräsenz immer zusammenhängend im Trainings- oder im Testbestand vorkommen, da es ansonsten zu Verzerrungen im Testergebnis kommen könnte.

[2] Hier wird durch eine Support Vektor Maschine mit einer radialen Basisfunktion (RBF) als Kernfunktion ein nicht-lineares Klassifikationsmodell erzeugt. Die Dokument-Scores werden direkt aus der Klassifikationsausgabe der SVM generiert und entsprechend skaliert. Auch hier wurde mittels Kreuzvalidierung auf eine strikte Trennung von Trainings- und Testdaten geachtet. Dieses Verfahren wird auch in den Experimenten der folgenden Abschnitte angewandt, da es in der Regel bessere Ergebnisse liefert als Keyword-Listen.



(a) *keywords-manual*: Verteilung der Textbewertung bei manuell erzeugten und gewichteten Keywords.



(b) *keywords-linearsvm*: Verteilung der Textbewertung mittels einer automatisch erzeugten Liste gewichteter Keywords.

Abbildung 5.1
Trennschärfe von gewichteten Keywords für »ab 6« und »ab 18« Inhalten. Die blaue Linie zeigt jeweils die Häufigkeitsverteilung der »ab 6«-Inhalte über die ermittelten Dokument-Scores. Die rote Linie zeigt entsprechend die Verteilung der »ab 18«-Inhalte.

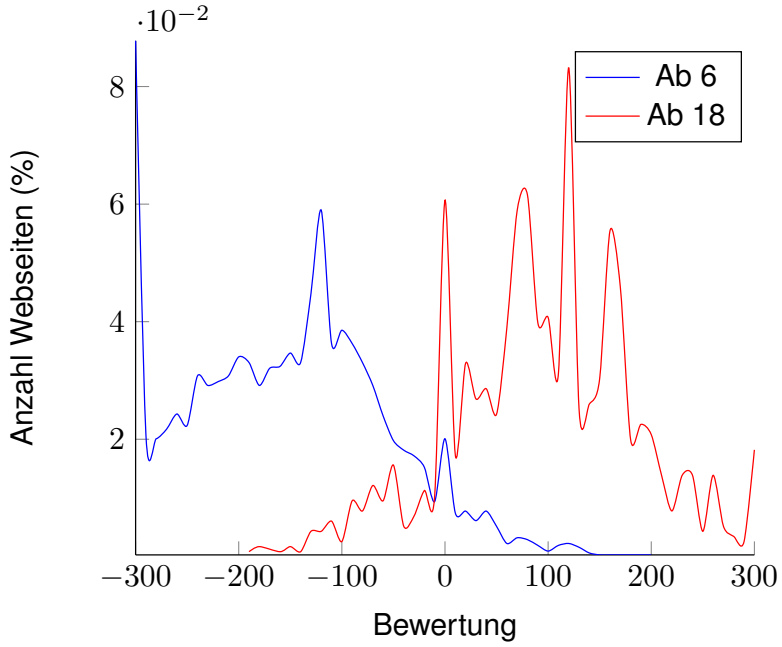


Abbildung 5.2
rbfsvm: Verteilung der Textbewertung für »ab 6« und »ab 18« Inhalte, durch eine SVM mit RBF-Kernfunktion. Das dargestellte Experiment erzielt bei einer Trennung bei null ein F-Maß von 81,8 % für die Zielkategorie »ab 18« mit einer Präzision von 90,9 % bei einer Trefferquote von 74,4 %. Diese hohe Präzision zeigt sich im Diagramm an dem relativ kleinen Bereich unterhalb der blauen Linie auf der rechten Seite einer gedachten vertikalen Trennlinie bei null.

- sie eine Reihe von scheinbar harmlosen Wörtern enthalten, etwa bei christlichen Fundamentalisten, die anhand von Bildern mit durch den Ausschabungsvorgang verstümmelten, abgetriebenen Föten gegen diese Praxis demonstrieren wollen,
- der jugendschutzrelevante Inhalt nicht ausschließlich über Schlagwörter, sondern verstärkt über den Kontext zu identifizieren ist, etwa bei Textdokumenten mit Holocaustleugnung. Zwar wird beispielsweise »Holocaust« erfasst, aber eben auch positiv besetzte Wörter wie etwa »Zeitung« und »government«, und
- für den Kontext eher ungewöhnliches Vokabular verwendet wird, wie etwa bei Liedtexten (Beispiel: »Verrottete Gliedmaßen beginnen zu gären, Körper entzwei gerissen / Haut abgezogen, um das Muskelgewebe bloßzustellen«).

Im Gegensatz zu manuell erstellten Keywords können mit der aus dem Ansatz *keywords-linearsvm* automatisch erzeugten Liste die Klassen schon besser getrennt werden (Abbildung 5.1(b)). Man kann klar sagen, dass Dokumente aus der »ab 18« Klasse tendenziell höhere Scores haben als die »ab 6«-Dokumente. Ein Großteil der »ab 6«-Dokumente hat nun sogar einen negativen Score, während die »ab 18«-Dokumente klar positive Scores haben. Es lässt sich daraus insbesondere schließen, dass die Abdeckung mit Keywords hier höher ist. Automatisch generierte Keyword-Listen enthalten also deutlich mehr Keywords, die in den Dokumenten tatsächlich vorkommen, und haben dabei den Vorteil, einfach aktualisiert und gewartet werden zu können.

Die mit Abstand beste Performanz liefert der Ansatz *rbfsvm* (Abbildung 5.2). Dies ist nicht verwunderlich, da hier keine explizite Keyword-Liste generiert wird, sondern ein auf strikte Diskriminierung der beiden Klassen ausgelegtes Klassifikationsmodell zum Einsatz kommt.

Zwar ist im Gegensatz zum Ansatz *keywords-manual* vor der Anwendung zunächst die Berechnung des Modells, das sogenannte Training, nötig. Dafür entfällt der redaktionelle Zeitaufwand zur Erstel-

lung der Keyword-Listen. Stattdessen wird nur eine Liste nach Jugendschutzrelevanz eingestufter Webadressen für das automatische Training benötigt. Auch die Rechenzeiten für das hier beschriebene Experiment sind absolut unproblematisch. Die Trainingszeit für dieses Experiment betrug (über alle Teilmodelle der Kreuzvalidierung aufsummiert) auf einem handelsüblichen Desktop-PC lediglich 78 Sekunden bei über 4 000 Beispielen. Die Anwendung der Modelle auf alle Beispiele betrug insgesamt lediglich 19 Sekunden, durchschnittlich also weniger als 5 Millisekunden pro Dokument.

Dabei ist zu erwähnen, dass diese Werte stark von der verwendeten Hardware, den Verfahren, ihrer Implementierung und weiteren Faktoren abhängen, so dass diese Zeitangaben eher zur Verdeutlichung der Größenordnung und für den internen Vergleich angegeben werden und sich nicht direkt auf andere Anwendungsbedingungen übertragen lassen. Abgesehen davon gilt generell, dass die Laufzeit des Trainings üblicherweise exponentiell in der Anzahl der Dokumente steigt. Für die meisten Verfahren existieren allerdings sehr effiziente Implementierungen, so dass auch Größenordnungen von mehreren hunderttausend Trainingsbeispielen kein Problem darstellen. Da sich die Dokumente bei der Anwendung der Klassifikation unabhängig voneinander verarbeiten lassen, steigt die Laufzeit hier nur linear. Das sequentielle Klassifizieren von 1 000 Webseiten dauert also in etwa 1 000 mal so lange wie die Klassifikation einer einzelnen.

Die vollständige Vorverarbeitungskette inklusive HTML-Parsing, Merkmalsextraktion und Umwandlung der Textinhalte in einen reellwertigen Vektor benötigt bei diesem Experiment insgesamt 67 Sekunden zusätzlich, also knapp 17 Millisekunden pro Dokument. Im laufenden Betrieb fallen somit etwa 22 Millisekunden für die Klassifikation eines einzelnen HTML-Dokuments an. Diese Werte zeigen, dass derartige Verfahren echtzeitfähig und durch ihren Einsatz keine Laufzeitprobleme zu erwarten sind.

Zusammenfassend lässt sich klar feststellen, dass Ansätze, die automatisch aus aktuellen Dokumenten Keywords und Modelle generieren, manuellen Ansätzen gegenüber deutliche Vorteile bieten. Mit automatisch generierten Keyword-Listen kann zielgerichteter nach neuen Inhalten gesucht werden und ihre Bewertung ist trennschärfer. Möchte man nicht nur für Suchportale geeignete Wörter vorliegen haben, sondern einen bereits vorliegenden Text beurteilen, ist der Einsatz von komplexeren Verfahren wesentlich zuverlässiger, allerdings sind die von den Verfahren erzeugten Regeln für einen menschlichen Betrachter nicht mehr ohne Weiteres verständlich darstellbar.

5.2.2 Themenmodelle

Mit Hilfe von Themenmodellen (siehe Abschnitt 4.4.3.3) lassen sich Dokumente automatisch thematisch charakterisieren. Dabei erfolgt nicht nur die Zuordnung eines neuen, bisher ungesehenen Dokuments in das Modell komplett automatisch, sondern auch die Erstellung des Modells, welche keinerlei manuelle Annotation erfordert. Diese Erstellung basiert auf der Analyse des gemeinsamen Auftretens von Wörtern. So entstehen Wortcluster mit Wörtern, welche oft zusammen in einem Dokument enthalten sind und die als Themen aufgefasst werden können. Auf ein neues Dokument angewendet, erhält man ein Themenprofil, also eine gewichtete Themenverteilung. Die Anzahl der Themenmodelle wird dabei vom Benutzer vorgegeben.

Die im letztem Abschnitt beschriebenen Inhalte »ab 18« wurden automatisch in 100 Themen geclustert. Die Ergebnisse sind in Tabelle 5.2 dargestellt. Hier ist klar zu erkennen, dass die Wörter nach Sprachen getrennt sind. Es gibt deutsche, englische, französische, italienische und türkische Themen. Aber auch eine gewichtete Zuordnung zu den möglichen Themen, zum Beispiel Sport, Autos, Sex, zweiter Weltkrieg und anderen, ist zu erkennen. Zu beachten ist hierbei, dass die Zuordnung implizit geschieht und besagt, dass diese Wörter oft in einem Kontext verwendet werden. Auch

Tabelle 5.2
Einige beispielhaft ausgewählte Themen aus 32 028 Webseiten der Kategorie »Allgemein, ab 18« bei 100 Themen ohne Sprachfilterung. Die einzelnen Themen besitzen keine konkreten Titel und sind daher durchnummeriert. Das Gewicht stellt den Einfluss, also die relative Größe des Themas im zugrundeliegenden Korpus dar.

Nr.	Gewicht	Stärkste Wörter
0	0,000028	gun rifle guns shooting rifles airsoft tactical firearms hunting …
2	0,000050	quot href www font body size target flickr style blank htm …
5	0,000016	car auto cars ford bmw audi mercedes honda porsche …
6	0,000018	sport nachrichten video fussball auto news wetter bild …
7	0,000041	music sat rock radio album tickets love band live dance …
10	0,000003	bir için son tüm ile yeni satılık göster istanbul çok video …
11	0,000062	games game play girl star world puzzle free funny super …
16	0,000018	seeds cannabis marijuana hemp buy legal feminized herbal …
17	0,000020	mobile sms handy iphone ringtones nokia free phone …
27	0,000021	del italia sito della pagina dei home giochi news web video …
28	0,000023	informationen fitness www thema gesundheit hilfe apotheke …
29	0,000008	waffen shop warenkorb messer eur walther softair produkt …
34	0,000013	npd deutschland deutschen deutsche euro politik artikel …
35	0,000018	sexe sur camping les site une rechercher rencontre pour …
36	0,000043	gay girls porn video free videos nude teen movies sexy sex …
39	0,000020	dvd blu ray film horror dvds angola movie season movies …
45	0,000026	sex erotik frauen video chat live geile virgin potsdam girls …
59	0,000030	big hot black sex pussy cock porn ass mature girl gets teen …
61	0,000016	german vol hitler military holocaust army jewish jews gas …
92	0,000028	games spiele game poker casino download spiel play flash …
95	0,000032	sex phone xxx adult porn free chat live girls dating toys …

wenn man als Mensch zumeist anhand weniger Wörter eine explizite Zuordnung vornehmen kann, unterscheiden sich bei diesem Ansatz manche Wörter sehr stark auf einer semantischen Ebene.

In Tabelle 5.3 wird eine Sprachfilterung vorgeschaltet und es werden nur deutsche Inhalte zugelassen. Um einen höheren Detailgrad der Themen zu erreichen, wird zusätzlich die Anzahl der Themen auf 200 verdoppelt. Dieser Wert ist hier manuell gewählt. Über die Minimierung der mittleren Likelihood (vgl. Abschnitt 4.4.2.2), der sogenannten Perplexität, lässt sich allerdings auch automatisch die Themenanzahl ermitteln, bei der die Themen am klarsten voneinander getrennt werden. Die Annahme ist dabei, dass dieser Wert der Anzahl der tatsächlich im Korpus enthaltenen Themen entspricht. Je nach dem, zu welchem Zweck die Themenmodelle eingesetzt werden, kann es aber durchaus sinnvoller sein, die Anzahl fest einzustellen.

Gibt es schon eine Reihe von Dokumenten mit annotierten Klassen, so kann man mit Hilfe von Klassen-Themenmodellen (vgl. Abschnitt 4.4.3.3) Themen erzeugen, welche stark an den existierenden Klassen orientiert sind. Hierdurch werden die Themen stärker fokussiert und können eher im Sinne der Klassifikation interpretiert werden.

Es ergibt sich also, dass regelmäßig erzeugte Themenmodelle für eine automatische thematische Einordnung auch neuer Themen geeignet sind. Sie können auch herangezogen werden, um sich einen Überblick über die Hauptinhalte einer großen Datenmenge zu verschaffen.

Tabelle 5.3
Einige beispielhaft ausgewählte Themen aus 32 028 Webseiten der Kategorie »Allgemein, ab 18« bei 200 Themen gefiltert nach deutscher Sprache. Auf diese Weise wird eine stärker inhaltliche statt einer sprachlichen Trennung der Themen erreicht. Die einzelnen Themen besitzen keine konkreten Titel und sind daher durchnummeriert. Das Gewicht stellt den Einfluss, also die relative Größe des Themas im zugrundeliegenden Korpus dar.

Nr.	Gewicht	Stärkste Wörter
4	0,000016	bravo sommer foto spezial lovestory tipps frage woche stars …
13	0,000018	handy lernen kostenlos klingeltöne real klingelton logo musik …
14	0,000007	absinthe geschmack wasser verte anis trinkempfehlung …
15	0,000025	waffen warenkorb shop walther eur softair produkt kal zippo …
19	0,000058	spiele games game spiel browsergames download kostenlos …
21	0,000059	handy iphone tarife handys apple dsl samsung mobile prepaid …
31	0,000045	npd demonstration widerstand nationale euro aktivisten …
44	0,000021	shisha headshop growshop bong zigarette tabak bongs shop …
45	0,000015	nackte jungs boys alkohol alkoholiker hilfe videos selbsthilfe …
47	0,000029	dvd blu ray film kaufen filme dvds uncut edition amazon…
50	0,000100	sex erotik girls geile frauen videos porno bilder sexy …
54	0,000023	wein weine stern wine rotwein weingut sekt warenkorb
59	0,000016	cam webcam live chat linux cams livecams livecamvoting …
65	0,000021	battlefield looki united spiele spieler panzer bundeswehr …
77	0,000021	bibel kirche hochzeit jesus christus gott papst sunrise dirty
79	0,000029	auto opel neuwagen gebrauchtwagen skoda reimport …

5.2.3 Zusammenfassung

Wir haben in diesem Abschnitt einen ersten Eindruck gewonnen, dass automatische Verfahren bezüglich ihrer Trennschärfe und dem Aktualisierungaufwand gegenüber einer manuellen Keyword-Liste erhebliche Vorteile bieten. Weiterhin konnte gezeigt werden, dass mit Hilfe von Themenmodellen entscheidende Wörter zusammengestellt werden können, was insbesondere bei fremdsprachigem Inhalt eine Vorsortierung ermöglicht. Generell kann gesagt werden, dass diese automatischen Verfahren leichter zu warten und besser auf aktuelle Dokumente anpassbar sind als entsprechende manuelle Ansätze. Zu beachten ist insbesondere bei Themenmodellen, dass die Anzahl der Themen als Parameter vorgegeben und ein optimaler Wert hier durch Experimente bestimmt werden muss. Neben der inhaltlichen Heterogenität des Textmaterials ist dabei auch die Anzahl der vorhandenen Dokumente zu berücksichtigen. Des Weiteren kann auch ein Blick auf die generierten Themen selbst Aufschluss geben, ob Themen zu allgemein oder zu speziell abgedeckt sind.

Eine Analyse des fehlklassifizierten Materials legt nahe, dass viele speziellere Seiten mit kategorieübergreifenden Modellen nur sehr schlecht erfasst werden beziehungsweise zu Fehlklassifikation führen können. So sind für manche Themen wie etwa der Selbstgefährdung häufig Aussteiger- und Beratungsseiten vorzufinden, die ähnliches Vokabular verwenden, aber eine andere Intention zugrunde liegen haben. Auch bei Extremismus unterscheiden sich die Webseiten zum Teil nur bei schärferem Hinsehen. Gerade im Bereich Gewalt sind es häufig die Multimedia-Inhalte, die vornehmlich zur Jugendschutzrelevanz beitragen.

Die Ergebnisse sind insgesamt also ermutigend, für eine umfassende Bewertung werden wir in den folgenden Abschnitten aber genauer auf Einzelaspekte und Herausforderungen in den jeweiligen Unterkategorien eingehen müssen. Wir werden uns daher im Folgenden mit den Unterkategorien Gewalt, Extremismus und Selbstgefährdung beschäftigen und gezielt die Methoden an diese Inhalte anpassen.

5.3 Betrachtung von Verfahren zur Klassifikation im Bereich Gewalt

Im Zuge einer themenbezogenen Klassifikation widmen wir uns zunächst der Betrachtung von technischen Verfahren anhand von Material aus dem Bereich Gewalt. Wie bereits in Abschnitt 3.2.3 beschrieben, gliedern sich die Inhalte in die Bereiche »real«, »fiktiv«, »Computerspiele« und »Musik«. Die visuellen Inhalte standen bei den von uns betrachteten Webseiten im Vordergrund. Daher werden wir uns in diesem Abschnitt zwar vornehmlich mit technischen Verfahren aus dem Bereich der Multimedia-Analyse beschäftigen, aber gleichzeitig auch untersuchen, inwieweit der textliche Inhalt für eine Klassifikation bereits ausreicht.

Für die Bewertung von technischen Verfahren im Bereich Gewalt gehen wir wie folgt vor:

- Zunächst führen wir einen qualitativen Vergleich von realer und fiktiver Gewalt im Video durch und sehen uns anhand von zwei Videos, die beide drastische Gewalt gegen Wehrlose beinhalten, die Möglichkeit und Grenzen von technischen Verfahren an (Abschnitt 5.3.1).
- Dann werden wir anhand des in Deutschland in der ungeschnittenen Fassung indizierten Films »Braindead« quantitativ die Wirkungsweise der Verfahren testen und evaluieren (Abschnitt 5.3.2).
- In einem zweiten großen Block stellen wir uns dann dem Thema Gewaltspiele, indem wir uns wiederum zunächst qualitativ zwei Videos aus dem Genre »Ego-Shooter« widmen und die technischen Verfahren gegeneinander halten (Abschnitt 5.3.3).
- Für den speziellen Unterbereich der nutzergenerierten »Let's Play« Videos untersuchen wir dann, inwieweit automatische Spracherkennung für eine Bewertung zielführend sein kann (Abschnitt 5.3.4).
- Dann werden wir anhand von Webseiten aus dem Bereich der Robbenjagd anhand des Textinhaltes untersuchen, inwieweit dieser auf die Grausamkeit der dargestellten Bilder Rückschlüsse erlaubt (Abschnitt 5.3.5).
- Schließlich werden wir untersuchen, wie exakt man Webseiten anhand textueller Klassifikation altersgenau klassifizieren kann (Abschnitt 5.3.6).

Der Abschnitt endet mit einer Zusammenfassung für den Themenbereich Gewalt.

5.3.1 Am Beispiel drastische Gewalt

Betrachtet wurden zwei Videos, die drastische Gewalt gegen Wehrlose zum Inhalt haben. Das erste ist ein Zusammenschnitt von einzelnen Videos, der durch das Logo einer Webseite kurz eingeleitet und mit harter Rockmusik unterlegt ist. Gezeigt werden mehrere, als real einzuschätzende Szenen extremer Schlägereien von zum Teil mit Schlagwaffen bewehrten Einzelnen oder Gruppen, die auf wehrlose, auf dem Boden liegende Personen einprügeln (im Folgenden *Prügelvideo*). Das zweite Video ist eine ungeschnittene Fassung eines älteren Zombiefilms, also eines fiktiven Spielfilms (im Folgenden *Zombievideo*). Exemplarisch sollen hier einige Szenen herausgegriffen werden, in der eine Gruppe nicht »infizierter« Menschen portraitiert wird, die nach der anfänglichen Freude über ihr Überleben in eine Mischung aus Langeweile und Dekadenz verfallen. Die untoten Gegner sind zwar potenziell tödliche Gegner, werden aber in dem Film als extrem behäbig dargestellt. In einigen Szenen, mit fröhlicher Musik unterlegt, wird durch eine einfallende Rockergruppe kaum eine Möglichkeit ausgelassen, die Zombies anzugreifen (etwa mittels Schusswaffen) oder zu verhöhnen (etwa mittels einer Torte im Gesicht). Die Folgen dieser Angriffe, darunter zum Beispiel Schüsse in den Kopf, werden in vollem Umfang gezeigt.

Unabhängig davon, dass hier eine andere Intention der »Autoren« angenommen werden darf, sind die Videos auch aus technischer Sicht sehr unterschiedlich. Das Zombievideo ist einer der bekannteren Vertreter seines Genres, zu dem mehrere Schnittfassungen existieren. In Deutschland sind einige davon indiziert oder sogar beschlagnahmt. Es ist ohne Weiteres möglich, auch aus kürzeren Schnittfassungen mittels Video-Fingerprinting eindeutige Szenen wiederzufinden. Inwieweit aber die Szenen des Prügelvideos wiedererkannt werden können, hängt stark davon ab, ob die einzelnen Videos bereits bekannt sind; in diesem Fall reicht es auch, wenn nur eines davon im Vorfeld bereits als jugendschutzrelevant eingestuft wurde. Die Qualität der Videos ist zum Teil denkbar schlecht, was darauf hindeutet, dass das Format mehrfach geändert worden ist. Wenn auch für einen menschlichen Betrachter eindeutig das gleiche Material zugrunde liegt, muss die Robustheit der Fingerprints besonders gewährleistet sein, das heißt, die Verfahren müssen auch abgeschnittene, skalierte oder sonst leicht verfremdete Szenen dennoch eindeutig wiedererkennen können.

Auch Audio-Fingerprinting für das Zombievideo ist prinzipiell denkbar, zu beachten ist aber, dass hier eine englische Fassung zugrunde lag. Zu prüfen wäre, inwieweit die Stellen des Films, in denen nicht gesprochen wird, groß und zahlreich genug sind, damit auch die Erkennung einer synchronisierten Fassung erfolgen kann.

Denkbar wäre auch eine automatische Spracherkennung auf dem Material, um zum Beispiel einige charakteristische Sätze wiederzufinden. Da in den entscheidenden Szenen allerdings sowohl Musik als auch ein hoher Anteil von Hintergrundmusik vorliegt und es sich im die englische Fassung des Films handelt, ist ein erfolgreicher Einsatz dieser Technologie hier eher fraglich. Bei dem Prügelvideo ist die Tonspur ohne Aussagekraft, da die Musik hineingeschnitten wurde und, abgesehen von einer aggressiven Grundstimmung, ansonsten in keiner inhaltlichen Relation zu den bewegten Bildern steht.

Inwieweit Konzeptdetektoren auf diesem Material ein vielversprechender Ansatz sind, erscheint fraglich. Im Prügelvideo sind manche Szenen mit einer Schwarz-Weiß-Überwachungskamera gefilmt, andere mit wackeliger Handkamera beziehungsweise Handykamera. Einige Angriffe sind stark verkleinert, andere wiederum, etwa bei Hinterhofkämpfen mit »sportlichem« Anspruch, sind in Großaufnahme, da der Filmende keine Übergriffe auf seine eigene Person befürchten musste. Das heißt aber auch, dass sowohl von der Farbgebung, von der Kameraführung oder vom Detailgrad der Objekte keine Kriterien für einen Detektor direkt ableitbar wären. Ähnliches gilt für das Zombievideo, wo durch den Grad der Professionalität zwar ein »besseres«, das heißt homogeneres Bildmaterial vorliegt, wo aber eine Übertragung auf anderes Filmmaterial ebenfalls ernüchternd sein kann. Zum einen spielen manche Horrorspielfilme ja gerade bewusst mit dem Eindruck eines Dokumentarfilms, etwa durch absichtlich verwackelte Aufnahmen, so dass auch hier von erschwerten Bedingungen ausgegangen werden kann. Zum anderen erscheint das Vorbereiten eines solchen Detektors als sehr mühevoll, da beispielsweise manche Szenen, wie etwa das Werfen einer Torte, ein bewusstes Spiel mit harmlos wirkender Metaphorik darstellt.

Vielversprechender wirkt hier der Einsatz einer Logodetektion. Im Prügelvideo wird vor dem Zusammenschnitt das Logo einer Webseite eingeblendet, dessen Aufmachung darauf hindeutet, dass sich deren Betreiber auf das Sammeln von derartigem Material spezialisiert haben. Beim Zombievideo dürfte das Titelbild zu Beginn einen wichtigen Hinweis liefern, der aber bei gleichzeitigem, wesentlich mächtigerem Einsatz von Video-Fingerprinting in diesem Fall als eher untergeordnet zu sehen ist. Hinzu kommt, dass manche Zusammenschnitte nur eine Auswahl von Szenen bereithalten, in denen der Titel nicht unbedingt angezeigt wird.

Auf Seiten der optischen Texterkennung kann im Prügelvideo eine Art Wasserzeichen erkannt werden, da bei manchen Szenen die ursprüngliche URL angegeben wird. Die URL wird in weißen Buchstaben über das Video gelegt und bewegt sich diagonal, was die Texterkennung zwar erschwert,

aber immer noch realistisch erscheinen lässt. Im Zombiefilm können zu Anfang und gegen Ende die Namen der beteiligten Personen automatisch erfasst werden. Insbesondere der Regisseur, der sehr viele vergleichbare Filme gedreht hat, dürfte hier eine interessante Wissensquelle darstellen.

5.3.1.1 Unmittelbares Erkennen von Gewalt

Wir haben an den beiden vorangegangenen Beispielen gesehen, wie unterschiedlich schon zwei Videos zu scheinbar ähnlicher Thematik sein können. Es stellt sich die Frage, ob ein Computerverfahren in der Lage ist, ein grausames Bild von einem harmlosen zu unterscheiden.

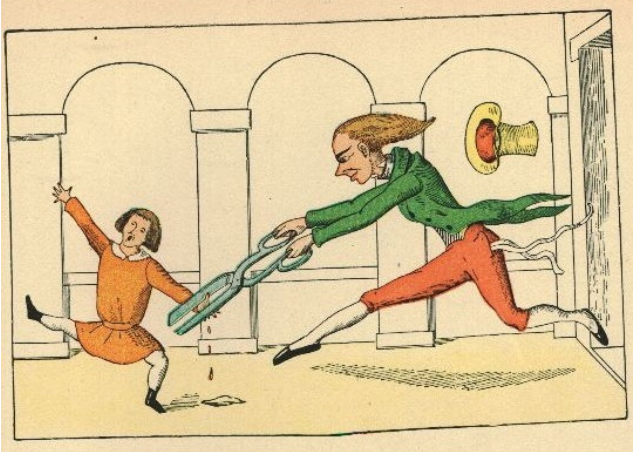
Betrachtet wird dabei nur, ob »Grausamkeit« durch ein Computerprogramm abschließend beurteilt werden kann. Wenn wir also »Grausames« betrachten, soll der Computer genau dieses Wort einem Web-Inhalt sicher zuordnen können. Eine bloße Erkennung von zum Beispiel Blut in einem Bild reicht noch nicht aus: eine blutende Wunde muss nicht automatisch durch Grausamkeit hinzugefügt worden sein, sondern kann auch bei einem Unfall entstanden sein. Umgekehrt fließt nicht bei jeder hinzugefügter Grausamkeit automatisch Blut.

Schon »Gewalt« ist für ein Computerprogramm nur schwer abstrakt zu erfassen. Auch für den Teilaspekt von physischer Gewalt am Menschen erscheint schon der Versuch ambitioniert, ihn algorithmisch zu beschreiben. Nochmal deutlich schwerer aus Computersicht zu erfassen sind die Begriffe »brutal« und »grausam«: auch hier ist eine große Bandbreite abzudecken, wobei zusätzlich auch eine emotionale Wertung hinzukommt.

Im Fragebogen zur Altersklassifizierung der FSM (vgl. Abschnitt 3.1.2.2) wird Gewalt fein differenziert: Unter dem Aspekt der Gewalt gegen Menschen oder menschenähnliche Wesen findet man die Aspekte Vergewaltigung, andere sexuelle Übergriffe, Foltern, Töten, wesentliche Verstümmelung, Blutvergießen und Verletzen.

Exemplarisch sei die Schwierigkeit der Erkennung eines grausamen Bildes an einem Bildbeispiel aus dem »Struwwelpeter« [Hoffmann, 1845] demonstriert. Insbesondere die Darstellung des Schneiders, der dem Daumen lutschenden Kind mit der Schere die Daumen abschneidet, gilt für viele als besonders drastische Gewaltdarstellung, zumal dies in einem Kinderbuch dargestellt ist, das ursprünglich für 3- bis 6-Jährige konzipiert wurde. In Abbildung 5.3 sind eine ganze Reihe von zu erwartenden Hürden aufgelistet, die bei der Erkennung einer grausamen Handlung in diesem Bild zu meistern wären. Wenn auch jede einzelne davon als prinzipiell möglich erscheint und zum Teil auch schon umsetzbar ist (z. B. die Emotionserkennung in Gesichtern), so ist hier eine große Anzahl von potenziell fehlerbehafteten Teilschlussfolgerungen und insbesondere ihre Verknüpfung mit heutiger Technik nicht zuverlässig umsetzbar. Hinzukommt, dass die erfolgreiche Erkennung eines Bildes noch lange keinen Generalisierungsanspruch erlaubt, vor allem auch, da bei anderen Bildern zum Teil deutlich unterschiedliche Teilschlussfolgerungen zu berücksichtigen wären.

Zu einzelnen Aspekten gibt es bereits Forschungsarbeiten: Beispielsweise lautet die Aufgabe bei der internationalen Text Retrieval Conference Retrieval Evaluation (TRECVid) Evaluierung von 2011, kurze Videobeiträgen in 500 semantische Begriffe einzuteilen. Von diesen Begriffen beschäftigen sich, neben einigen Vokabeln aus dem militärischen Umfeld wie etwa »Maschinengewehr« oder »Explosion«, zwei Bezeichner konkret mit physischer Gewalt: »Physischer Kampf« und »Gewaltsame Aktion«. Dennoch kann auch hier von keiner allgemeinen Erkennung von Gewalt gesprochen werden. Vielmehr lernt der Computer anhand eines limitiert großen Sets von vormarkierten Bildern Gemeinsamkeiten. Sollte dieses Material aber beispielsweise auf Gewaltbilder aus dem bewaffneten Nah-Ost-Konflikt beschränkt sein, wird das Programm noch nicht automatisch eine Kneipenschlägerei erkennen können.



Frage: Wer ist auf dem Bild zu sehen?
Aus Sicht eines Menschen: »Ein Mann mit einer Schere und ein Kind.«
Aus Sicht eines Computers: 1. Menschenerkennung: zwei, davon ein Kind. 2. Objekterkennung: Hut, Schere, Tür. 3. Erkennen, dass der Mann die Schere in den Händen hält.
Frage: Was passiert in dem Bild?
Mensch: »Der Mann schneidet dem Kind den Daumen ab.«
Computer: 1. Kenntnis, dass durch eine Schere auch Schnitte erfolgen. 2. Schwierigkeit: Schneidevorgang nur durch (farblose) Bluttropfen angedeutet. 3. Schwierigkeit: Hand bzw. Daumen als solche nicht zu erkennen, da von Schere verdeckt. 4. Körpermodell notwendig, um zu wissen, dass an dieser Stelle die Hand sein muss.
Frage: Ist das Bild grausam?
Mensch: »Das Abschneiden des Daumens ist besonders schwere Körperverletzung. Das Kind schreit um Hilfe, der Mann hingegen zeigt keine Emotionsregung, was besonders verstörend ist.«
Computer: 1. Kenntnis, dass Daumen wichtiges Körperteil ist. 2. Erkennung der Emotion über Gesichtsanalyse notwendig. 3. Es muss erkannt werden, dass der Mann in voller Absicht die Schere bedient.

Abbildung 5.3
Hürden bei der automatischen Erkennung von Grausamkeit, am Beispiel des Struwwelpeter [Hoffmann, 1845].

Zusammenfassend kann man festhalten, dass ein Computerprogramm menschlich wertende Begriffe wie beispielsweise »Grausamkeit«, »Unmenschlichkeit« oder »Agitation« nicht allgemein wird erkennen können. Vielmehr wird im Folgenden versucht werden, anhand von vorklassifizierten Daten andere ähnliche Daten erkennen zu können.

5.3.2 Fingerprinting bei Videoschnitten: Braindead

Aus der vorangegangenen qualitativen Bewertung von Filmen gegen Wehrlose erscheint bei Spielfilmen insbesondere die Audio-Fingerprinting Methode als vielversprechend. Wir werden ihre Effizienz nun an einem Horrorfilm exemplarisch untersuchen. Als Material soll uns hierbei der Spielfilm »Braindead« (engl.: »Dead Alive«) dienen. »Braindead« entstand 1992 unter der Regie von Peter Jackson (»Der Herr der Ringe«). Er handelt von einem Virus, der zunächst durch einen Affenbiss auf Menschen übertragen wird, so dass diese sich in untote, Menschenfleisch verzehrende Monster verwandeln. Verschiedene Fassungen sind in Deutschland indiziert oder beschlagnahmt. Die mutmaßliche Begründung für die Beschlagnahmung einer ungeschnittenen Fassung kursiert an verschiedenen Stellen im Internet, war für uns aber trotz mehrfacher Anfrage nicht verifizierbar. Da sie einer Inhaltsangabe der jugendschutzrelevanten Szenen gleicht, geben wir sie hier im Original wieder, um dann später zu versuchen, die angesprochenen Szenen automatisch anhand der Schnittfassung wiederzufinden (Tabelle 5.4).

In Deutschland frei erhältlich ist die jugendfreie Fassung (FSK 16) des Films, die gegenüber der ungeschnittenen Fassung um circa 25 Minuten gekürzt ist. Leicht geschnitten ist die amerikanische Fassung (Altersfreigabe: unrated), ungeschnitten ist sie in England und Österreich ab 18, in den Niederlanden ab 16 erhältlich.[3]

5.3.2.1 Verfügbarkeit des Films im Internet

Wir haben geprüft, inwieweit sich der Film oder Teile des Films im Internet ausfindig machen lassen. Eine Suche nach dem Film auf einem großen Videoportal, mit dem Suchbegriff »Brain Dead« im September 2012 zeigt 34 Treffer aus dem Film innerhalb der ersten 75 Resultate. Neben der ungeschnittenen spanischen und englischen Fassung ist auch die ungeschnittene deutsche Version durch leichte Variation der Suchanfrage vollständig verfügbar gewesen. Es existiert ebenfalls eine in kleinere Segmente zerteilte, zehnteilige Version, von denen einige als »bedenklich« in YouTube gemeldet sind, aber drei von den zehn Teilen sind auch ohne besondere Anmeldung zugänglich. Verschiedenste Ausschnitte und Zusammenschnitte zeigen einzelne, besonders drastische Szenen aus dem Film, zum Teil durch Text kommentiert (etwa: »die Top 10 menschlichen Opfer in Braindead« oder »die Rasenmäherszene«). Drei Videos sind kommentierte Inhaltsangaben des Films, zum Teil mit Bild-/Ton-Material, zum Teil nur bei gezeigtem DVD-Cover. Die FSK 16 Version des Films ließ sich auch nach längerem Durchforsten der Suchergebnisse nicht ausfindig machen. Einzig ein ausländischer Versandhandel machte gezielt Werbung mit einer besonders stark geschnittenen Szene des Films, die von einem jungen Mann auf dem Fernseher geschaut wird, mit dem Verweis, dass die ungeschnittene Fassung im Ausland vorrätig sei. Tabelle 5.5(a) zeigt eine Übersicht über die Natur der Treffer.

Videos, die als Resultate ausgegeben wurden aber nichts mit dem von uns gesuchten Film gemein hatten, waren vornehmlich Musikstücke. Da Braindead bei vielen als Klassiker seines Genres gilt, wohl aber auch aus Provokationsgründen wegen der eigentlichen Bedeutung des Wortes (»Hirntod«) tragen viele Bands »Braindead« schon im Bandnamen, und noch mehr Bands verwenden ihn

[3] Quelle: http://www.medienzensur.de, Stand 10.10.2012

Tabelle 5.4
Einzelbegründungen zum Beschlagnahmungsbeschluss des Filmes »Braindead« in der Fassung von ACME-Video, München, in der Ausgabe von 1995. Beschlagnahmung auf Grundlage des § 131, durch das AG Tiergarten, vom 1.3.2000.

6. min.	Im Vorspann des Films wird einem am Boden liegenden Mann die rechte Hand und dann der linke Arm mit einer Axt abgetrennt. Die Gliedmaßen sind in Nahaufnahme zu sehen.
18. min.	In einem Zoo wird einem Affen der Arm durch ein undefinierbares exotisches Tier abgerissen und danach verspeist. Dies ist in Nahaufnahme zu sehen.
19. min.	Das gleiche Tier beißt Lionels Mutter in den Nacken.[a] Die Wunde ist in Nahaufnahme zu sehen. Anschließend wird das Tier zertreten bis nur noch ein Brei zu sehen ist.
29. min.	Am Esstisch fällt der Mutter von Lionel das blutige Ohr ab. Aus dem Pudding wird es herausgeholt und dann von der Mutter genüsslich verspeist.
34. min.	Die Mutter reißt einer Krankenschwester von hinten die Wangen auseinander und den Kopf nach hinten, so dass eine klaffende Halswunde zu sehen ist. Unmittelbar danach wird diese Krankenschwester durch Lionel gefüttert, indem er ihr Brei in den geöffneten Hals gibt.
58. min.	Ein entstelltes Zombie-Baby wird in einen Mixer gesteckt. Der wird angestellt, bis das Baby nur noch ein blutiger Brei ist.
68. min.	Einem jungen Mann wird ein Rechen in den Kopf gerammt. Dies ist in Nahaufnahme zu sehen.
72. min.	Einem jungen Mann wird der Brustkorb aufgerissen und das Rippengerüst herausgeholt. Einem weiteren jungen Mann wird die Kopfhaut vom Kopf gerissen. Einer jungen Frau werden Stücke aus dem Hals gebissen und einer anderen jungen Frau werden die Gedärme aus dem Bauch gezogen. Diese Handlungen sind jeweils in Nahaufnahme zu sehen.
74. min.	Ein Mann wird durch ein Fenster gezogen, wobei das Unterteil skelettiert wird.
75. min.	Ein Zombie stößt einer Frau die Faust in den Nacken. Die Faust tritt durch den Mund wieder aus.
78. min.	Ein Mann trennt einem Zombie den Kopf mittels einer Zange vom Rumpf. Der Kopf wird als Fußball benutzt.
87. min.	Lionel zerstört alle Zombies mit einem Rasenmäher. Diese werden kleingemetzelt, wobei einzelne Körperteile wahllos herumfliegen.

[a] Die Beschreibung weicht bisweilen (in unwesentlichen Teilen) vom Filminhalt ab. Faktisch wird die Mutter in die Hand gebissen.

Tabelle 5.5
Die 75 ersten Treffer auf die Anfrage »Braindead«, in einem großen Videoportal.

(a) Echte Treffer aus dem Film »Braindead« und die Anzahl aufgeteilt nach Sprache: (D)eutsch, (E)nglisch, (F)ranzösisch und (S)panisch.

Was	D	E	F	S
Komplette, ungeschnittene Fassungen	1	1		1
Zusammenschnitte besonders drastischer Szenen	1	1		1
Filmkritiken mit Ausschnitten (aus der indizierten Fassung)	3			
Vergleichende Arbeit der Schnittversionen	1			
Kleine Segmente der ungeschnittenen Versionen	5			1
Einzelne, besonders drastische Szenen	5	5	1	
Trailer des Films	2	2	1	1
Werbung für einen ungeschnittenen Versand von Horrorfilmen aus dem Ausland, motiviert durch die geschnittene deutsche Fassung (!)	1			

(b) Videos, die nicht dem Film »Braindead« entstammen, aber als Resultat auf eine entsprechende Suchfrage gelistet werden.

Was	Anzahl
Musikgruppen mit »Braindead« im Namen	16
Musiklieder mit »Braindead« im Liedtitel	20
Filme mit »Braindead« im Titel	5
Reportagen zum Thema Hirntod	2
Sonstiges (Tanzvideo, Paintball-Teamname, Humorbeiträge, …)	11

als Liedtitel. Andere Treffer kamen durch Namen von Gruppierungen (als Paintball-Team) zustande. Auch gibt es andere Zombiefilme mit identischen oder ähnlichen Namen, was zu einigen zusätzlichen, wenn auch sehr wenigen, Treffern führte. Lediglich zwei Beiträge beschäftigten sich mit der medizinischen Kondition des Hirntods. Tabelle 5.5(b) beschreibt die Videos, die sich nicht auf den Film beziehen.

Zusammenfassend lässt sich festhalten, dass der Zugang zur beschlagnahmten Fassung des Films für Kinder und Jugendliche ohne großen Aufwand möglich ist.

5.3.2.2 Zeitliches Zuordnen der Szenen

In diesem Unterabschnitt widmen wir uns der Frage, inwieweit man durch Audio-Fingerprints die im vorigen Abschnitt beschriebenen Funde der ungeschnittenen Version zuordnen kann, und vor allem, inwieweit man auseinanderhalten kann, welche Ausschnitte davon noch aus dem FSK 16 Bereich stammen. Dazu gehen wir wie folgt vor: alle 10 Sekunden erzeugen wir einen Fingerprint auf Grundlage der ungeschnittenen deutschen Version und speichern diesen in einer Datenbank. Wir verwenden dazu das Verfahren wie in Abschnitt 4.5.1.2 beschrieben.[4]

Da es eine FSK 16 Version der Films gibt, ist zunächst die Frage, wie sich diese zur ungeschnittenen Version des Films verhält, das heißt, welche Segmente keine derartige Altersfreigabe aufweisen. Die Suche auf der FSK 16 Fassung mit Audio-Fingerprints der ungeschnittenen Fassung gestaltet sich

[4]Das heißt, der Fingerprint erfolgt anhand von temporaler und frequenzieller Nachbarschaftsanalyse und es wird ein 16 Bit Vektor angelegt, der aus Gründen der Robustheit mittels des Golay-(23,12,8) Verfahrens auf einen 12 Bit Vektor reduziert wird. Vektoren, die zu wenig Information enthalten, werden hierbei weggeschnitten; insgesamt bleiben hierdurch 576 Zeitabschnitte von insgesamt 589 erhalten. Das Verfahren ist in [Bardeli et al., 2012] vorgestellt.

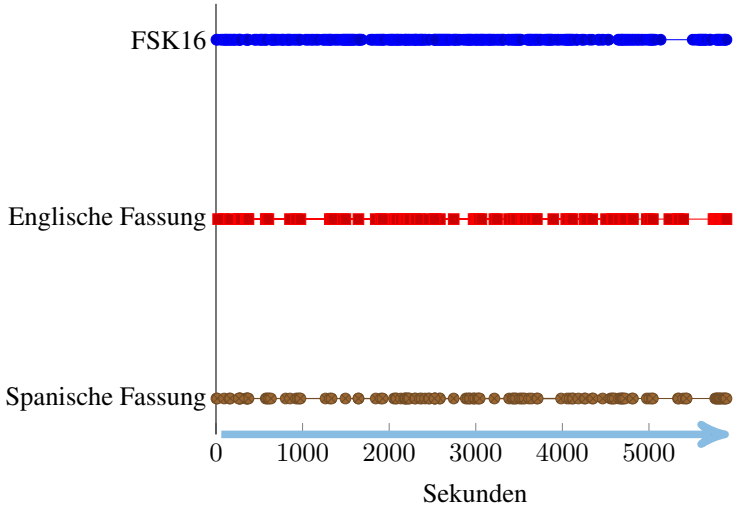


Abbildung 5.4
Anhand von Audio-Fingerprints aus der deutschen ungeschnittenen Fassung gefundene Abschnitte.

hierbei unproblematisch (78,4 % Treffer). Interessanterweise lassen sich zudem auch große Segmente der ausländischen Fassung wiederfinden (Englisch: 18,4 %, Spanisch: 22,6 %, beide ungeschnitten), was wohl daran liegt, dass in vielen Passagen (insbesondere den Kämpfen) kaum gesprochen wird, so dass die Soundeffekte annähernd identisch sind. In Abbildung 5.4 sind die gefundenen Abschnitte graphisch dargestellt. Mit vergleichbarer Präzision lassen sich so auch ein Großteil der Filmteile sowie der kleineren Teilsegmente, etwa in Szenenausschnitten oder in Best-of Zusammenstellungen, altersklassifiziert zuordnen (Abbildung 5.5).

Wir sind also mit dieser Technik in der Lage, folgende Aussage zur Jugendschutzrelevanz zu treffen: keinem der hier untersuchten Filmabschnitte liegen ausschließlich Szenen aus der geschnittenen FSK 16 Fassung des Films zugrunde. Es wird vornehmlich auf die expliziten und drastischen Ausschnitte der indizierten Version zurückgegriffen.

Im allgemeinen funktioniert das Verfahren sehr zuverlässig, mit folgenden Ausnahmen:

Tonqualität zu schlecht: Bei einer Besprechung des Films wurde eine offensichtlich sehr ausgeleierte VHS-Kassette abgefilmt. Die Frequenzen waren dadurch derart gestört (in diesem Fall Verschiebung zu tieferen Frequenzen), dass die verwendete Fingerprint-Methode nicht mehr zuverlässig funktioniert hat.

Zu kurze Segmente: Sind die Ausschnitte zu kurz, so lassen sie sich schlecht wiederfinden. In einem Extremfall wurde ein Videoschnipsel von drei Sekunden Länge nicht gefunden.

Kommentare: Die Filmkritiken sind zum Teil kommentiert und das Audiosignal ist dadurch derart verändert, dass es nicht mehr wiedererkannt wird.

Trailer: Gleiches gilt für die Trailer, die meist von einem Erzähler besprochen werden und damit nur noch Bild-Ähnlichkeit mit dem Originalfilm aufweisen.

Für die kommentierten Filme schauen wir im nächsten Abschnitt noch einmal explizit auf das Bildmaterial am Beispiel der »Braindead«-Trailer.

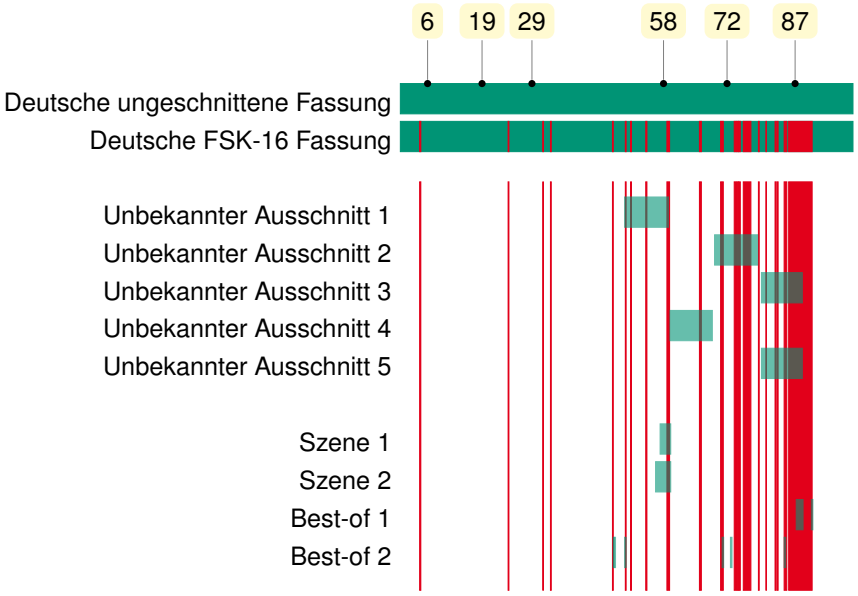


Abbildung 5.5
Abgleich der FSK 16 Fassung von »Braindead« mit der ungeschnittenen Fassung. Die Minuten über dem obersten Balken entsprechen den in Tabelle 5.4 beschriebenen Zeitangaben. Die automatisch lokalisierten Szenen (rot) finden sich in fast allen unbekannten Ausschnitten wieder und sind damit nicht ab 16 freizugeben.

5.3.2.3 Trailer

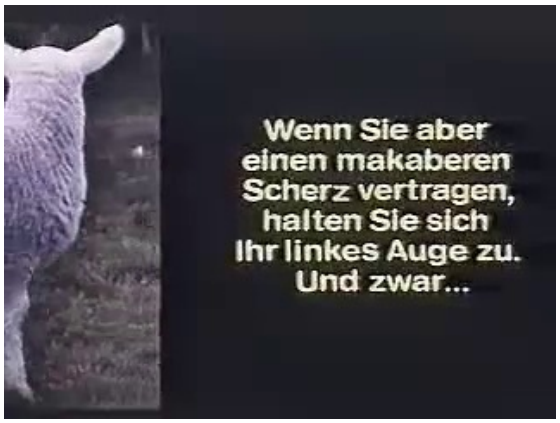
Die Trailer sind unterschiedlicher Machart und auf das Publikum zugeschnitten. Audio-Fingerprints helfen hier nicht viel, zumal selten der Originalton abgespielt wird, sondern das Geschehen durch einen Erzähler kommentiert wird. Keiner der Trailer hat einen einzigen Treffer auf die deutsche ungeschnittene Fassung. Auch herkömmliche Video-Fingerprints wie etwa die von pHash[5] bereitgestellten Programme finden hier ebenfalls keine Übereinstimmung zu dem Originalfilm. Dies ist nicht verwunderlich, da es sich bei den Filmen auch um stark unterschiedliches Material handelt; die Trailer unterscheiden sich zudem in ihren jeweiligen Landesfassungen voneinander:

- Da die deutsche Version schon in der Kinofassung stark gekürzt wurde, wurde hier durch »Warnungen« vor dem Material stärker auf die Vorstellungskraft gesetzt (Abb. 5.6(a)). Bei einem anderen Video ist die Alterskennzeichnung vor den Film geschaltet (Abb. 5.6(b)).
- Im Spanischen werden zunächst die Charaktere eingeführt, der humorvolle Aspekt des Films wird stärker betont (Abb. 5.6(c)).
- Im Englischen lautet der Titel »Dead Alive« (Abb. 5.6(d)).

Wir wollen uns in diesem Abschnitt daher mit der Frage auseinandersetzen, inwieweit sich die Videos dennoch über Video-Fingerprints der ungeschnittenen Fassung zuordnen lassen. Es wird also gezielt nach einzelnen Filmpassagen gesucht, die sich auffallend ähneln. Generell ist in den Trailern die Bildqualität sehr stark unterschiedlich, was auch bei den anderen Videos zu beobachten ist (Abb. 5.7). Die folgenden Experimente wurden von der Forschungseinrichtung »Institut national de recherche en informatique et en automatique« (INRIA) an dem von Fraunhofer IAIS zusammengestellten Material durchgeführt. Das eingesetzte Verfahren ähnelt dem in [Douze et al., 2010].

Zu den ursprünglichen Anfragen (Videos in Abbildung 5.7), wurden jeweils übereinstimmende Segmente gefunden, allerdings gab es hierbei auch einige Lücken, etwa, wenn der Inhalt in kurzen Abschnitten neu zusammengestellt wurde. Insgesamt wurden 200 Dateien untersucht; neben den

[5] siehe http://www.phash.org, Stand: November 2012



(a) Zweiteilung des Bildes im deutschen Trailer



(b) Alterskennzeichnung in einem der deutschen Trailer



(c) Charaktereinführung im spanischen Trailer



(d) Anderer Titel im englischen Trailer

Abbildung 5.6
Unterschiedliche Aspekte in den Trailern für den Film »Braindead«, die in den jeweils anderen Sprachen nicht vorkommen.



(a) Deutscher Trailer 1



(b) Deutscher Trailer 2



(c) Spanischer Trailer



(d) Englischer Trailer 1



(e) Englischer Trailer 2



(f) Versionsvergleich

Abbildung 5.7
Unterschiedliche Bildqualität in den Trailern für den Film »Braindead«. Bei dem deutschen Trailer ist die jeweils linke Seite mit anderem Filmmaterial gefüllt, damit sich »empfindsame Zuschauer« (Zitat Trailer-Werbung) das rechte Auge zuhalten können.

oben erwähnten 6 Trailern sowie zwei Kritikfilmen, die das Material enthielten, wurden zusätzlich Videos eines großen Videoportals verwendet, die auf die Anfrage »Brain Dead Alive Trailer« ausgegeben wurden; hier waren allerdings auch einige von Fans zusammengestellte Versionen enthalten. Insgesamt gab das Verfahren 342 Treffer in insgesamt 107 von diesen 200 Filmen wieder, das heißt, hier wurden Segmente aus der ungeschnittenen deutschen Version identifiziert. Bei der manuellen Überprüfung konnte bei 87 von den übrigen Filmen verifiziert werden, dass sie keine Segmente von Braindead enthielten. 6 Filmschnipsel wurden nicht gefunden. Erwähnenswert ist noch, dass keines dieser Segmente – bedingt durch den direkten Vergleich mit einem anders geschnittenen Vollfilm – länger als 2 Sekunden war.

Durch die zeitliche Zuordnung zur ungeschnittenen Version ist auch hier bei den entsprechenden Versionen eine Beurteilung bezüglich der Altersbeschränkung möglich (Abb. 5.8). Allerdings zeigt sich bei der Betrachtung im Detail, dass diese Aussage nicht im Umkehrschluss gelten muss, da manchmal nur die Szenen aus der geschnittenen Fassung gefunden wurden, die Gewaltszenen aus der beschlagnahmten Fassung jedoch nicht, obwohl sie enthalten waren.

Um dies zu zeigen, haben wir Szenen aus einem Video zum Thema »Braindead«, genauer eine von einem Fan erstellte Gegenüberstellung von Szenen aus den verschiedenen Schnittfassungen, näher überprüft. In vielen Einstellungen werden die Szenen nur als kleine Einblendung vor einem großen Hintergrund gezeigt. Die hier verwendete ungeschnittene deutsche Fassung, ebenfalls auf dem gleichen Videoportal verfügbar, hat durchweg einen großen schwarzen Rahmen um das Bild und einen in der oberen linken Ecke eingeblendeten Schriftzug. Wenn auch sieben einzelne Szenen erkannt werden, so ist insbesondere die wohl drastischste Szene des Films, in der ein Rasenmäher als Nahkampfwaffe eingesetzt wird, nicht erkannt worden (Abb. 5.9).

Das Verfahren läuft auf einem herkömmlichen Rechner mit einem Prozessor für die Analyse eines Videos etwa in Echtzeit (das heißt, die Analyse von einer Minute Video dauert ebenfalls eine Minute) und belegt dabei etwa 30 Kilobyte pro Sekunde. Die Dauer der Suche ist abhängig von der Größe der Datenbank. Bei einer Datengröße von 200 Stunden dauert die Suche nach einem Clip der Größe von 1 Minute etwa 30 Sekunden. Die Suchdauer wächst dabei linear mit der Größe der Datenbank beziehungsweise der Größe des Clips.

Zusammenfassend lässt sich sagen, dass die Video-Fingerprints ein sehr mächtiges Werkzeug zur Erkennung von jugendschutzrelevantem Material sind und durch zeitliche Zuordnung auch eine Altersklassifizierung bei verschiedenen Schnittfassungen erlauben. Bei einem Einsatz in der Praxis sind aber mitunter Ungenauigkeiten bei einzelnen Szenen sowie eine relativ hohe Ressourcenauslastung in der Analyse zu erwarten.

5.3.3 Am Beispiel von Gewaltspielen

Wir haben uns bislang in diesem Abschnitt mit Spielfilmen auseinandergesetzt, also statischem Inhalt, dessen zugrunde liegendes Material sich nicht mehr verändert, der aber durch Videoschnitt vom Regisseur (z. B. »Director's Cut«), Vertrieb (Schnittfassung) oder Benutzer (»Best-of«) in neuer Zusammenstellung auftauchen kann. Für Computerspiele mit Gewaltdarstellung gelten hier andere Voraussetzungen, weswegen wir nun qualitativ zwei Videos zu diesem Thema betrachten wollen. Die Videos sind entweder offiziell vom Spielehersteller herausgegeben oder von einem Fan erstellt und zeigen den Mitschnitt eines Gewaltspieles. Letztere unterscheiden sich damit von anderen Gewaltvideos insofern, als dass keine identischen Szenen angenommen werden können. Die Videos werden vielmehr in Echtzeit, also während eines Durchspielens erzeugt und sind abhängig von der Spielweise des Nutzers immer wieder einzigartig.

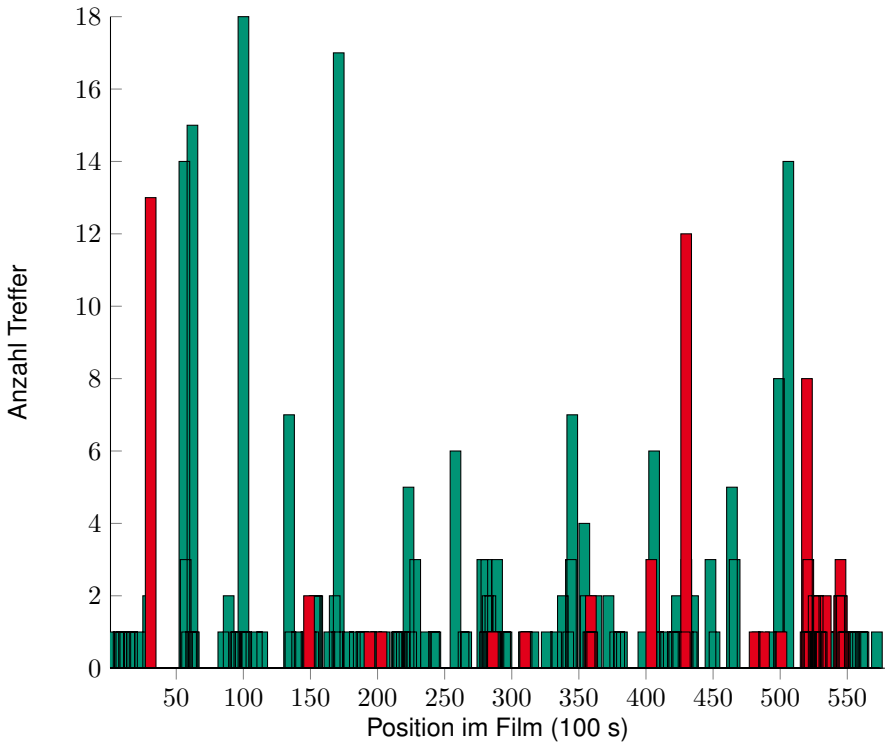


Abbildung 5.8
Gefundene Treffer von Ausschnitten aus 200 verschiedenen Trailern, in der ungeschnittenen Version. Treffer aus den ungeschnittenen Szenen, die nicht in der FSK 16 enthalten sind, sind rot markiert.



(a) Szenen aus dem Versionsvergleich-Video »Braindead«…



(b) … die in der ungeschnittenen Fassung gefunden wurden.



(c) Im Versionsvergleich-Video hinzugefügte Einstellungen ohne Entsprechung im Original (etwa, um die verschiedenen Schnittfassungen optisch einzuleiten).



(d) Szenen aus dem Versionsvergleich-Video, die nicht in der ungeschnittenen Fassung gefunden wurden.

Abbildung 5.9
Ergebnisse der Video Copy Detection bei einem Fan-erstellten Video zum Thema »Braindead«.

Betrachtet wurden zwei Videos, in denen aus der Ich-Perspektive ein Spieler durch eine Landschaft wandert – in einem Fall eine ausgebombte Hochhaussiedlung, im anderen Fall eine verlassene Insel – dort auf verschiedene Gegner trifft und diese mit einem großen Arsenal an Waffen bekämpft. Das erste Video ist dabei militärischer Natur und zeigt besonders realistische Bilder aus einem kriegsähnlichen Einsatzszenario (im Folgenden *Soldatenvideo*). Das zweite Video handelt vom Überleben auf einer Insel, auf der mehrere menschenähnliche Monster von verschiedenen Seiten auf den Spieler einstürmen, mit besonderem Augenmerk auf möglichst spektakulär inszenierte Kampfeinstellungen (im Folgenden *Monstervideo*).

Das Soldatenvideo ist eine Demonstration der Vorabversion des Spiels. Es ist also vom Spielehersteller professionell aufbereitet und zeigt exemplarisch die Spielfunktionalität. Das heißt zum einen, dass das Material nicht nur zum Zeitpunkt des Erscheinens absolut neuartig ist, sondern auch, dass das fertige Spiel noch nicht auf dem Markt ist und damit auch noch nicht offiziell auf jugendschutzrelevanten Inhalt untersucht werden konnte. Hinweis kann hier lediglich sein, dass es sich bei dem vorliegenden Spiel um den Teil einer Serie handelt, deren Vorgänger bereits klassifiziert wurden. Neben der Möglichkeit der Texterkennung erscheint hier jedoch die Logodetektion für den Spieletitel als sinnvoller, weil dieser oftmals ähnliche Elemente wie die Vorgängertitel enthält. Ebenfalls erkannt werden können Logos von Spieleherstellern, was bei großer Angebotspalette aber sicher noch kein hinreichendes Merkmal für den Inhalt darstellt. Denkbar wäre hier auch wiederum, mit Hilfe von Audio-Fingerprints die Hintergrundmusik des Spiels zu erkennen und mit bekanntem Material zu vergleichen, was aber aufgrund der Dynamik des Inhaltes als schwer erscheint; so werden oft durch die Handlung des Spielers Geräusche, Sprachschnipsel oder auch eine angepasste Musikkulisse eingebaut, die das Wiederauffinden von bekannten Charakteristika deutlich erschwert. Von gesondertem Interesse sind hier die wenigen vom Spiel vorgesehen Passagen, die weitestgehend gleich bleiben, sogenannte *cut-scenes*. Inwieweit sich diese automatisch identifizieren lassen, muss gesondert untersucht werden.

Das Monstervideo fällt in die stets größer werdende Kategorie der sogenannten »Let's Play«-Videos, in der die Lautstärke des ursprünglichen Videos oftmals heruntergeregelt wird und der Spieler selbst Kommentare zu den Handlungen im Spiel abgibt. Das ermöglicht prinzipiell den Einsatz von automatischer Spracherkennung, um anhand des gesprochenen Inhalts den Namen des Spiels oder zumindest dessen Handlung ausfindig zu machen. Das vorliegende Video stellt an das Verfahren allerdings einige, wenn auch nicht unlösbare Anforderungen. Zum einen sind die relevantesten Kommentare – etwa »Den tret ich tot, da machen wir uns keine Waffe kaputt hier« – immer dann am häufigsten, wenn die Hintergrundgeräusche besonders laut sind, also bei Kampfhandlungen. Zum anderen wird insbesondere im Spielebereich häufig ein eigenes Fachvokabular verwendet, dass sich zudem mit der Zeit wandelt (z. B. »fraggen«). Ein bloßes Sammeln dieser Begriffe durch das Durchsuchen von einschlägigen Foren reicht im Falle der Spracherkennung nicht aus, da man auch die Aussprache (Phonetisierung) kennen muss, um sie gezielt in einer Tonspur wiederzufinden; in diesem Falle wird *fraggen* englisch (oder je nach Spieler englisch-ähnlich) ausgesprochen.

Angezeigt werden in beiden Videos genreübliche Übersichtstafeln, die die verbleibende Lebensenergie des Spielers sowie die Munition und/oder den Zustand seiner derzeit gezückten Waffe anzeigen. Die Waffe wird dabei ebenfalls aus der Ich-Perspektive angezeigt. Zusätzlich ist in der Mitte des Bildschirms ein Fadenkreuz zu sehen. Im Monstervideo wird zudem oft in ein Menü gewechselt, dass die derzeitige Mission darstellt oder den Zustand aller mitgeführten Waffen angibt. Da diese Objekte im Vergleich zum bewegten Hintergrund oftmals statisch sind, können sie durch Bewegungsfilter gesondert erkannt und damit als Objekte abgespeichert werden, die dann wiederum in anderen Filmen des gleichen Spiels wiedererkannt werden können. Zu beachten ist hierbei aber, dass manche Objekte animiert oder durchscheinend sind, zum Teil vom Spieler verändert werden können (z. B. die Größe oder Form des Fadenkreuzes) und sich von Spiel zu Spiel unterscheiden. Sie können also nur

ein zusätzliches, vermutlich eher unzuverlässiges Merkmal bilden, so dass der Computer lediglich einen ähnlichen Inhalt vermuten kann.

Auch hier scheint der Einsatz eines *Konzeptdetektors* aufgrund der Heterogenität des Materials zumindest als schwierig. Prinzipiell könnte eine Geräuschdetektion von zum Beispiel Schussgeräuschen, Schmerzensschreien oder Explosionen einen ersten Hinweis geben. Ebenfalls denkbar wäre das Erkennen eines Konzepts, dass am unteren Bildschirmrand eine Waffe geschwungen würde. Allerdings sind in vielen Spielen besonders realistisch aussehende Waffen vorzufinden (z. B. in Kriegsspielen), andere weichen bewusst von »klassischen« Darstellungsformen ab und versuchen, möglichst ungewöhnliche Waffen hervorzubringen (z. B. in futuristischen Spielen). Auch ein zumindest technisch denkbarer Blutdetektor dürfte, wenn überhaupt, nur einen Teil des Materials abdecken. Im Falle des Monstervideos finden sehr viele Angriffe über stumpfe Schlagwaffen statt, das Soldatenvideo zeigt eine Reihe von Raketenschüssen etwa auf einen weit entfernten Scharfschützen. Zudem gibt es eine Reihe von Spielen, in denen dem Spieler über weite Strecken eine stark veränderte Sicht simuliert wird, etwa durch eine simulierte Gasmaske, eine Nachtsichtbrille oder durch eine »magische« Fähigkeit.

5.3.4 Spracherkennung: Let's play-Videos

Wir haben im vorherigen Unterabschnitt Videos von Gewaltspielen untersucht und die automatische Spracherkennung als eine mögliche Form der Datenextraktion bei »Let's play«-Videos identifiziert. Da wir bei dem Material mit einigen Schwierigkeiten rechnen, soll nun geprüft werden, ob eine herkömmliche Spracherkennung dafür ausreichend ist, um das Material altersgerecht einzustufen, oder ob die Analysequalität den Anforderungen hierfür nicht genügt.

Let's Play-Videos erfreuen sich wachsender Beliebtheit und zu praktisch allen neueren sowie zu hinreichend bekannten alten Spielen findet sich schnell eine große Anzahl an Beiträgen. Um eine möglichst große Bandbreite an Altersabstufungen abzudecken, haben wir uns zu jeweils einem Let's Play Video zu den folgenden Spielen entschieden:

Minecraft In Minecraft (2009, Entwickler: Mojang) kann man aus kubischen Blöcken seine eigene Welt zusammenbauen. Die Spielfigur trifft im Laufe des Spiels auf stark verpixelte Gegner, die Kämpfe sind dabei aber harmlos und ohne Blut. Das Spiel hat die Einstufung USK 6.

Doom Das Spiel Doom (1993, Entwickler: id Software) kam bereits Anfang der 90er auf den Markt und war wegen der damals neuartigen Technik ein großer Erfolg. Inhaltlich kämpft ein namenloser Soldat sowohl gegen menschliche Gegner als auch gegen Monster. Als Waffen dienen unter anderem Kettensäge, Shotguns und Panzerfaust. Das Spiel war in Deutschland bis August 2011 indiziert; dieser Beschluss wurde dann aufgehoben und das Spiel ist heute als USK 16 eingestuft.

Battlefield 3 (2011, Entwickler: Digital Illusions CE) Das Spiel ist im Militärbereich angesiedelt (man spielt entweder amerikanische oder russische Soldaten) und setzt auf möglichst realistische, menschliche Darstellung von Kriegsgeschehen. Es ist in Deutschland als USK 18 freigegeben.

Borderlands (2009, Entwickler: Gearbox Software) Dieses Spiel zeichnet sich durch eine überzogen Comic-artige Darstellung aus. Alleine oder in einem Team von bis zu vier Mitspielern sucht man beständig nach neueren, stärkeren Schusswaffen, um gegen Monster und Menschen anzukämpfen. Die entschärfte deutsche Fassung, in der unter anderem keine Gliedmaßen der Gegner abgetrennt werden, hat die Freigabe USK 18, wohingegen die ungeschnittene Fassung indiziert ist.

Dead Rising 2 (2011, Entwickler: Capcom) In diesem Spiel werden verschiedene Zombies mit einer Großzahl an improvisierten Waffen bekämpft. Das Spiel ist bundesweit beschlagnahmt.

Beim ersten Durchhören der Videos zu diesen Spielen finden sich hier eine Menge Wörter, die nicht im normalen Wortschatz enthalten sind. Unbekannte Wörter können prinzipiell aus Textforen oder den Spielbeschreibungen gesammelt und dann automatisch phonetisiert werden, indem die Aussprachеregeln aufgrund des sonstigen deutschen Vokabulars gelernt werden. Das mag auch für manche Neologismen noch funktionieren (etwa: »Gölemchen« als Verniedlichungsform von »Golem«), bei anglizistischen Wortkreationen wie etwa »Animations-Fail«, wo der erste Teil des Wortes deutsch ausgesprochen wird und der zweite englisch, ist das deutlich schwieriger und müsste größtenteils von Hand definiert werden. Auch für rein englische Wörter – die Spieltitel sind bereits alle auf Englisch – stößt die Phonetisierung auf Schwierigkeiten. Die Verwendung eines Phonetisierers für die englische Sprache ist nicht unbedingt zielführend, da die Aussprache eher eine Mischung beider Sprachen ist. So würde man nach englischen Aussprachеregelungen für den Ausdruck »Ultra Violence« am ehesten die Aussprache »altra ua-iolens« erwarten, tatsächlich wird es von dem Spieler hier aber »ultra wai-olentz« ausgesprochen.

Exemplarische Ergebnisssätze der Spracherkennung sind in Tabelle 5.6 dargestellt. Wörter, die nur in einem zweiten Anlauf mit manuell angepasster Phonetisierung erkannt werden konnten, sind entsprechend markiert. Aufgrund der schwierigen Grundbedingungen sind die Ergebnisse aus der Spracherkennung ernüchternd. Faktoren, die die Spracherkennung zusätzlich zu der Aussprache und dem Vokabular erschweren, sind dabei:

Ungeübte Sprecher: Die Spieler zögern oft im Redefluss und gerade bei Handlungen, die Konzentration erfordern, werden die Wörter sehr gehetzt ausgesprochen.

Schlechte Mikrophone: Die Sprache wird stark verrauscht, häufig entstehen zudem Clipping-Artefakte, wenn für das Gerät zu laut gesprochen wird.

Geräusche vom Spiel: Zwar ist die Lautstärke in den meisten Spielen heruntergeregelt, dennoch sind die Geräusche im Spiel klar hörbar. In einem Fall ist zusätzlich noch Rockmusik im Hintergrund zu hören, die wohl von der Stereo-Anlage parallel zum Spielekonsum abgespielt wird.

Zusammenfassend kann man sagen, dass Spracherkennung auf Let's Play-Videos nicht von ausreichender Qualität ist, um zuverlässig auf den Inhalt schließen zu können. Dennoch sind viele der Probleme gezielt lösbar, wobei ein relativ hoher Aufwand einzuplanen ist. So kann man zum Beispiel versuchen, eine »denglische« Aussprache gezielt zu trainieren. Das Vokabular könnte wie oben beschrieben durch regelmäßiges Durchforsten von Spieleforen aufgefrischt und auch typische Wortfolgen könnten gezielt für diese Spiele im Sprachmodell trainiert werden. Abzuwägen ist also, inwieweit Let's Play-Videos ein wesentlicher Bestandteil von jugendschutzrelevantem Material darstellen. Es ist jedenfalls festzuhalten, dass eine bloße Eingruppierung von Let's Play Videos nach der Alterseinstufung des zugrunde liegenden Spiels mitunter nicht ausreicht, wie man beispielsweise an dem Minecraft-Beispiel sehen kann. Das Spiel wird auch von Erwachsenen gerne gespielt, so dass Wortbeiträge wie das noch vergleichsweise harmlose »am Arsch lecken« trotz des ursprünglichen USK 6 Inhaltes vorkommen. Dies dürfte bei den meisten Videos aber eine eher untergeordnete Rolle spielen.

Tabelle 5.6
Ergebnisse der Spracherkennung auf den Let's Play Videos.

Minecraft	So, die Spinne kann mich mal kreuzweise mit ihren Drüsen am Arsch lecken, denn wir verpieseln uns jetzt in Richtung Bauernhaus und probieren mal det Gölemchen zu bauen
Doom	…zumindest die ersten drei Episoden, danach kann die Sintflut kommen. Also, dann starten wir jetzt ein Spiel in der dritten Episode, nämlich Inferno. Und, wie immer, Schwierigkeitsgrad Ultra Violence.
Battlefield 3	Es geht endlich los mit Let's Play Battlefield 3 [..] Ich werde euch durch dieses actionlastige Shooter abenteuer führen, das Game aus dem Jahre 2011
Borderlands	Das könnte ich auch einfach da runterhauen – ja … das funktioniert nicht. Animations-Fail!
Dead Rising 2	Ja anscheinend ist er gerade gesnapt, weil seine Kollegen oder Untergebenen oder so was ges getötet wurden auf uh auf … Granate …ich weiß genau wie ich dich [?] Scheiße … ah shit … aua, das tut doppelt weh

Legende:
ohne Farbe = nicht erkannt
mit hellgrüner Farbe = von der Spracherkennung erkannt
mit dunkelgrüner Farbe = gefundene Wörter bei angepasster Phonetisierung

5.3.5 Brutalitätserkennung: Robbenjagd

Wir haben uns bislang mit Gewaltdarstellungen beschäftigt, die durch den Titel des Materials, also durch den Namen des Spielfilms oder des Computerspiels, bereits bekannt sind. Bei Gewaltdarstellung in Bildern ist ein derartiges Vorwissen für gewöhnlich nicht vorhanden. Eine automatische Erkennung des Inhaltes nur auf Grundlage des visuellen Inhaltes scheint zudem nur in Ausnahmefällen möglich.[6]

Bilder, auf denen zum Beispiel Tierfolter zu sehen ist, sind so mannigfaltig, dass eine automatische Erkennung hier als nicht zielführend erscheint (vgl. Abb. 5.10). In diesem Unterabschnitt wollen wir uns daher der Frage widmen, ob eine Klassifikation der Webseiten alleine aufgrund ihres Textinhaltes möglich erscheint.

Zu diesem Zweck wurden 225 Webseiten zum Thema »Robbenjagd« manuell zusammengestellt und bewertet. Dabei wurden sie jeweils nur anhand der enthaltenen Bilder in eine der Kategorien »harmlos« und »drastisch« eingeordnet. Die Annahme ist, dass die Seiten mit eher drastischen Gewaltdarstellungen entsprechende Merkmale aufweisen wie »diese grausamen Bilder« oder »Schaut Euch das an!«. Jedoch ist die vorhandene Datengrundlage zu klein als dass sich signifikante Häufungen aufdecken ließen.

Manche der betrachteten Webseiten sind Teil desselben Webauftritts. Eine Gruppierung der Merkmale der einzelnen Dokumente nach ihrem jeweiligen Webauftritt ist daher notwendig, um zum Beispiel deren Namen und andere Webauftritt-spezifische Merkmale auszuschließen, die sich nicht zur allgemeinen Charakterisierung der Zielkategorie eignen. Nach diesem Schritt bleiben von den 109 drastischen und 116 harmlosen Webseiten noch 62 drastische und 66 harmlose Webseiten-Gruppen übrig. Daraus lassen sich zwar relevante Häufungen einzelner Wörter aufdecken, komplexere Aussagen sind bei derart reduziertem Material allerdings nur sehr eingeschränkt möglich. Die stärksten Merkmale sind in Tabelle 5.7 dargestellt.

[6] Vergleiche hierzu generelle Überlegungen in Abschnitt 5.3.1.1 sowie die Konzepterkennung speziell für Ritzbilder in Abschnitt 5.5.4.



(a) Bildquelle:
http://upload.wikimedia.org/wikipedia/commons/7/79/Inuit_sealhunter.jpg,
abgerufen am 29.11.2012 im deutschsprachigen Artikel »Robbenjagd«.



(b) Bildquelle:
http://www.waldorf-elmshorn.de/Impressionen/Impressionen2007/Robbenjagd/repectforanimals6b_520.jpg,
abgerufen am 29.11.2012

Abbildung 5.10
Unterschiedliche Gewaltdarstellung bei Webseiten zum Thema Robbenjagd. Das Bild in (b) wurde mit einer Altersfreigabe von 14 Jahren eingestuft.

Tabelle 5.7
Die relevantesten Merkmale für harmlose und drastische Gewaltdarstellungen zum Thema »Robbenjagd« .

»harmlos«	»drastisch«
kolonie	erschlägt
projekt	von fellen
bejagt	tiere getötet werden
rückgang derart	qualvollen tod
ziehen schwangere	an die kanadische regierung
firmen	der internationale tierschutz fonds
system	verantwortung
jungtiere	robben abgeschlachtet werden
im februar	kanadischen premierminister stephen harper
lebensräume	knüppeln
zum schutz	importverbot von robbenprodukten
für touristen	von robbenjägern
seebär	brutalität
dezimieren	kanadas schande

Während die Erkennung des »Robbenjagd«-Materials gegenüber allgemeinen Webseitenkollektionen mit etwa 97 % sowohl bei der Präzision als auch der Trefferquote recht einfach scheint, ist die Trennung in »enthält harmlose Bilder« und »enthält drastische Bilder« wesentlich komplizierter. Die in Tabelle 5.7 dargestellten Merkmale sehen zunächst gut nachvollziehbar aus. Diese Begriffe sind zwar stark an die entsprechende Kategorie gebunden, treten allerdings insgesamt recht selten auf. Selbst die häufigsten dieser Merkmale treten in nicht mehr als fünf bis sechs Webseiten-Gruppen auf. Das ist zu wenig, um daraus ein gutes Modell abzuleiten.

Ein zusätzliches Problem stellt bei diesem Experiment die schwammige Trennung in »harmlos« und »drastisch« dar. Eine Stichprobe auf den Trainingsdaten ergab eine Einigkeit der menschlichen Annotatoren von nur rund zwei Dritteln. Das bedeutet, jedes dritte Dokument hätte ein anderer Annotator anders eingestuft. Dies führt zu einem nicht unerheblichen Rauschen in den Daten und erschwert ihre maschinelle Auswertung.

Daher ist es auch nicht verwunderlich, dass der automatische Klassifizierer mit per Kreuzvalidierung ermittelten Genauigkeiten um die 50 % nicht besser als eine zufällige Auswahl ist. Um dieses Ergebnis zu verbessern, ist in erster Linie eine breitere Datengrundlage notwendig, um daraus eindeutigere Muster ableiten zu können. Zum anderen erscheint ein Versuch interessant, inhaltlich abweichende aber funktional ähnliche Kategorien zusammenzuführen. Dies könnten alle erdenklichen Informationsangebote zum Tierschutz sein, die mit schockierenden Bildern illustriert sind. Beispielsweise kämen dafür Bereiche wie die Geflügelzucht, Tierhaltung im Zirkus oder schlechte Verhältnisse in Tierheimen infrage. Das Wichtigste aber bleibt – wie bei jedem der hier vorgestellten Versuche – eine möglichst große Anzahl von Dokumenten zu berücksichtigen, die mit Blick auf die Zielkategorie homogen und bezüglich ihrer Struktur heterogen sind, also aus vielen verschiedenen Webauftritten stammen.

Zusammenfassend lässt sich also festhalten, dass die Erkennung von drastisch brutalen Bildern durch den sie umgebenen Text auf Grundlage dieses Experiments nicht abschließend geklärt werden kann, aber prinzipiell bei geeigneten Daten als möglich erscheint.

5.3.6 Textklassifikation

Zur Beantwortung der Frage, inwieweit sich Webseiten nur anhand textueller Informationen klassifizieren lassen, wurden Webseiten manuell in Kategorien einsortiert. Zunächst werden diese Kategorien zusammengefasst betrachtet. Das bedeutet, die jeweiligen Unterkategorien wie zum Beispiel »Folter«, »Krieg« oder »Vergewaltigung« werden ihrer Altersfreigabe entsprechend zu den Kategorien »Gewalt ab 18«, »Gewalt ab 16«, »Gewalt ab 14«[7] und »Gewalt ab 12« zusammengefasst. Dies ergibt die folgenden Kategorien:

- *Gewalt ab 18*: Eine Sammlung von 889 manuell annotierten Webseiten.
- *Gewalt ab 16*: Eine Sammlung von 206 manuell annotierten Webseiten.
- *Gewalt ab 14*: Eine Sammlung von 152 manuell annotierten Webseiten.
- *Gewalt ab 12*: Eine Sammlung von 256 manuell annotierten Webseiten.

Da im realen Einsatz die genannten Kategorien von allgemeinen, unbedenklichen Webseiten zu trennen sind, wird eine zusätzliche Kategorie ergänzt:

[7] Das zur Verfügung gestellte Datenmaterial enthielt diese zusätzliche Unterteilung »ab 14«, die in der Praxis dann der Altersstufe »ab 16« mit zugeordnet wird.

Tabelle 5.8
Die relevantesten Sprachmerkmale für Unterkategorien von Gewalt.

Tierquälerei	Abtreibung	Ogrish	Vergewaltigung
galgo	abortion	ogrish	vergewaltigung
hund	abtreibung	death	rape
animals	pro life	murder	vergewaltigt
vivisection	baby	killed	vergewaltigungsszene
abusers	mord	execution	brutale
tierheim	zellklumpen	dead	gewalt
tier	ungeboren	life	abgehackt
entsorgt	baby	suicide	abstumpfen
jäger	embryo	accident	totgeschwiegen
tierschützer	mutterleib	gore	sexual
überfahren	abgetrieben	fuck	verschließen
spanien	unborn	uncensored	victim
traurig	abtreiben	morgue	schreien

- *Allgemein ab 6*: Eine Sammlung von 3 139 manuell annotierten Webseiten. Davon wurden 210 gezielt zum Thema »Gewalt« gesucht, gesichtet und für nicht jugendschutzrelevant befunden.

Die Experimente[8] wurden mit fünffacher Kreuzvalidierung (siehe Abschnitt 4.2.3) über den genannten Webseiten-Kollektionen durchgeführt. Dabei wurde darauf geachtet, dass Webseiten derselben Domäne (z. B. »fraunhofer.de«) immer auch derselben Teilmenge der Kreuzvalidierung angehören. Dies ist wichtig, um möglichst realistische Ergebnisse zu erzielen, da Webseiten derselben Domäne zumeist auch derselben Kategorie angehören. In diesem Fall würde beispielsweise der Name der Webseite unrealistisch viel über die Kategoriezugehörigkeit aussagen.

Als die Webseiten beschreibende Merkmale (vgl. Tabelle 5.8) werden folgende Gruppen definiert:

- *nur URL*: Dies sind alle Wörter, die in der Webadresse vorkommen. Zusätzlich werden die zusammenhängenden Wörter, also die Wort-N-Gramme mit einem N bis 10 betrachtet. Kommt in der Webadresse also »/gewalt/text.html« vor, so werden die Wörter »gewalt text« auch zusammenhängend als Merkmal gesehen.
- *nur Text*: Dies beinhaltet den Titel der Webseite, die META-Tags zu »keywords« und »description« und den eigentlichen Inhalt der Webseite.
- *Text+URL*: Eine Kombination beider Merkmalsgruppen.
- *nur Haupttext*: Mittels einer Software-Bibliothek[9] wird der Hauptteil der Webseite extrahiert. Durch das Weglassen der Navigationselemente und Werbeeinblendungen wird der wesentliche Inhalt der Webseite hervorgehoben.
- *Haupttext+URL*: Eine Kombination beider Merkmalsgruppen.

Für den praktischen Betrieb der Anwendung wäre es hilfreich, wenn sich bereits anhand der Webadresse die Jugendschutzrelevanz ablesen lassen würde. Auf diese Weise ließe sich direkt schon die Anwahl des entsprechenden Inhalts blockieren. Die in Tabelle 5.9 gezeigten Ergebnisse lassen allerdings den Schluss zu, dass die Adressen alleine zu wenig Auskunft für eine vollständige Klassifikation liefern. Aufgrund der hohen Präzision von gut 85 % bei den für unter 18-Jährige gesperrten Inhal-

[8] Für die konkrete Klassifikation wird eine Support Vektor Maschine (SVM, siehe Abschnitt 4.4.2.5) mit der Radialen Basisfunktion als Kernfunktion verwendet, da dieses Klassifikationsverfahren sich in der Regel bestens für die beschriebene Problemstellung eignet. Der SVM-Parameter c wird jeweils auf den Trainingsdaten jedes einzelnen Kreuzvalidierungsschritts per verschachtelter Kreuzvalidierung optimiert.

[9] In diesem Experiment wurde Boilerpipe (http://code.google.com/p/boilerpipe/, Stand: November 2012) verwendet.

Tabelle 5.9
Die Klassifikationsergebnisse für das Thema Gewalt, getrennt nach Kategorien und Merkmalsgruppen. Als Gegenkategorie wird jeweils *Allgemein ab 6* verwendet. Die Spalte Prä@90 bezeichnet die Präzision bei einer Trefferquote von mindestens 90 %.

Merkmale	F-Maß	Präzision	Trefferquote	Prä@90
Gewalt ab 18				
nur URL	74,8 %	86,7 %	65,8 %	48,6 %
nur Text	76,5 %	91,8 %	65,6 %	80,2 %
Text+URL	88,4 %	94,2 %	83,2 %	82,5 %
nur Haupttext	65,8 %	82,6 %	54,7 %	40,9 %
Haupttext+URL	83,7 %	91,1 %	77,4 %	68,7 %
Gewalt ab 16				
nur URL	4,1 %	13,5 %	2,4 %	11,4 %
nur Text	49,3 %	63,4 %	40,3 %	12,4 %
Text+URL	48,9 %	56,3 %	43,2 %	15,3 %
nur Haupttext	36,6 %	53,8 %	27,8 %	9,1 %
Haupttext+URL	48,0 %	54,7 %	42,7 %	10,0 %
Gewalt ab 14				
nur URL	24,6 %	55,8 %	15,8 %	4,9 %
nur Text	52,9 %	62,7 %	45,7 %	20,6 %
Text+URL	50,9 %	56,9 %	46,1 %	19,8 %
nur Haupttext	30,0 %	44,3 %	22,7 %	6,9 %
Haupttext+URL	40,6 %	51,5 %	33,6 %	11,0 %
Gewalt ab 12				
nur URL	37,3 %	67,4 %	25,8 %	8,7 %
nur Text	53,2 %	77,0 %	40,6 %	22,2 %
Text+URL	54,4 %	76,6 %	42,2 %	32,7 %
nur Haupttext	38,7 %	56,9 %	29,3 %	12,7 %
Haupttext+URL	47,4 %	66,7 %	36,7 %	17,1 %

ten ließe sich ein solcher Filter jedoch vorschalten, da die als gesperrt erkannten Inhalte in der Regel tatsächlich jugendschutzrelevant sind.

Weitere Experimente haben zudem gezeigt, dass sich durch Verschieben der Klassifikationsschwelle (vgl. Abschnitt 4.2.1.1) bei nur gering abfallender Trefferquote die Präzision noch einmal deutlich anheben lässt. So ergibt sich eine Präzision von 96,4 % bei immerhin noch 60,8 % Trefferquote allein anhand der Webadresse. Verschiebt man die Klassifikationsschwelle in die Gegenrichtung, erhält man nur noch eine Präzision von 34,9 % bei 97,0 % Trefferquote. Dies lässt sich dadurch erklären, dass nur ein bestimmter Teil der Webseiten relevante Terme in seiner Adresse trägt. Sind diese aber vorhanden, lassen sie sich offenbar sehr sicher zur Klassifikation von gefährdenden Inhalt einsetzen. Wird die Webseite anhand ihrer Adresse hingegen als unkritisch eingestuft, sollte nicht auf eine weitergehende Analyse des Inhalts verzichtet werden. Als Ergänzung zu den aus dem textuellen Inhalt der Webseite gewonnenen Merkmalen lässt sich die Webadresse allerdings sinnvoll verwenden.

In den Ergebnistabellen wird zusätzlich die Präzision bei einer Trefferquote von mindestens 90 % als »Prä@90« dargestellt. Dadurch wird dem Umstand Rechnung getragen, dass in einigen Einsatzszenarien derartiger Filter das Underblocking, also das Übersehen von relevantem Inhalt, ein größeres Problem darstellt als das Overblocking, also das fälschlicherweise als relevant Markieren von eigentlich unbedenklichem Inhalt. Das bedeutet, dass bei einem Wert von 80 % für »Prä@90« nur eine von zehn bedenklichen Webseiten übersehen wird, während eine von fünf unbedenklichen Webseiten fälschlicherweise als relevant erkannt wird.

Der am besten trennende Klassifizierer in diesem Experiment ist der für *Gewalt ab 18* mit den Merkmalen »Text+URL«. Dieser erreicht einen »Prä@90«-Wert von 82,5 %, also werden 7 von 40 unbedenklichen Webseiten fälschlicherweise gesperrt, während nur 4 von 40 gesperrten Webseiten übersehen werden. Der »Prä@95«-Wert liegt bei 73,7 %, also immer noch in einem akzeptablen Bereich.

Die vorgestellten Ergebnisse hängen stark von der Zusammensetzung der Negativ-Klasse ab. Inhaltlich ähnliche Webseiten für unterschiedliche Altersstufen sind aus nachvollziehbaren Gründen wesentlich schwerer zu trennen als inhaltlich unterschiedliche Seiten. Das bedeutet, dass eine maschinelle Trennung der Dokumente in Altersstufen, also zum Beispiel in »geeignet ab 16« gegenüber »geeignet ab 18« sehr fehleranfällig ist. Allerdings muss dazu gesagt werden, dass dies auch menschlichen Entscheidern in vielen Fällen deutlich schwerer fällt als die Entscheidung, dass eine Webseite generell gesperrt werden sollte. Die in Tabelle 5.9 dargestellen Ergebnisse beziehen sich also grundsätzlich auf Experimente, bei denen jeweils versucht wurde, die entsprechende Zielkategorie von der Klasse *Allgemein ab 6* zu trennen.

Bei dem Versuch, zusätzlich zu *Allgemein ab 6* auch die anderen Altersstufen der Gewalt-Inhalte von *Gewalt ab 18* zu trennen, fällt das F-Maß direkt um mehr als 20 Prozentpunkte auf 61,7 % ab. Der Optimierungsalgorithmus lernt bei der Erstellung des Klassifikationsmodells also, dass bestimmte Begriffe aus dem Gewalt-Bereich nicht »ab 18« sind und stattdessen zur Negativ-Klasse gehören. Dadurch verliert der daraus resultierende Klassifizierer an Trennschärfe. Ein weiteres Experiment zeigt aber auch, dass speziell auf die Trennung zweier Altersstufen ausgelegte Modelle die beiden Klassen *Gewalt ab 16* und *Gewalt ab 18* mit einer Präzision von 81,1 % unterscheiden können. Für *Gewalt ab 12* gegen *Gewalt ab 14* wird eine Präzision von 83,6 % erreicht. Dies legt nahe, dass im praktischen Einsatz ein zweistufiger Klassifizierer eingesetzt werden sollte. Zunächst werden die potentiell gesperrten Inhalte vom allgemeinen Material getrennt und im nächsten Schritt wird über die Altersfreigabe dieser Kandidaten entschieden.

Um gute Ergebnisse zu erzielen, werden je nach Komplexität der Klassifikationsaufgabe einige hundert bis über tausend Beispieldokumente für die Zielkategorie benötigt. Da das Thema »Gewalt«

viele inhaltlich sehr unterschiedliche Unterkategorien enthält, sind einige davon stark unterrepräsentiert. Von allen *Gewalt ab 18*-Dokumenten stehen 343 zum Thema »Abtreibung« und 99 zu »Games« nur 13 zu »Vergewaltigung« und eins zum Thema »Unfall« gegenüber. Da die für diese Unterkategorien relevanten Merkmale sich kaum decken, ist für die kleinen Unterkategorien keine sinnvolle Klassifikation zu erwarten.

Dieser Effekt zeigt sich noch gravierender in den Resultaten der niedrigeren Altersstufen. Das generell schlechtere Abschneiden gegenüber den ab 18 freigegebenen Inhalten lässt sich nur teilweise damit erklären, dass die Inhalte weniger drastisch und daher schwerer von allgemein freigegebenem Material zu unterscheiden sind. Das Ergebnis von *Gewalt ab 12* mit Text und URL als Merkmale ist besser als die von *Gewalt ab 14* und auch *Gewalt ab 16*. Dies lässt neben einem Verweis auf zufällige Artefakte der Datenbasis den Schluss zu, dass schlicht die Anzahl der Trainingsdokumente den Ausschlag bei der Klassifikationsqualität gibt.

Zu Beobachten ist allerdings, dass quer durch alle Experimente die beste Repräsentation der Webseiten durch den komplett enthaltenen Text und den in der Web-Adresse enthaltenen Termen gegeben ist. Ein Entfernen des strukturellen Inhalts wie Navigationsbereich, Kopf- und Fußzeilen führt hier zu teilweise recht starken Qualitätseinbußen. Das Gruppieren der Webseiten nach Webauftritt bei der Merkmalsauswahl hat zu einer Verbesserung von bis zu 10 % im F-Maß geführt. Der beste Klassifizierer mit den Merkmalen *Text+URL* für *Gewalt ab 18* verbesserte das F-Maß von 83,2 % auf 88,4 %. Ohne diese Gruppierung wurden sehr viele Merkmale als relevant erkannt, obwohl sie nichts mit der Zielkategorie zu tun hatten.

Zusammenfassend kann man sagen, dass die Textklassifikation insbesondere für Inhalte ab 18 sehr gut funktioniert. Nur etwa jede zwanzigste unbedenkliche Webseite wird fälschlicherweise als jugendschutzrelevant eingestuft, so dass hier bei vertretbar erscheinendem Overblocking eine große Anzahl von Webseiten einfach vorsortiert werden kann, insbesondere dann, wenn schnell reagiert werden muss. Die Gründe für einen Anstieg der Qualität bei höheren Altersstufen liegen sowohl in der zunehmenden Anzahl von eindeutigen Wörtern und Kontexten (schon allein durch explizite Merkmale in der Webadresse), zum anderen aber auch in einer wesentlich größeren Anzahl an vorliegenden Daten. Die vorliegenden Trainingsdaten sind allerdings bei Weitem nicht ausreichend, um die relevanten Themenbereiche und Altersgruppen repräsentativ zu erfassen. Eine weitergehende Analyse und Trennung dieser Effekte ist bei der gegebenen Datenlage allerdings nicht möglich.

5.3.7 Zusammenfassung: Betrachtung von Verfahren im Bereich Gewalt

Wir haben uns in diesem Abschnitt mit technischen Verfahren zur Erkennung von Web-Inhalten aus dem Bereich Gewalt auseinandergesetzt. Die betrachteten Inhalte sind insbesondere durch visuelle Informationen, also Bilder und Videos, jugendschutzrelevant. Generell lässt sich sagen, dass, wie in Abschnitt 5.3.1.1 erörtert, das Spektrum von Gewalt viel zu groß erscheint, um automatisch, zum Beispiel über Konzepterkennung, erkannt zu werden. Nur für spezielle Unterkategorien scheint ein zielgerichteter Detektor sinnvoll; es sei hier auf Konzepterkennung für Ritzbilder in Abschnitt 5.5.4 verwiesen.

Auf Multimedia-Ebene lohnen sich insbesondere bei ungewöhnlich drastischen und damit etwas selteneren Darstellungen Fingerprint-Techniken. Neben dem Wiedererkennen von bereits bekanntem Material kann man die Datenbank nach und nach aufstocken, etwa, indem man bei Kompilationen mit teilweise neuem Material dieses in den Datenbestand mit aufnimmt. Sind diese Inhalte bereits altersdifferenziert klassifiziert, etwa bei Gewaltfilmen, so stellen auch unterschiedliche Schnittfassungen keine größere Hürde dar, sofern diese Fassungen einmal zum Training vorliegen. Fingerprint-Techniken auf Audio Ebene sind zuverlässig, sofern der Originalton in etwa erhalten bleibt; nicht

erfasst werden können kommentierte, neu synchronisierte oder mit neuer Tonspur unterlegte Filme. Eine Erkennung auf visueller Ebene durch Methoden des Video-Fingerprints liefern auch bei schwierigem Material immer noch hinreichend gute Ergebnisse, allerdings bei einem höheren Preis in der Ressourcenlast. Gerade bei einschlägig bekanntem Material wie etwa berühmten Horrorfilmen könnte sich der Einsatz hier aber lohnen, da eine kurze Überprüfung von Videoportalen schon eine große Anzahl an Variationen und Neuverwertungen durch Fans zeigt, die nur zum Teil aufgrund ihres Videotitels eindeutig erkannt werden können.

Vielversprechend ist die Analyse des umliegenden Textes, welche aufgrund der Präzision insbesondere in den »ab 18« Bereichen in den seltensten Fällen zu Overblocking neigt. Ein Rückschluss vom Text auf die Natur der enthaltenen Bilder ist hingegen nicht ohne weiteres möglich.

Der immer größer werdende Bereich der Let's Play-Videos ist in mehrerer Hinsicht problematisch. Zum einen stellt jedes Video einen neuen Inhalt mit wenig übereinstimmenden Merkmalen dar, zum anderen ist durch den von Nutzern zusätzlich erstellten Inhalt, hier die Kommentare, eine Jugendschutzrelevanz manchmal nur aufgrund des gesprochenen Wortes ersichtlich. Allerdings ist die Spracherkennung hier durch verschiedene Faktoren wie etwa die Umgebungsgeräusche des Spiels stark beeinträchtigt, so dass gezielt angepasste Modelle erforderlich wären und ein sinnvoller Einsatz eher fraglich ist.

5.4 Betrachtung von Verfahren zur Klassifikation im Bereich Extremismus

Wie bereits in Abschnitt 3.3.3 beschrieben, gliedert sich ein Großteil der Extremismusvideos in die Bereiche »Bericht«, »Interviews« und »Musik«. Anders als bei Gewalt ist hier das gesprochene oder geschriebene Wort von zentraler Bedeutung. Visuelle Informationen können zwar wichtig sein (etwa bei verfassungsfeindlicher Symbolik), spielen aber eher eine untergeordnete Rolle.

Auch wenn der Extremismus viele gefährliche Facetten aufweist, konzentrieren wir uns in diesem Kapitel auf rechtsextremes Gedankengut, da uns nur zu diesem Bereich Daten zur Verfügung standen. Allerdings fallen einige Videodaten aus dem Bereich Gewalt ebenfalls unter die Kategorie Extremismus (z.B. Enthauptungen in islamistischen Propagandavideos).

Für die Bewertung von technischen Verfahren im Bereich Extremismus gehen wir wie folgt vor:

- Zunächst führen wir einen qualitativen Vergleich einer geschichtsrevisionistischen Webseite und einer Informationsseite durch (Abschnitt 5.4.1).
- Dann werden wir quantitative Untersuchungen anhand von zwei Enzyklopädien durchführen. Zu diesem Zweck werden Inhalte aus einem in Deutschland indizierten rechtsextremen Wiki thematisch entsprechenden Artikeln der deutschsprachigen Wikipedia gegenüber gestellt. (Abschnitt 5.4.2).
- In Abschnitt 5.4.3 wird untersucht, inwieweit sich Webseiten mit ähnlich extremistischen Tendenzen aber unterschiedlicher Ausprägung in der Textklassifikation verhalten.
- In Abschnitt 5.4.4 untersuchen wir anhand von Stichproben, inwieweit die Blockierung von eingehenden Links seitens einer Suchmaschine bereits Aufschluss über die Natur der Webseite geben kann.
- Für den Bereich der rechtsextremistischen Musik untersuchen wir am Beispiel einer großen Datensammlung rund um die rechtsextreme Band »Skrewdriver« quantitativ Techniken, wie man die im Web auffindbaren Musikstücke einander zuordnen kann (Abschnitt 5.4.5).
- Den folgenden Abschnitt 5.4.6 widmen wir der Fragestellung, wie nachgesungene Lieder extremistischer Bands erkannt werden können.

- In Abschnitt 5.4.7 gehen wir der Frage nach, wie bestimmte strafrechtliche Symbole und Banner wiedererkannt werden können.
- Abschließend führen wir bei drei Videos zu geschichtsrevisionistischen Themen, konkret der Holocaustleugnung, eine Spracherkennung durch und prüfen die Eignung für eine Klassifizierung der Inhalte (Abschnitt 5.4.8).

Der Abschnitt endet mit einer Zusammenfassung.

5.4.1 Am Beispiel Geschichtsrevisionismus

Das erste Beispiel dieses Abschnitts beschäftigt sich mit dem Thema Geschichtsrevisionismus. Im Internet finden sich zahlreiche Webseiten, auf denen geschichtliche Tatsachen als unwahr dargestellt werden und stattdessen verfälschte Geschichtsschreibung betrieben wird.

So wird auf einigen geschichtsrevisionistischen Seiten der Holocaust infrage gestellt. Es finden sich darauf Phrasen wie »Holocaust-Lüge«, »Idiotengeschichten«, »hätten sie [..] massenweise Menschen vergasen wollen«, »Was für Beweise gibt es also« in Textnähe zu »Judenmord« und so weiter. Jede für sich genommen sagt noch nicht viel aus, aber die Anzahl und die Kombination dieser Phrasen in Konjunktiv-Formulierungen lässt sich als Indikator zur Erkennung verwenden. Interessant ist dabei auch, dass bestimmte Begriffe wie »Vernichtungslager« oder »Gaskammer« häufig in Anführungszeichen gestellt werden.

Will man eine sichere Erkennung dieser Inhalte erzielen, reicht es nicht aus, solche speziellen Formulierungen finden zu können. Bereits eine leicht geänderte Formulierung kann dazu führen, dass ein Dokument nicht entsprechend erkannt wird. Die komplexen Phrasen kommen also relativ selten vor und die einfachen Begriffe wie »Holocaust-Lüge« für sich genommen könnten auch in einem ganz anderen Kontext stehen. Es lohnt sich also ein Blick auf tiefergehehende Analyseverfahren. Dazu wäre es denkbar, verschiedene Wissensquellen zusammenzuführen, um eine Art »Verständnis« der politischen Orientierung des Autors zu gewinnen.

Dieses Vorgehen soll exemplarisch an einem konkreten geschichtsrevisionistischen Artikel demonstriert werden. Mittels eines Verfahrens zur Eigennamenerkennung werden Personennamen im Text automatisch erkannt. Im Fall des betrachteten Artikels sind es die Namen »Stefan Szende« und »Udo Walendy«.

Angenommen, die Wikipedia-Artikel zu allen Personen wären aus ihrem reinen, natürlichsprachlichen Text in eine strukturierte Form überführt und entsprechend gespeichert.[10] Im Artikel über »Stefan Szende« findet sich als erster Satz:

> »Stefan Szende war ein ... Widerstandskämpfer gegen den Nationalsozialismus.«

Demgegenüber steht im Artikel über »Udo Walendy« folgender erster Satz:

> »Udo Walendy ... gehört zu der Gruppe der Geschichtsrevisionisten, die die nationalsozialistische Kriegsschuld bestreiten.«

[10] siehe http://de.dbpedia.org/, Stand: November 2012

Mittels geeigneter Verfahren werden diese Textbausteine als Eigenschaften, also »Widerstandskämpfer gegen den Nationalsozialismus« beziehungsweise »Geschichtsrevisionist«, den entsprechenden Personen in der resultierenden Personen-Datenbank zugeordnet.

Nun werden die Personennamen im aktuell betrachteten Artikel in der Datenbank gesucht. Falls mehrere Datenbankeinträge desselben Namens vorkommen, wird üblicherweise der aufgrund des Kontexts wahrscheinlichere Kandidat ausgewählt. Dieses Verfahren nennt sich Disambiguierung, es ist also die Auflösung mehrdeutiger Bezeichnungen.

Zuletzt wird die Tendenz des Kontexts untersucht, in dem die Personen im Artikel genannt werden, um daraus Rückschlüsse auf die Haltung des Autors gegenüber der genannten Person zu ziehen. Im konkreten Fall sind es die folgenden Kontexte:

> »Wie uns der Dr. phil. Stefan Szende weismachen will ...«

und

> »Wie auch Udo Walendy in der vorzüglichen Nummer 44 der Historischen Tatsachen erwähnt ...«.

Daraus lässt sich ableiten, wie der Autor zu den Werken der zitierten Personen steht. Der Kontext der Nennung eines Namens enthält also wertvolle Informationen zur Bewertung des Textes. Dieses Vorgehen ist jedoch durch die erforderliche Verknüpfung verschiedener Technologien mit Hintergrundwissen deutlich aufwändiger als eine einfache Suche nach bestimmten Schlagwörtern.

Da die Formulierungen im Kontext der Namensnennungen sehr vielfältig sein können, ist es nicht leicht, eine hohe Trefferquote zu erzielen, also möglichst viele der Formulierungen korrekt zu erkennen. Häufig genannte Fehlerquellen sind in diesem Zusammenhang zum Beispiel nicht erkannte Negationen oder auch Satire, welche das Ergebnis direkt umkehren. Wird jedoch mehr Wert auf eine hohe Präzision gelegt, so lassen sich bereits mit einfachen Mustern die Orientierungen der dazugehörigen Aussagen erkennen, zum Beispiel mit dem genannten Muster »Wie uns *PERSON* weismachen will...«. Eine Liste solcher Formulierungen lässt sich natürlich manuell erstellen. Es wird allerdings empfohlen, auch an dieser Stelle auf maschinelle Lernverfahren zu setzen, da sich auf diese Weise eine höhere Robustheit gegenüber geringfügigen Abweichungen der Formulierungen erreichen lässt. Bestimmte Inhalte aus dem rechtsextremen Bereich lassen sich hingegen bereits mit relativ einfachen Mitteln entdecken, da sich auch hier Häufungen charakteristischer Wörter und Phrasen finden. Dies wird im folgenden Abschnitt illustriert.

Zusammenfassend kann man sagen, dass Kontextinformationen potentiell sehr wertvolle Informationen zur Identifizierung von extremistischem Gedankengut enthalten, die von Standardverfahren nicht abgedeckt werden. Eine Anpassung an die Inhalte erscheint somit vielversprechend, aber auch sehr aufwändig.

5.4.2 Textanalyse: Rechtsextremes Wiki

Anhand zweier Online-Enzyklopädien soll demonstriert werden, welche Informationen über deren Inhalt aus der Verteilung der Wörter über die einzelnen Artikel gewonnen werden können. Im Beispiel

Tabelle 5.10
Ein Vergleich der Wortverteilungen über die Artikel aus einem rechtsextremen Wiki (R) gegenüber Artikeln zum Thema »Nationalsozialismus« der Wikipedia (W).

	Verhältnis	$\neg W$	$\neg R$	W	R	Wort
01.	283,7	21 840	33 337	1	888	vsa
02.	240,0	21 841	33 850	0	375	auswahlseite
03.	213,8	21 841	33 891	0	334	weltnetz
04.	114,9	21 841	34 046	0	179	warnhinweis
05.	89,7	21 826	31 977	15	2 248	familienname
06.	89,3	21 841	34 086	0	139	altnord
07.	84,2	21 841	34 094	0	131	pluspedia
08.	74,7	21 841	34 109	0	116	nordgerm
09.	72,8	21 841	34 112	0	113	weltnetzseite
10.	66,4	21 841	34 122	0	103	rechtsrockgruppe
11.	58,1	21 840	34 044	1	181	ebenrode
12.	56,2	21 841	34 138	0	87	neuyork
13.	51,7	21 841	34 145	0	80	netzseite
14.	51,4	21 840	34 065	1	160	altgr
15.	49,8	21 841	34 148	0	77	teilvereinigung
16.	45,9	21 841	34 154	0	71	altnd
17.	45,9	21 841	34 154	0	71	brö
18.	44,0	21 841	34 157	0	68	lamellen
19.	42,1	21 841	34 160	0	65	nordger
20.	40,5	21 840	34 099	1	126	wodan

in Tabelle 5.10 werden dazu die Inhalte des deutschsprachigen Teils eines in Deutschland indizierten rechtsextremen Wikis (R) mit den Inhalten aller Artikel zum Thema »Nationalsozialismus« der deutschsprachigen Wikipedia (W) verglichen. Dabei wird für jedes Wort die relative Häufigkeit, also die Wahrscheinlichkeit, dass das Wort in einem Artikel enthalten ist, berechnet.

Für dieses Beispiel wird die Korrelation eines Wortes mit einem der beiden Korpora als Verhältnis der beiden Wahrscheinlichkeiten berechnet. Um unendliche Werte zu vermeiden, wird angenommen, dass jedes überhaupt gefundene Wort mindestens einmal auch im jeweils anderen Korpus enthalten ist. Vereinfachend wird intern zu jeder Worthäufigkeit eins addiert. Somit ergibt sich beispielsweise für das Wort »Weltnetz«, das ausschließlich im Wortschatz des rechtsextremen Wikis Verwendung findet:

$$\frac{\frac{R+1}{\neg R+R+1}}{\frac{W+1}{\neg W+W+1}} = \frac{\frac{334+1}{34\,225+1}}{\frac{0+1}{21\,841+1}} = \frac{\frac{335}{34\,226}}{\frac{1}{21\,842}} = 213,8$$

Dabei stehen R und $\neg R$ für die Anzahl der Dokumente des rechtsextremen Wikis, die das entsprechende Wort enthalten, beziehungsweise für die Anzahl der Dokumente, die das Wort nicht enthalten. Analog gilt das für W und $\neg W$ für die Anzahl der Dokumente aus der Wikipedia.

Dieses Verfahren wird als Merkmalsselektion bezeichnet. Die so gewonnenen Merkmale – in diesem Fall einfach Wörter – lassen sich in der weiteren Verarbeitung als Eingabe für statistische Verfahren verwenden, mithilfe derer sich beliebige Texte automatisch kategorisieren lassen. Üblicherweise werden für die Vorbereitung der konkreten Klassifikation andere Korrelationsmaße verwendet. Zum Beispiel wird bei der »Mutual Information«[11] auch berücksichtigt, wie häufig ein Merkmal überhaupt

[11] auf deutsch auch Transinformation, siehe http://de.wikipedia.org/wiki/Transinformation, Stand: November 2012

vorkommt. Der Informationsgehalt eines Wortes, das zwar nur in einer der beiden Klassen auftaucht, insgesamt aber auch nur drei Mal vorkommt, ist also relativ gering. Ein Wort hingegen, das in beiden Klassen mehrfach auftaucht, in der einen aber noch wesentlich häufiger, ist zur Klassifikation natürlich wertvoller.

Um die Funktionsweise zu veranschaulichen, eignet sich das einfache Verhältnis der Häufigkeiten der Wörter jedoch sehr gut. Hier wird deutlich, dass im rechtsextremen Wiki durch die konsequente Vermeidung von Anglizismen eine ganz eigene Terminologie entsteht. Anstatt also wie üblich die »Vereinigten Staaten von Amerika« mit »USA« abzukürzen, wird hier die deutsche Variante »VSA« verwendet. Der Begriff »Teilvereinigung« bezeichnet die deutsche Wiedervereinigung, die nach Ansicht der Autoren dieses Wikis nicht vollständig stattgefunden hat. Selbsterklärend sind erscheinende Begriffe wie »Weltnetz« oder »Neuyork«.

Besser als ein Vergleich zweier Webauftritte ist es, wenn die beiden Kategorien *extremistisch* und *nicht extremistisch* möglichst viele verschiedene Webauftritte umfassen. Dadurch lassen sich stärker spezielle Wörter der jeweiligen Auftritte vom eigentlichen Inhalt der Zielkategorie trennen. In diesem Fall finden sich neben einigen Abkürzungen zum Beispiel die Begriffe »Warnhinweis« und »Familienname« als stark korreliert mit dem Inhalt des rechtsextremen Wikis. Solche Verzerrungen entstehen dadurch, dass immer auch einige Begriffe in der Struktur und nicht dem eigentlichen Inhalt eines Webauftritts berücksichtigt werden. Hier sind es zum Beispiel bestimmte Formulieren in Biographien, die häufig das Wort »Familienname« enthalten.

Zwar lassen sich automatisch weitgehend vollständig die Navigationselemente und einige andere strukturelle Inhalte von Webseiten entfernen, spezielle Formulierungen sind allerdings nur recht schwierig und damit fehleranfällig von inhaltlich relevanten Wörtern zu trennen. Leider ist es zeitaufwändig und damit teuer, diese manuell kategorisierten Kollektionen von Webseiten in entsprechender Größe zu erhalten. Daher wurde für diese Auswertung vereinfachend auf die Trennung durch die beiden Webauftritte zurückgegriffen.

Zusammenfassend kann man sagen, dass auch bei thematisch sehr ähnlichen Textsammlungen mit unterschiedlicher politischer Orientierung eine Differenzierung gut möglich ist. Für das vorliegende Experiment war insbesondere das verwendete Vokabular, also etwa der konsequente Verzicht von Lehnwörtern aus dem Englischen, für eine gute Qualität der Merkmalsextraktion entscheidend. Allerdings sind bei Textinhalten, die ähnliches Vokabular verwenden, auffällige Häufungen charakteristischer Phrasen trotzdem zu erwarten. Dies erfordert jedoch eine größere Anzahl an Dokumenten, da Phrasen weniger gehäuft auftreten als einzelne Wörter. Das bedeutet: je ähnlicher zwei Dokumentklassen sind, desto mehr Dokumente werden für eine erfolgreiche Textklassifikation benötigt.

5.4.3 Textklassifikation

Wir haben uns im vorangegangenen Abschnitt mit Inhalten beschäftigt, die zwar das gleiche Thema behandeln, in ihrer Ausrichtung allerdings sehr unterschiedlich sind (aufklärend beziehungsweise geschichtsrevisionistisch). Nun wollen wir die Grenze noch etwas enger ziehen und Inhalte betrachten, die beide extremistische Tendenzen haben, aber eine unterschiedliche Ausprägung aufweisen. Untersucht werden zum einen Textinhalte mit indiziertem oder strafrechtlich relevantem extremistischem Material und zum anderen Textinhalte, die zwar extremistischem Gedankengut zuzuordnen sind, aber für ältere Jugendliche zugänglich bleiben können. Diese thematisch sehr dünne Grenze äußert sich auch in der eng aneinander liegenden Altersfreigabe »ab 14« und »ab 18«.

Tabelle 5.11
Die Klassifikationsergebnisse im Bereich Rechtsextremismus, für die Zielkategorien gegenüber *Allgemein ab 6*, getrennt nach Merkmalsgruppen.

Merkmale	F-Maß	Präzision	Trefferquote	Prä@90
Rechtsextremismus ab 18				
nur URL	54,4 %	83,7 %	40,3 %	24,6 %
nur Text	82,9 %	88,6 %	77,9 %	51,0 %
Text+URL	81,8 %	86,6 %	77,5 %	60,2 %
nur Haupttext	63,9 %	73,0 %	56,8 %	12,9 %
Haupttext+URL	67,1 %	78,3 %	58,6 %	28,9 %
Rechtsextremismus ab 14				
nur URL	11,5 %	58,3 %	6,4 %	5,8 %
nur Text	55,1 %	69,0 %	45,8 %	15,9 %
Text+URL	55,4 %	68,9 %	46,4 %	17,0 %
nur Haupttext	51,2 %	70,2 %	40,2 %	8,1 %
Haupttext+URL	41,6 %	54,4 %	33,6 %	8,0 %

Die Rahmenbedingungen für die in diesem Abschnitt vorgestellten Klassifikationsexperimente entsprechen denen zum Thema Gewalt (Abschnitt 5.3.6). Die Kollektion *Allgemein ab 6* wird auch in diesem Experiment wieder als negative Klasse betrachtet. Die Zielkategorien sind definiert wie folgt:

- *Rechtsextremismus ab 18*: Eine Sammlung von 191 manuell annotierten Webseiten (z. B. indizierte Musikwebseiten oder die unkommentierte Fassung von Hitlers »Mein Kampf«).
- *Rechtsextremismus ab 14*: Eine Sammlung von 110 manuell annotierten Webseiten. Diese Webseiten sind aufgrund ihres rechtsextremistischen Inhaltes zwar eindeutig, enthalten aber keine strafrechtlich relevanten oder indizierten Inhalte (z. B. die Webauftritte von Vereinsseiten, die einem rechtsextremen Gedankengut zugeordnet werden können).

Trotz der, verglichen mit *Gewalt ab 18*, deutlich kleineren Kollektion *Rechtsextremismus ab 18* werden relativ gute Ergebnisse erzielt. Dies lässt sich damit erklären, dass das Referenzkorpus *Allgemein ab 6* nur wenige Inhalte zum Thema »Rechtsextremismus« beinhaltet. Daher ist anzunehmen, dass die Dokumente in diesem Fall weniger nach extremistischen Tendenzen getrennt werden, sondern schlicht nach dem Thema »Rechtsextremismus«, was ein wesentlich einfacheres Problem darstellt.

Benötigt würde daher ein Korpus, der aus möglichst vielen Inhalten besteht, die nicht als rechtsextrem eingestuft sind, sich aber mit der Thematik befassen. So ließe sich eine rein thematische Trennung verhindern und damit ein für die Praxis realistischeres Ergebnis erzielen.

Wie schon bei den Experimenten zum Thema Gewalt zu erkennen war, lassen sich auch in diesem Bereich allein anhand der Webadresse bestimmte Inhalte offenbar recht gut erkennen, was die hohe Präzision von über 80 % erklärt. Die geringe Trefferquote von etwa 40 % zeigt hingegen, dass die breite Masse der Webadressen nicht genug Informationen für eine Einstufung preisgibt.

Für das mit nur 55,4 % F-Maß wesentlich schlechtere Ergebnis für die Inhalte »ab 14« lassen sich zwei wesentliche Gründe nennen. Zum einen stehen mit 110 Webseiten nur sehr wenige Trainingsbeispiele zur Verfügung. Eine sinnvolle Abstraktion vom konkreten Text ist damit nur schwer möglich. Andererseits ist anzunehmen, dass die für 14-Jährige freigegebenen Textinhalte weniger explizit sind als die für 18-Jährige, was eine maschinelle Erkennung zusätzlich erschwert.

Zusammenfassend kann man sagen, dass die Präzision der Textklassifikation für den Bereich extremistische Inhalte, der für ältere Jugendliche zugänglich sein darf, nicht sehr hoch sein wird. Klare Ergebnisse sind hingegen bei eindeutigen, mitunter strafrechtlich relevanten Inhalten zu erwarten.

5.4.4 Analyse eingehender Links

Gerade im Bereich des geschichtsrevisionistischen Rechtsextremismus lohnt sich zusätzlich ein Blick auf die Linkumgebung der aktuell betrachteten Webseite. Für den folgenden Versuch wurde eine Stichprobe mit als »ab 18« bekannten Webseiten zum Bereich »Rechtsextremismus« als Gruppe »rechtsextremistisch/gesperrt« markiert. Als Vergleichsgruppe »holocaust/nicht gesperrt« werden als nicht rechtsextrem zu erkennende Webseiten aus den Suchergebnissen zu den Begriffen »Holocaust« und »Holocaust-Lüge« verwendet.

Für die Information, welche Webseite auf die aktuell betrachtete verweist, wird theoretisch der komplette Webgraph benötigt (vgl. Abschnitt 4.6). In diesem Experiment wird stattdessen auf die Suchfunktion von Google zurückgegriffen, bei der sich die eingehenden Verweise mit der Syntax »link:URL« suchen lassen. Interessant ist dabei, dass aus dem Google-Suchergebnis automatisch bei »ChillingEffects.org« als illegal gemeldete Inhalte entfernt werden.

Tabelle 5.12 zeigt für jede Webseite beider Gruppen jeweils die Anzahl der als »frei« und »illegal« gekennzeichneten eingehenden Verweise. Aufgeführt werden nur die URLs, zu denen es laut Google überhaupt eingehende Verweise gibt. Trotz der relativ kleinen Stichprobe geht aus der Tabelle deutlich hervor, dass die Anzahl der eingehenden Verweise von als illegal bekannten Webseiten ein sinnvolles Kriterium bei der Einstufung der aktuell betrachteten Webseite ist.

Für ein automatisches Ermitteln dieses Merkmals sind allerdings einige Hürden zu nehmen, da Google es nicht erlaubt, die Suchergebnisse maschinell auszuwerten. Zunächst wird eine Programmierschnittstelle (API) zu einer Suchmaschine benötigt, um die eingehenden Verweise automatisch zu erhalten und nicht für jede Webseite händisch zu überprüfen. Derzeit (Stand: November 2012) unterstützen die wesentlichen Suchmaschinen jedoch diese Funktion entweder nicht oder sind kostenpflichtig. Auch steht derzeit die URL-Überprüfungsfunktionalität von »ChillingEffects.org« nicht zur Verfügung und kann daher nicht praktisch nachvollzogen werden.

Zusammenfassend lässt sich sagen, dass die Analyse eingehender Links prinzipiell sehr sinnvoll erscheint und ein mächtiges Werkzeug bei der Filterung jugendschutzrelevanter Inhalte sein kann. Die Abhängigkeit von externer Funktionalität kann dabei allerdings, auch unter finanziellen Gesichtspunkten, eine Praxistauglichkeit erschweren.

5.4.5 Musikklassifikation: Identifikation bekannter extremistischer Lieder

In diesem Abschnitt soll der Frage nachgegangen werden, wie gut man Lieder wiederfinden kann, wenn sie bekannt sind und in einer Datenbank verwaltet werden. Als untersuchtes Material verwenden wir hierbei Lieder aus dem Umfeld der rechtsextremen Band Skrewdriver.

Durch die Gründung des in Deutschland verbotenen[12] rechtsextremen Netzwerks »Blood & Honour« erlangte die Band internationale Bekanntheit. Die erste Veröffentlichung der Band, »All skrewed up«, ist weitestgehend unpolitisch und in Deutschland zum Beispiel über große Onlineversandhandel legal bestellbar (Stand: September 2012). Im Gegensatz dazu wurde zum Beispiel das Album »Hail the new Dawn« indiziert.[13]

Für einen Test von Audio-Fingerprinting hat das Fraunhofer IAIS einen umfangreichen Datensatz mit rechtsextremen Musikstücken verwendet. Insgesamt umfasst die Sammlung 523 Alben und 6 631

[12] Quelle: http://www.aktion-zivilcourage.de/Braune_Symbole.472/, Stand: 25.9.2012
[13] Bundesanzeiger 63, vom 30.3.2007

Tabelle 5.12
Eine Stichprobe zur Frage, wie viele nach »ChillingEffects.org« als illegal bekannte Inhalte auf gesperrte rechtsextreme Inhalte verweisen.

Gruppe 1: rechtsextremistisch/gesperrt		Gruppe 2: holocaust/nicht gesperrt	
Anzahl eingehender Verweise		Anzahl eingehender Verweise	
frei	illegal	frei	illegal
16	4	18	0
1	0	337	0
18	5	37	0
145	4	1 520	0
1	1	4	0
17	11	1	0
98	11	2	0
10	6	14	0
1	0	13	0
2	0	76	0
2	8	25	0
0	2	10	0

Lieder mit einer Gesamtlänge von knapp 400 Stunden. Jedes Album enthält im Schnitt 13 Lieder. 3 000 Lieder wurden benutzt, um mittels der frei verfügbaren Software Echoprint [Ellis et al., 2011] einen Audio-Fingerprint-Server aufzusetzen. Im Anschluss wurde über den Server getestet, inwieweit die 3 000 Lieder wiedergefunden werden. Anhand von weiteren 3 000 Liedern, die allerdings nicht in den Server eingespeist wurden, wurde überprüft, wie exakt das Verfahren arbeitet. Die Ergebnisse sind in Tabelle 5.13 dargestellt.

Die Verfahren liefern die erwarteten Ergebnisse. Zwar ist eine große Anzahl von Liedern als Treffer markiert worden, obwohl sie eigentlich nicht in der Datenbank vorkam, doch diese Werte sind irreführend, da sich auch einige Lieder aus Sammlungen, »Best-Of« oder »Tribute« darunter befanden und die Lieder damit auch in der Sammlung doppelt vorkommen konnten. So enthalten von den 348 als »false positive« gemeldeten Liedern knapp 100 den Begriff »tribute«, »sammler«, »sampler« oder »collection« schon im Titel.

Wenn man die Fingerprints auf die Länge von 30 Sekunden beschränkt, funktioniert das Verfahren deutlich schneller; hier brauchte die Abfrage aller 6 000 Lieder etwa 2 Stunden auf einem handelsüblichen Laptop, wohingegen die Verarbeitung der vollständigen Lieder etwa 6 Stunden benötigte (bei 400 Stunden Originalmaterial). Was bei den Tabellen noch auffällt ist, dass sich die einzelnen Zeilen nicht jeweils auf 3 000 aufaddieren. Durch einen Fehler in der eingesetzten Software ist die Verarbeitung zum Teil abgebrochen, was zu fehlenden Ergebnissen bei einigen Liedern geführt hat. Die Werte in der Tabelle sind daher nur als Tendenzen zu interpretieren.

Bei den Experimenten zu beachten ist, dass diese Fingerprinting Methode insbesondere auf große, zusammenhängende Segmente schaut und weniger kleinere Ausschnitte beurteilt. Sie unterscheidet sich damit wesentlich von der Audio-Fingerprinting Methode, die bei der Wiedererkennung für den Film Braindead eingesetzt wurde (vgl. Abschnitt 5.3.2). Das hat hier insbesondere den Vorteil, dass Aufklärungsinhalte mit kurzen Segmenten dieser Lieder nicht erkannt werden, sondern lediglich solche, die die Lieder unkommentiert in voller Länge abspielen.

Zusammenfassend kann man sagen, dass Audio-Fingerprinting im Bereich der Musikklassifikation einen erheblichen Mehrwert bei der Identifizierung von extremistischen Inhalten darstellt.

Tabelle 5.13
Audio-Fingerprint Experiment auf rechtsextremen Liedern.

(a) Ergebnisse der Erkennung, bei einem Fingerprint der Länge 30 s

	als Treffer markiert	nicht als Treffer markiert
enthalten	2597	242
nicht enthalten	393	2432

(b) Ergebnisse der Erkennung, bei einem Fingerprint über die gesamte Länge der Datei

	als Treffer markiert	nicht als Treffer markiert
enthalten	2355	119
nicht enthalten	348	2087

Tabelle 5.14
Ausgewählte Coverlieder für die Musikklassifikation.

Idx	Group	Band	Anmerkung
1	1	Skrewdriver	Original
2	1	Skrewdriver	Demo-Tape
3	1	Skrewdriver	Live-Fassung, sehr schlechte Tonqualität
4	1	Englische Band	Coversong, sehr grölend
5	1	Amerikanische Band	Coversong
6	1	Schwedische Band	Coversong, weibliche Sängerin
7	1	Schwedische Band	Coversong, weibliche Sängerin, relativ melodisch
8	1	Britische Band	Coversong, recht nah am Original
9	2	Landser	Original
10	2	Deutsche Band	Parodie, unterschiedlicher Text, fehlendes Gitarrensolo

5.4.6 Musikklassifikation: Live-Versionen, Cover und Parodien

Die beschriebenen Daten enthielten neben dem oben erwähnten doppelten Auftreten von Liedern, etwa in Samplern, auch eine ganze Reihe an Cover-Versionen. Wir wollen in diesem Abschnitt der Frage nachgehen, inwieweit diese Lieder von Fingerprinting erfasst werden oder ob es andere Techniken gibt, die diese Lieder identifizieren können. Die folgenden Experimente sind in Zusammenarbeit mit der Universität des Saarlandes und den International Audio Laboratories Erlangen entstanden.

Für eine exemplarische Bewertung des Konzepts haben wir zwei strafrechtlich relevante Lieder gewählt: zum einen das Lied »Hail the new Dawn« von Skrewdriver. Von diesem Skrewdriver Lied kursieren einige Fassungen der Band selbst, etwa als Demo-Aufnahme oder als Live-Version, und eine ganze Reihe von unkritisch nachgesungenen Coverversionen von einschlägig bekannten Bands. Als zweites Lied haben wir »Landser, Landser« der Band Landser gewählt, das insofern interessant erscheint, weil es eine satirische Fassung davon gibt, die den Ton dieser Band nachahmt, aber die Texte durch harmlosen Unsinn ersetzt und somit der Lächerlichkeit preisgibt – so wird aus »Brave Bürger sind blind, aber ihr könnt es seh'n/Das Deutsche Reich wird wieder aufersteh'n« (Landser) in der Satirefassung die Textzeile »Mütter mit Kindern woll'n nach draußen geh'n / Um 'nen schönen Teich mit Pflanzen zu sehn«. Tabelle 5.14 fasst die Liederauswahl zusammen.

Abbildung 5.11(a) zeigt die Ergebnisse von pHash [Zauner, 2010] Audio-Fingerprints auf den zehn Aufnahmen. Laut dieser Technik weist keine der Aufnahmen eine Ähnlichkeit zu einer anderen auf. Abbildung 5.11(b) hingegen zeigt die Ähnlichkeit der Aufnahmen, wenn man über Chroma-Merkmale und andere musikbezogene Merkmale vergleichen lässt (s. [Müller et al., 2005,

Serrà et al., 2008]). Hier zeigt sich, dass insbesondere die anderen Versionen der Band selber gut identifiziert werden können. Schwieriger wird es für die Aufnahmen 5 beziehungsweise 6, die von den jeweiligen Sängern stärker neu interpretiert wurden. Die Parodie von Landser wird nur als schwach ähnlich zum Original angesehen und weist naheliegenderweise noch weniger Ähnlichkeit zu den Skrewdriver Liedern auf.

Erneut betont sei hier der exemplarische Charakter des Cover-Experiments. Die Anfrage auf 10 Liedern ist recht spärlich für generalisierende Aussagen und unterschiedliche Parameter der Verfahren könnten die Ergebnisse noch einmal stark beeinflussen. Experimente mit anderen Einstellungen[14] zeigen zum Beispiel bessere Ergebnisse für die »Landser« Gruppe, also eine höher bewertete Ähnlichkeit, allerdings sinkt dadurch die Ähnlichkeitsbewertung für die Skrewdriver Cover. Gut erkannt werden Lieder insbesondere dann, wenn sie ungefähr der harmonischen Struktur des Originalliedes folgen; die Passagen dafür sollten hier mindestens 40 bis 60 Sekunden lang sein. Gibt es keinen oder nur sehr untergeordneten harmonischen Inhalt im Lied oder wird die Melodie stark variiert, wie es bei Musikrichtungen wie Punk oder Hard Rock der Fall sein kann, stoßen diese Verfahren bisweilen an ihre Grenzen (vgl. [Serrà et al., 2008, Serrà et al., 2009]).

Neben einer unmittelbaren Bewertung von unterschiedlichen Versionen eines Liedes ist auch ein Vergleich zu anderem Audiomaterial wichtig, das nicht in unmittelbarem Bezug zu den Cover-Aufnahmen steht. In einem zweiten Experiment haben wir daher die zehn Cover-Aufnahmen mit anderen Aufnahmen einer Musikdatenbank vermischt. Ziel ist es nun, die gewünschten Cover-Aufnahmen anhand einer Referenzaufnahme automatisiert zu identifzieren. Hierzu haben wir ein System zur Coversong-Identifikation implementiert, das in weiten Teilen dem in [Serrà et al., 2008, Serrà et al., 2009] beschrieben Verfahren folgt. Es ist zu erwarten, dass die Identifikation bei einer Vergrößerung des Datenbestand schwieriger wird. Um diesen Effekt aufzeigen zu können, haben wir drei Datenbestände mit 112, 340 und etwa 2 000 Aufnahmen zusammengestellt, die jeweils die 10 Cover-Aufnahmen beinhalten. Mit Hilfe des Identifikationssystems wurden dann diese Aufnahmen entlang absteigender Ähnlichkeit in Bezug auf eine Referenzaufnahme sortiert.

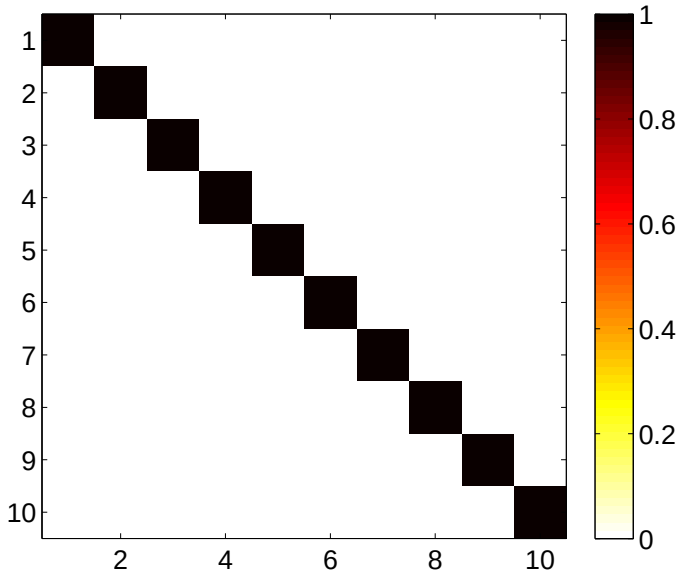
Die Ergebnisse sind in Tabelle 5.15 dargestellt. Während die durchschnittliche Genauigkeit für das Originallied sowie die Demofassung unverändert bleibt, nimmt sie für die Cover-Versionen bei steigender Datensatzgröße stetig ab. Abbildung 5.12 stellt das Ranking der einzelnen Lieder für den Bestand mit 112 Aufnahmen graphisch dar.

Zusammenfassend kann man sagen, dass Cover-Erkennung die Suche nach relevanten Aufnahmen insbesondere bei Live- und Demoversionen wesentlich erleichtern kann. Von anderen Bands nachgesungene Lieder mit jugendschutzrelevantem Inhalt können gut erkannt werden, wenn sie sich stark am Ursprungsmaterial orientieren. Neuinterpretationen mit veränderten Musikstrukturen sind hingegen schwer zu erfassen, was aber bei Parodien auch wünschenswert sein kann.

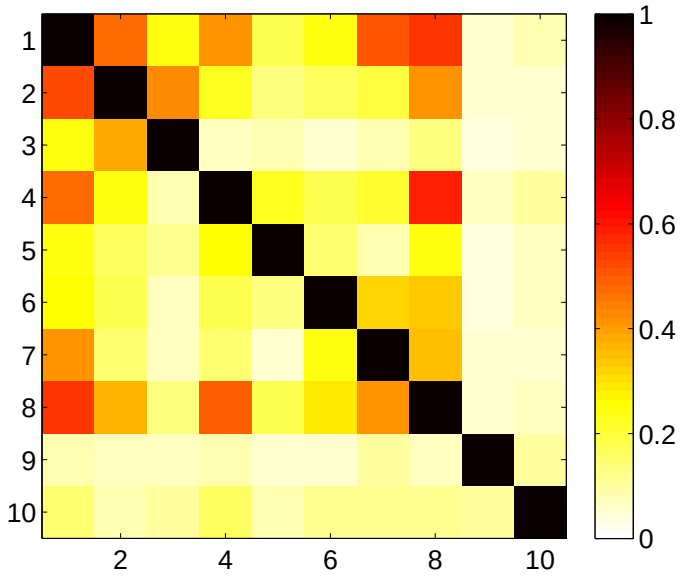
5.4.7 Logoerkennung: Objekte und Symbole

Beim Wiedererkennen von einschlägigen Symbolen der extremistischen Szene wie etwa dem Hakenkreuz oder dem Logo der in Deutschland verbotenen Gruppierung »Blood & Honour« ist es zunächst ein technischer Unterschied, ob das gesamte Logo unverändert bleibt (Objekt-Wiedererkennung), etwa bei gleichbleibenden Fahnen, oder ob eine bestimmte Form wiedererkannt werden soll (Symbolerkennung), etwa bei Graffiti von Hakenkreuzen an Wänden. Die zu erwartende höhere Genauigkeit der Objekt-Wiedererkennung und der geringe Trainingsaufwand steht hierbei der höheren

[14] unter Verwendung von transpositionsinvarianten Matrizen wie in [Müller und Clausen, 2007], bei längeren Glättungsfiltern.



(a) Ähnlichkeitsmatrix für Coverexperiment, unter Benutzung von Audiofingerprints von pHash. Die Ergebnisse sind auf den Bereich 0 bis 1 normiert.



(b) Ähnlichkeitsmatrix für Coverexperiment, unter Benutzung von Liedähnlichkeitsmerkmalen.

Abbildung 5.11
Ergebnisse der Klassifikation bei Coverliedern.

Tabelle 5.15
Durchschnittliche Präzision (*Average precision*; AP) und deren Mittelwert auf alle Anfragen (*mean of average precision*; MAP), für die 10 Liederanfragen bezüglich der drei Datensätze.

Idx	Cover Gruppe	AP_{112}	AP_{340}	AP_{2000}
1	1	1	1	1
2	1	1	1	1
3	1	0.831	0.789	0.620
4	1	0.982	0.924	0.883
5	1	0.948	0.929	0.878
6	1	0.891	0.875	0.860
7	1	0.916	0.849	0.777
8	1	1	1	1
9	2	0.167	0.044	0.009
10	2	0.125	0.050	0.015
MAP		0.786	0.746	0.704

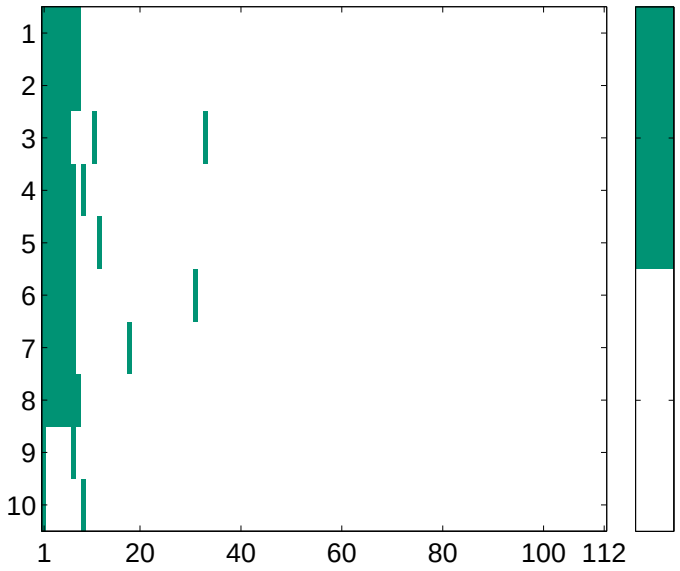


Abbildung 5.12
Bewertung von Ähnlichkeiten für 10 Lieder im rechtsextremen Kontext, verglichen mit 102 unpolitischen Liedern. Farbige Einträge zeigen die Nähe zu relevanten Liedern, weiße Einträge stehen für nicht relevante Lieder aus den anderen Datensätzen.

Flexibilität der Symbolerkennung gegenüber und muss entsprechend der Zielsetzung der Anwendung abgewogen werden.

Die Techniken in diesem Bereich sind gut erforscht und für beide Arten der Anwendung ausgereift. In diesem Abschnitt zeigen wir exemplarisch die Ergebnisse für beide Methoden.

Für die Objekt-Wiedererkennung auf »Blood & Honour« greifen wir hier auf das Know-How des griechischen Forschungsinstituts *Information Technology Institut* (CERTH-ITI) zurück.[15] Die Ergebnisse sind in Abbildung 5.13 dargestellt. Als Suchanfrage verwenden wir das Gruppenlogo, das in Wikipedia vorliegt 5.13(a), und als Beispielbilder haben wir die ersten Treffer einer Bildanfrage auf Google genommen, die auf den Suchterm »Blood & Honour« angezeigt werden. Der Algorithmus reagiert recht robust auf verschiedene Herausforderungen wie etwa hochgehobene Banner (Abb. 5.13(c), 5.13(g)), auch dann, wenn das Transparent stark reflektiert. Dadurch wird auch ein Großteil der gefundenen Rockkonzert-Bilder mit diesem Banner im Hintergrund identifiziert. Ebenfalls gefunden werden Artikel aus Online-Shops, also etwa T-Shirts (Abb. 5.13(b)) oder CD-Cover (Abb. 5.13(e)) – solange das Symbol groß genug ist (vergleiche Abb. 5.13(d)). Problematisch wird es aber zum Beispiel dann, wenn der Inhalt des Banners stark variiert, etwa durch eine andere Schriftart oder ein nachgezeichnetes Logo (Abb. 5.13(f)).

Für die Symbolerkennung von Hakenkreuzen verwenden wir einen hauseigenen Detektor. Exemplarische Ergebnisse sind in Abbildung 5.14 dargestellt. Positive Treffer gibt es für Hakenkreuze auf Fahnen und T-Shirts (Abb. 5.14(a), 5.14(b), 5.14(d), 5.14(e)), sofern diese nicht zu stark gekrümmt sind (Abb. 5.14(c)). Auch bei Tätowierungen funktioniert die Erkennung gut (Abb. 5.14(f)). Zu beachten ist in diesem Fall aber, dass auch durchgestrichene Hakenkreuze, die gemeinhin als nicht verboten gelten, von dem Algorithmus als positiver Treffer gewertet werden (Abb. 5.14(d)).

Das Verfahren selbst basiert auf einer vergleichsweise alten Technik,[16] liefert aber wie gesehen robuste Resultate. In der Grundeinstellung wurden jedoch zunächst recht viele falsch positive Treffer gemeldet (insbesondere bei monochromatischem Hintergrund). Durch eine Verschärfung der Schwellwerte konnte das Ergebnis verbessert werden. Allein auf dem T-Shirt in Abb. 5.14(d) findet sich noch eine derartige Fehlklassifikation direkt unterhalb des durchgestrichenen Hakenkreuzes. Ein weiterer gravierender Nachteil des Verfahrens ist ein Mangel an Robustheit gegenüber Rotation; für den verwendeten Detektor konnten nur Hakenkreuze erkannt werden, die in einem 45° Winkel zur X-Achse, also auf der Spitze, standen.

Zusammenfassend kann man sagen, dass die Erkennung der meisten Logos und Symbole bei vorliegendem Verfahren unproblematisch ist und wenig Aufwand bedeutet. Neben der Erkennung von extremistischen Symbolen erscheint vor allem das Erkennen von Logos einschlägig bekannter Gewaltwebseiten (vgl. Abschnitt 5.3.1) interessant.

5.4.8 Spracherkennung: Holocaust-Leugnung

Die Leugnung des Holocaust ist in vielen Ländern, darunter in allen deutschsprachigen Ländern, illegal.[17] In Deutschland greift hier § 130 des Strafgesetzbuches (Volksverhetzung). In diesem Abschnitt wollen wir untersuchen, inwieweit die wichtigen Schlagwörter, die auf die Leugnung hindeuten, in

[15] Verwendet werden hier SURF Deskriptoren [Bay et al., 2008] und einige geometrische Restriktionen, die durch den RANSAC Algorithmus [Fischler und Bolles, 1981] gegeben sind.

[16] Viola–Jones object detection framework, [Viola und Jones, 2001].

[17] Eine Übersicht über die verschiedenen nationalen und das europäische Gesetz bietet die Wikipedia, siehe http://de.wikipedia.org/wiki/Gesetze_gegen_Holocaustleugnung, Stand 30.11.2012



(a) Bildquelle: http://upload.wikimedia.org/wikipedia/de/c/c7/B%26h.JPG, abgerufen am 29.11.2012 im deutschsprachigen Artikel »Blood and Honour«.



(b) Aus einem englischen Online-Versand für rechtsextreme Fanartikel.



(c) Aufmarsch von Rechtsradikalen. Das Logo kann hier gut erkannt werden.



(d) Verbotenes Propagandamaterial, ausgestellt vom Landeskriminalamt Baden-Württemberg. Für eine gute Logoerkennung sind die Objekte zu klein.



(e) CD-Cover eines Samplers mit rechtsradikalen Liedern. Trotz einer anderen Hintergrundfarbe wird das Logo gut erkannt.

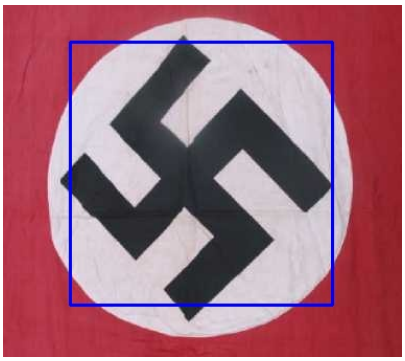


(f) Flyer für ein rechtsradikales Musikkonzert in Belgien. Die Schriftart und das Symbol sind so stark verändert, dass sie nicht erkannt werden.



(g) Bild einer Tauschbörse. Selbst bei ungünstigen Lichtverhältnissen und reflektierendem Transparent wird das Logo erkannt.

Abbildung 5.13
Erkennungsergebnisse auf dem Blood & Honour Logo. Die Verwendung von Schriftzug und Symbol ist in Deutschland seit dem Verbot der »Blood & Honour – Division Deutschland« im September 2000 strafbar (§ 86a StGB), d. h. die Verbreitung im Inland ist außer zu Zwecken z. B. der Wissenschaft bzw. der Aufklärung verboten.



(a) Fahne mit Hakenkreuz.



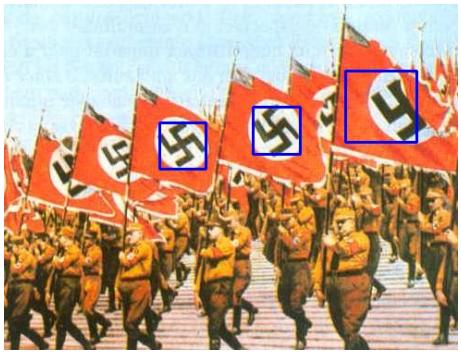
(b) Historisches Bild eines Naziaufmarsches. Eins der Hakenkreuze kann gut erkannt werden.



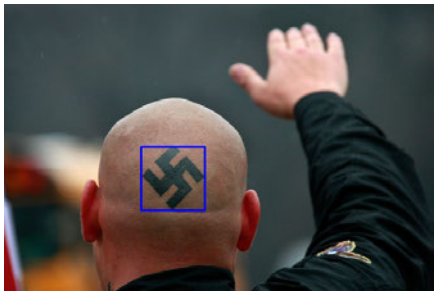
(c) Auf dieser historischen Aufnahme einer Naziflagge ist das Symbol zu stark gestreckt, um vom Erkenner erfasst zu werden.



(d) Auch erlaubte Symbole wie etwa das durchgestrichene Hakenkreuz werden als Treffer gerechnet.



(e) Auch bei teilverdeckten Hakenkreuzen funktioniert die Erkennung gut.



(f) Erkanntes Hakenkreuz bei einem Tattoo auf dem Hinterkopf. Hier hätte eine striktere Objekt-Wiedererkennung Schwierigkeiten.

Abbildung 5.14
Erkennungsergebnisse für den Hakenkreuz-Detektor. Die Verwendung des Symbols ist durch den eindeutigen Bezug zum Nationalsozialismus in Deutschland nach § 86a StGB strafbar.

einem Video automatisch gefunden werden können. Als Technik verwenden wir hierfür einen deutschen Spracherkenner wie in [Schneider et al., 2008] beschrieben.

Als Material dienen uns Videos von Horst Mahler, Sylvia Stolz und Ernst Zündel. Horst Mahler (*1936) ist ein ehemaliger Rechtsanwalt und bereits mehrfach wegen Volksverhetzung, Terrorismus und Raub verurteilt, zuletzt 2009 für sechs Jahre wegen Holocaust-Leugnung [Meiritz, 2009]. Auch seine Lebensgefährtin Sylvia Stolz (*1963) leugnet die gezielte Ermordung von sechs Millionen Juden während der Zeit des dritten Reichs und wurde 2009 für 3 Jahre verurteilt [Jüttner, 2009]. Von beiden Personen ist eine große Anzahl an Interviews und Reportagen im Internet zu finden. Wir verwenden jeweils die ersten 15 Sätze aus einem dieser Beiträge, um die Qualität einer automatischen Spracherkennung zu messen. Als drittes Video verwenden wir das von Ernst Zündel (*1939) produzierte Video zum zweiten Leuchter-Report, einer geschichtsrevisionistischen Hinterfragung des Holocaust, die als widerlegt gilt.[18]

Die Ergebnisse der Erkennung finden sich in Tabelle 5.16. Die Spracherkennung funktioniert aufgrund ungünstiger Umstände (Hintergrundgeräusche, starker Akzent bei Frau Stolz, schlechte Tonqualität) nicht fehlerfrei. Dennoch werden wichtige Keywords wie »Fremdherrschaft«, »Auschwitz« und »Gaskammer« gut erkannt. Die Analyse der Daten hat etwa einen Echtzeitfaktor von 2, das heißt für den zweiten Leuchter-Report (Dauer: circa eine Stunde) wurden insgesamt zwei Prozessorstunden verwendet.

Zusammenfassend kann man für diese drei Videos konstatieren, dass das Ergebnis im Gegensatz zu den Erkennungsraten für Let's Play Gewaltspiele (vgl. Abschnitt 5.3.4) noch ausreichend dazu verwendet werden kann, um eindeutige Passagen zu finden oder automatisch auf Abschnitte mit gegebenenfalls relevantem Inhalt hinzuweisen. Inwieweit eine automatische Erkennung aller Verdachtsvideos praxistauglich ist, kann im Rahmen dieser Stichprobe allerdings nicht abschließend bewertet werden.

5.4.9 Zusammenfassung: Betrachtung von Verfahren im Bereich Extremismus

Zusammenfassend lässt sich sagen, dass Extremismus in Text und Multimedia mit kleinen Einschränkungen relativ leicht zu erkennen ist. Erneut angemerkt sei hier, dass wir uns in den vorangegangenen Experimenten mit Rechtsextremismus beschäftigt haben und zum Beispiel zu Aussagen über Gewaltvideos aus dem islamistischen Kontext auf den Abschnitt zu Gewalt verweisen.

Schon textlich ist aufgrund der verwendeten Vokabeln schnell eine Unterscheidung zwischen verherrlichenden und aufklärerischen Inhalten möglich. Auch kann eine redaktionelle Klassifikation deutlich erleichtert werden, da automatisch auf problematische Inhalte wie etwa Symbole und Logos hingewiesen werden kann. Mit automatischer Spracherkennung können Passagen mit entscheidenden Wörtern wie etwa »Gaskammer« suchbar gemacht und so leichter redaktionell gefunden und eingestuft werden. Auch bei Musik greifen die Verfahren gut: mit einer geeigneten Datenbank ist das Erkennen von bekannten Liedern mit strafrechtlich relevanten Inhalten unproblematisch. Durch Cover-Erkennung kann eine Vorsortierung vorgenommen werden, so dass auch nachgesungene Lieder schneller redaktionell identifiziert werden können.

[18] Vergleiche z. B. http://www.lehrer-online.de/363569.php, abgerufen am 29.11.2012

Tabelle 5.16
Ergebnisse der Spracherkennung auf Holocaust-Leugnungsvideos.

Name	Korrekte Wörter	Beispielsatz
Horst Mahler	103 von 253 (40,7 %)	Die Fremdherrschaft kann nur durch Täuschung der Betroffenen sich an der Macht halten. Zur Täuschung gehört, dass man uns vormacht, die Bundesrepublik ist ein Staat, es ist der deutsche Staat, und dass die Justizbehörden deutsche Justizbehörden sind, deutsche Gerichte. Aber was sie in Wahrheit tun: sie vollstrecken den Befehl der Fremdherrschaft.
Leuchter Report	72 von 242 (32,6 %)	Nach weiteren Studien und Besprechungen in Toronto, bei dem Leuchter mit Blueprints, Dokumenten und Luft aufnahmen und Anderem konfrontiert wurde, akzeptierte Leuchter einen Vertrag über eine Untersuchung einzugehen, die ihn nach Polen bringen sollte, um die als Gaskammern ausgewiesenen Gebäude in Auschwitz, Birkenau und Majdanek in Augenschein zu nehmen.
Interview Sylvia Stolz	81 von 173 (46,8 %)	Hitler und die Nationalsozialisten gelten als böse, als Teufel. Sie haben ein paar gute Dinge gemacht, aber prinzipiell sind sie als Verbrecher anzusehen. So ist, so denken sehr viele Jugendliche und natürlich auch sehr viele Erwachsene. Wie hat das alles angefangen? Diese Frage stellt sich.

Legende:
ohne Farbe = nicht erkannt
mit Farbe = von der Spracherkennung erkannt

5.5 Betrachtung von Verfahren zur Klassifikation im Bereich Selbstgefährdung

Wie bereits in Abschnitt 3.4.3 beschrieben, gliedern sich die Daten in die Bereiche Pro-Ana/Pro-Mia, selbstverletzendes Verhalten, Anleitung zum Selbstmord sowie Verherrlichung von Rauschzuständen und gefährliche Handlungsanregungen.

Für die Bewertung gehen wir wie folgt vor:

- Zunächst führen wir einen qualitativen Vergleich von zwei Videos mit selbstgefährdendem Inhalt zum Thema Pro-Ana und zum Thema Ritzen durch (Abschnitt 5.5.1).
- Dann bewerten wir quantitativ anhand von einigen Videos aus dem Bereich »Ritzen«, wie häufig hier auf (urheberrechtlich geschützte) Lieder als Untermalung der selbsterstellten Inhalte zurückgegriffen wird und sich somit eine geeignete Datenbank zum Abgleich anbietet (Abschnitt 5.5.2).
- Die Form von selbst eingestellten Texten, sowohl mit als auch ohne Hintergrundbild, deckt eine große Bandbreite ab. In diesem Teilabschnitt untersuchen wir exemplarisch einfache und herausfordernde Texterkennungsaufgaben auf ihre Machbarkeit (Abschnitt 5.5.3).
- Eine Hauterkennung ist einfach möglich. Ob es sich bei gezeigten Hautpartien aber um Tattoos oder um Schnittwunden handelt, scheint bei Sichtung des Materials schwierig. Wir gehen der Frage nach, inwieweit hier eine semantische Konzepterkennung auf Bildebene funktionieren kann (Abschnitt 5.5.4).
- Bei Bildinhalten, die nicht von Nutzern selbst erstellt sondern lediglich abgewandelt werden, untersuchen wir dann, wie robust ein Fingerprint-Verfahren auf diese Veränderungen reagiert (Abschnitt 5.5.5).
- Daran anschließend bewerten wir ausgehend von typischen Textinhalten von Anorexie-verherrlichenden Foren, inwieweit sich mit wenig Trainingsmaterial relativ sicher entsprechende Inhalte identifizieren lassen können (Abschnitt 5.5.6).

Der Abschnitt endet mit einer Zusammenfassung.

5.5.1 Am Beispiel Ritzen und Pro-Ana

Betrachtet wurden zwei Videos aus dem Bereich Selbstgefährdung, die auf den ersten Blick ähnlich aufgebaut zu sein scheinen. In beiden wird nicht gesprochen, sondern durchgängig jeweils ein Musikstück wiedergegeben. Beide sind ein Aneinanderfügen von Bildern mit selbstgefährdendem Inhalt. In einem Fall handelt es sich um Bilder eines durch einen scharfen Gegenstand aufgeritzten Unterarms, in den Botschaften geschrieben wurden (im Folgenden *Ritzvideo*), im anderen Fall wird eine Ansammlung von stark abgemagerten Mädchen und jungen Frauen gezeigt (im Folgenden *Anorexievideo*). Beide Ansammlungen von Fotos werden durch Text der mutmaßlich jugendlichen Autoren des Videos ergänzt.

Aus technischer Sicht sind die Videos jedoch hochgradig unterschiedlich. Das Musikstück, das der erste Autor für die Ritzbilder gewählt hat, ist ein relativ bekanntes, melancholisches Rockstück, dass häufiger auch im Radio zu hören ist. Eine Wiedererkennung dieses Stücks, etwa durch Audio-Fingerprints, ist daher nicht unbedingt zielführend. Im anderen Video hingegen ist ein Musikstück wiedergegeben, das eine offene Befürwortung und Aufforderung zum Abmagern und zur Anorexie beinhaltet. Hier wäre ein Wiedererkennen des Titels wesentlich aufschlussreicher. Ergänzend hierzu könnte eine Audiosegmentierung sowie eine Sprachdetektion aufzeigen, dass das Stück in voller Länge und, zumindest auf Audio-Ebene, gänzlich unkommentiert wiedergegeben wird.

Der Text, der im Video dargestellt wird, unterscheidet sich in beiden Videos zunächst einmal von seiner Ausführung und seiner Platzierung. Im Ritzvideo wird eine schnörkellose Schriftart verwendet, die vor schwarzem Hintergrund dargestellt wird, was für automatische Texterkennung besonders günstige Voraussetzungen bietet. Im Anorexievideo hingegen wird eine stark stilisierte Schrift in verschiedenen Farben über die Bilder gelegt, was die direkte Erkennung deutlich erschwert. Inhaltlich ist die Sprache im Ritzvideo von schlechter Rechtschreibung und Grammatik gekennzeichnet, während im Anorexievideo auf Englisch relativ uneindeutige, fast träumerisch anmutende Allgemeinaussagen getroffen werden. Hinzu kommt, dass im Anorexievideo manche Sätze über mehrere Bilder hinweg fortgeführt werden. Verfahren, die Wörter aufgrund ihres Vorkommens innerhalb eines Satzes oder eines Dokuments bewerten müssen, können hier möglicherweise auf Schwierigkeiten stoßen.

Die Bilder selbst zeigen ebenfalls einen entscheidenden Unterschied: die Fotos im Anorexievideo offenbaren größtenteils erst im Kontext ihren selbstgefährdenden Inhalt, da sie zunächst nur aus einer Ansammlung von extrem dünnen bis abgemagerten Mädchen und jungen Frauen bestehen. Viele davon dürften allerdings auch in weiteren Videos derselben Art auftauchen und ein gehäuftes Auftreten von solchen bekannten Bildern dürfte ein starker Hinweis sein, dass das Thema des Videos Magersucht ist. Inwieweit diese in einem verherrlichenden oder eher abschreckenden Kontext steht, bedarf jedoch einer genaueren und kritischen Betrachtung, die eventuell eine Kontextanalyse des Videos liefern kann.

Man kann argumentieren, dass von Magersucht betroffene Jugendliche nicht sehr häufig Fotos von ihrem eigenen Körper hochladen werden, da eine selbst empfundene Hässlichkeit der Betroffenen oft Teil der Krankheit ist. Das könnte sich zum Beispiel in Foren mit Hungerwettbewerben und eingeforderten »Beweisfotos« allerdings auch anders darstellen. Im Falle des Ritzvideos sind alle Bilder Verletzungen des eigenen Unterarms. Technisch bedeutet dies vor allem, dass auf keine bekannte Datenbank mit einschlägigem Bildmaterial zurückgegriffen werden kann. Vielmehr müsste hier mit Konzepterkennung versucht werden, ein mutmaßlich unbekanntes Bild eines selbstgeritzten Körperteils automatisch zu erkennen. Ein gangbarer Weg wäre beispielsweise die Identifikation einer größeren Hautpartie, möglicherweise unter Rückgriff auf Hautdetektoren aus der Klassifikation von Pornographie. In diesen Hautpartien müssten dann in einem zweiten Schritt rote, schnörkellose beziehungsweise gradlinige Linien zu finden sein, etwa in der Form von serifenlosen Buchstaben.

5.5.2 Musikcluster: Ritzvideos

Bei fast allen untersuchten Videos wurde unterlegte Musik beobachtet. Es handelt sich um mutmaßlich urheberrechtlich geschütztes Material, das aber oft allgemeinen Charakter aufweist und als getragenes Rockstück durchaus auch im Radio zu hören ist. Wir wollen in diesem Abschnitt der Frage nachgehen, ob sich eine entsprechend aufgebaute Datenbank von Musikstücken bei der Erkennung von Selbstgefährdungs-Videos lohnen kann. Zu diesem Zweck haben wir auf einem großen Online-Videoportal zu der Suchanfrage »Ritzen« 75 Videos untersucht, von denen mit 6 Ausnahmen alle Selbstverletzungen darstellten oder behandelten. (Ausnahmen: »Ritzen« als Begriff aus dem Kunstgewerbe sowie Beiträge des belgischen Pianisten Peter Ritzen und des niederländischen Ökonoms Jo Ritzen)

Mit dem pHash Algorithmus [Zauner, 2010] haben wir hier einen paarweisen Audiovergleich durchgeführt, der bei dieser Datenmenge und einer eher naiven Anwendung etwa 2 Stunden auf einem herkömmlichen Rechner benötigt. Abbildung 5.15 stellt die Ähnlichkeit in einer gefärbten Matrix graphisch dar. Jede Auskristallisation jenseits der Diagonale stellt dabei eine mehrfache Verwendung des gleichen Liedes innerhalb dieser Videos dar. Für die sechs großen Blöcke ist hier ein Auszug aus den Liedtexten angegeben (wenn vorhanden):

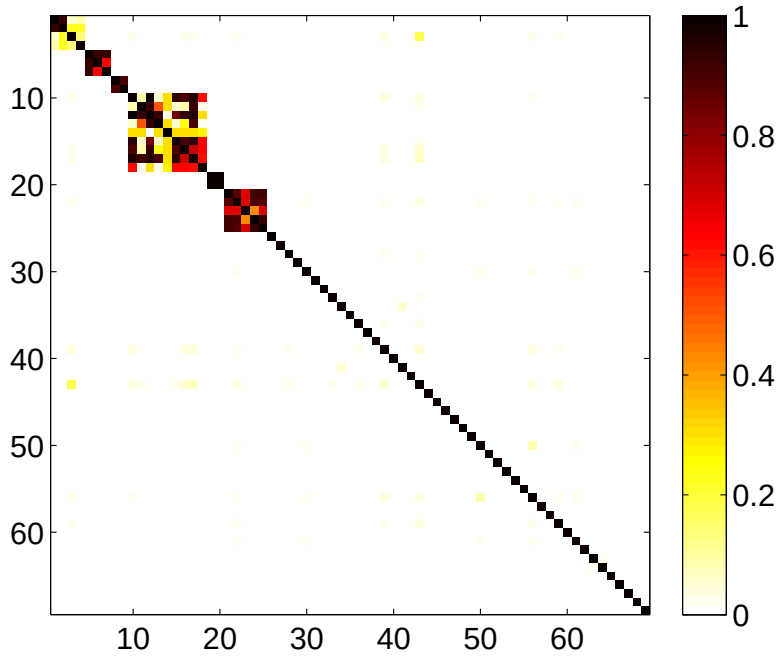


Abbildung 5.15
Ähnlichkeitsmatrix für Ritzlieder, unter Benutzung von Audio-Fingerprints von pHash.

Block 1 – verwendet von 4 Videos. Textauszug: »Sie sitzt allein in ihrem Zimmer es ist niemand da Sie hält das Messer in der Hand und führt es an ihren Arm«
Block 2 – verwendet von 4 Videos. Textauszug: »Sometimes you wonder if this fight is worth while / The precious moments are all lost in the tide«
Block 3 – verwendet von 2 Videos. Getragenes Instrumental-Stück.
Block 4 – verwendet von 10 Videos, kommt mit Abstand am häufigsten vor. Textauszug: »keiner würde sie je verstehen, ihre Liebe zur Klinge / sie ging ein Schritt weiter / ein Schnitt weiter«
Block 5 – verwendet von 2 Videos. Textauszug: »I'll show you mine if you show me yours first / Let's compare scars, I'll tell you whose is worse«
Block 6 – verwendet von 5 Videos. Textauszug: »For all the pain I caused / I didn't mean it / Oh, please believe me / I really mess things up«

Block 1 und 4 (4+10 Videos) sind also vom Text her explizit zum Thema Selbstverletzung. Die anderen sind mutmaßlich eher metaphorisch zu verstehen, werden aber aufgrund eines passenden Textinhaltes ebenfalls häufig verwendet. Es ist auch in den Videokommentaren zu beobachten, dass häufig nach dem Namen des verwendeten Liedes von anderen Nutzern gefragt wird beziehungsweise die kompletten Lyrics schon als Kommentar zum Video mitgeliefert werden. Allerdings ist davon auszugehen, dass die allgemeiner gehaltenen Lieder auch in anderen Videobeiträgen verwendet werden, so dass hier nur bei 14 von 69 Videos ein relativ eindeutiges Merkmal vorliegt.

Zusammenfassend lässt sich also vermuten, dass eine Audio-Datenbank zur Erkennung von Videos mit selbstverletzendes Verhalten prinzipiell möglich erscheint, aber nur einen kleineren Teil der Videos wird zuordnen können. Das Erkennen von allgemeiner gehaltenen Liedern kann einen zusätzlicher Hinweis auf den Inhalt liefern, ist aber kein hinreichendes Kriterium. Solche Musiktitel können natürlich auch in aufklärerischen Angeboten genutzt werden, jedoch werden sie in diesem Fall kaum in der vollen Länge wiedergegeben werden und werden daher nicht zum Overblocking führen.

5.5.3 Erkennung von Text in Bildern

Neben der im letzten Unterabschnitt untersuchten Musik ist die Verwendung von Text, entweder über Bilder gelegt oder alleinstehend, sehr häufig anzutreffen. Während manche Autoren die Krankheit nur umschreiben, finden sich bei den meisten doch entsprechende Keywords wie etwa »Ritzen«, »Klinge«, »rote Tränen« oder ähnliches. Zudem ist eine Textklassifikation auch auf Texten mit stärkerer Umschreibung denkbar.

Wir wollen in diesem Abschnitt exemplarisch die Schwierigkeiten bei der Texterkennung anhand der 69 Ritzvideos ausloten. Die Texterkennung spaltet sich in zwei Teilaufgaben: zum einen die Lokalisation der Schrift, zum anderen die Erkennung der einzelnen Buchstaben. Um die Arbeitsweise von einem hauseigenen Algorithmus zu demonstrieren, haben wir 2×4 Schnappschüsse von Text im Video gewählt, jeweils 4 ohne (Abbildung 5.16) und 4 mit untergelegtem Hintergrundbild (Abbildung 5.17). Der Fokus dieser Zusammenstellung liegt hierbei darauf, mögliche Hindernisse aufzuzeigen und weniger einen repräsentativen Durchschnitt zu ermitteln. So sind weiße oder rote Schrift auf schwarzem Grund (Abbildung 5.16(a) und 5.16(c)) die häufigsten Vertreter von Schrift ohne Hintergrundbild. Bei Schrift mit unterlegtem Hintergrundbild wird häufig normale Schrift auf schwarz-weißem Hintergrund verwendet (Abbildung 5.17(c)).

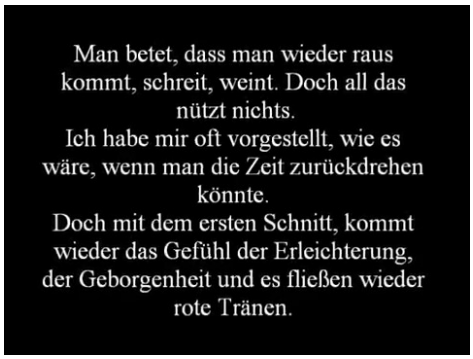
Jeweils rechts neben den Abbildungen sind die Ergebnisse der Textlokalisation mit einem grünen Rechteck markiert. Man kann hier gut sehen, dass der unmodifizierte Ansatz die Farbe der Schrift nicht gesondert berücksichtigt und daher bei Bildern mit schwacher Sättigung, etwa bei Abbildung 5.16(c), die Schrift grundsätzlich nur schlecht erkennt. Hier wäre eine gezielte Anpassung des Verfahrens notwendig. Die besten Ergebnisse sind erwartungsgemäß bei weißer Schrift auf schwarzem Grund (Abbildung 5.16(a)) oder auf schwach ausgeleuchtetem Untergrund (Abbildung 5.17(a) und 5.17(c)) zu sehen.

In einem zweiten Schritt wird auf den lokalisierten oder vermeintlich lokalisierten Textfeldern eine Texterkennung nachgeschaltet. Die exemplarischen Ergebnisse sind in Tabelle 5.17 dargestellt. So sind ungewöhnliche Schriftarten wie etwa in Abbildung 5.16(e) prinzipiell sehr gut zu lokalisieren und zu erkennen, sofern sie klar abgegrenzt und genügend groß sind. Sobald aber die Schriftart selbst nicht in der Datenbank trainiert wurde, ist das Ergebnis dennoch unleserlich. Hier wäre also eine entsprechende Vorarbeit auf den typischerweise verwendeten Schriftarten zu leisten. Die Ergebnisse sind also insgesamt eher gemischt, wobei aber bei einigen der Bilder wichtige Keywords wie »Schnitt« oder »Schmerzen« richtig erkannt wurden. Zu beachten ist aber, dass die Schreibweise von den Autoren oft schon – gewollt oder ungewollt – fehlerhaft ist. So ist die Erkennung der Phrase »flüstèrn hörèn äBèrdü Hörst mich nicht mäl schrèièn« in Abbildung 5.17(a) sehr nah am Original, eine naive Abfrage von orthographisch richtig geschriebenen Keywords würde hier aber nicht anschlagen.

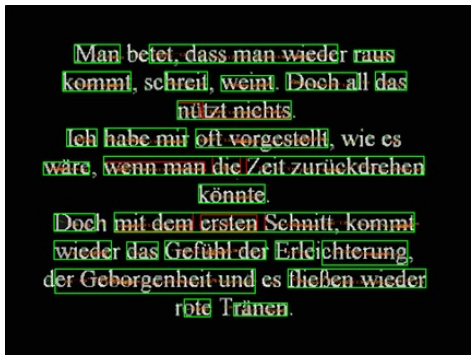
Zusammenfassend lässt sich sagen, dass die Texterkennung einen wichtigen Beitrag zur Erkennung von selbstgefährdendem Verhalten liefern kann, die Standardalgorithmen aber nicht ohne Weiteres auf dem Material funktionieren, sondern einer gezielten Anpassung bedürfen.

5.5.4 Konzepterkennung Ritzbild

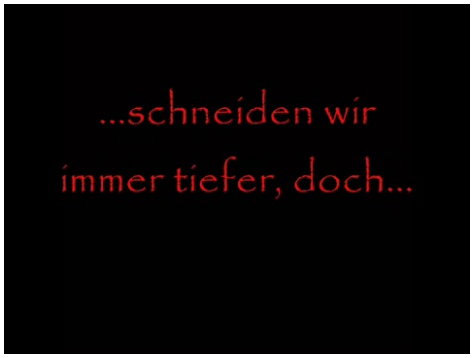
Neben dem eingeblendeten Text und der unterlegten Musik ist insbesondere die Natur der dargestellten Bilder am aussagekräftigsten. Auf einem frei zugänglichen Server zum Thema Selbstverletzung befand sich eine Sammlung von Bildern mit selbstverletzendem Verhalten; insgesamt befanden sich hier 5 383 Bilder, hauptsächlich Selbstverletzung durch Schnitte, zumeist am Oberarm, aber



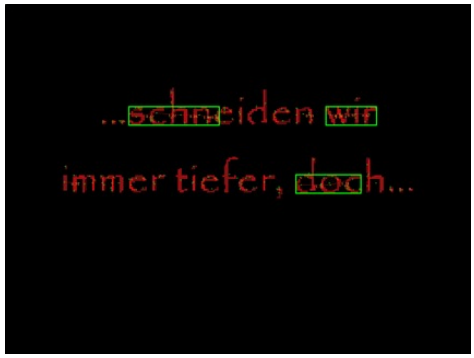
(a) Ideale Voraussetzungen: weiße Schrift auf schwarzem Grund



(b) Textlokalisation: weiße Schrift auf schwarzem Grund



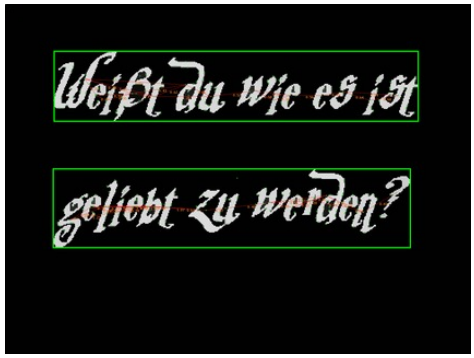
(c) Andere Schriftfarbe



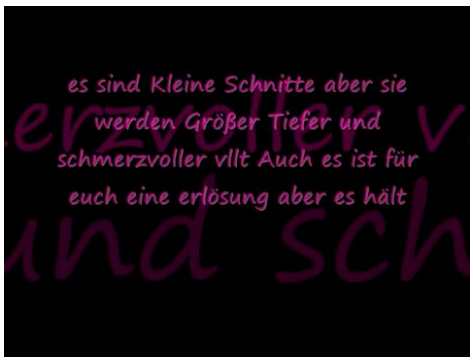
(d) Textlokalisation: andere Schriftfarbe



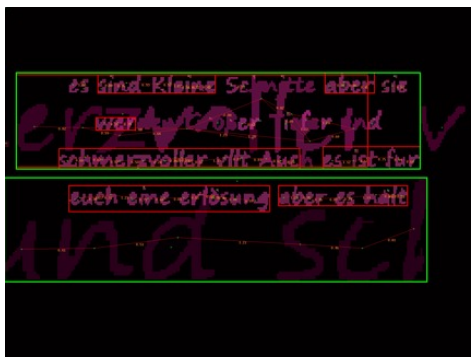
(e) Andere Schriftart



(f) Textlokalisation: andere Schriftart



(g) Schrift hinter Schrift



(h) Textlokalisation: Schrift hinter Schrift

Abbildung 5.16
Screenshots mit Text aus Videos zum Thema Ritzen.



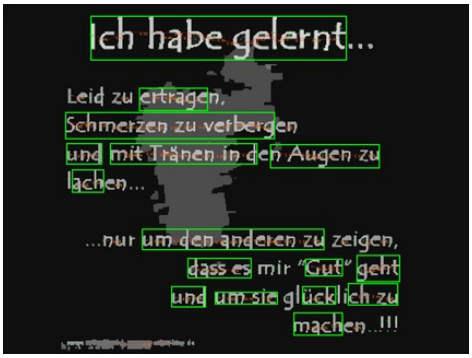
(a) Klar abgegrenzte Schrift vor Hintergrundbild, bei merkwürdiger Orthographie



(b) Textlokalisation: klar abgegrenzte Schrift vor Hintergrundbild



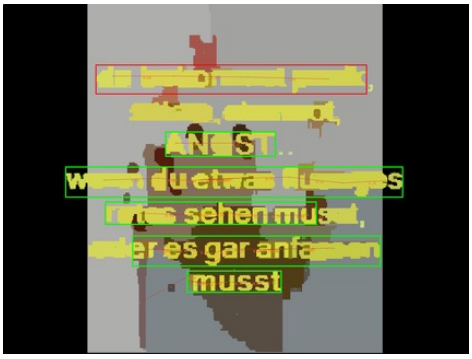
(c) Schwach abgegrenzte Schrift vor Hintergrundbild



(d) Textlokalisation: schwach abgegrenzte Schrift vor Hintergrundbild



(e) Schwierige Kontraste bei schwieriger Farbe



(f) Textlokalisation: schwierige Kontraste bei schwieriger Farbe



(g) Schwierige Kontraste, Rechtschreibfehler



(h) Textlokalisation: schwierige Kontraste

Abbildung 5.17
Screenshots mit Text über Bild aus Videos zum Thema Ritzen.

Tabelle 5.17
Schwierigkeiten der optischen Texterkennung bei verschiedenen Farbtypen, Schriftarten und unerwarteter Orthographie.

(a) Texterkennung bei reinen Textbildern

Bild	Texterkennung
5.16(a)	Man :tet. dass man wiede TZIUS ;hreit Doch a das kommt wemt nützt nichts Ich habe mir oft vorgestellt wenn man die Zeit zurückdrehen Wäl'C könnte mit dem ersten Schnitt. kommt Doc wiede* Gefühl der nchterung. Erle das fließen wieder ler Geborgenheit und 'ote fällen
5.16(c)	*unleserlich*
5.16(e)	*unleserlich*
5.16(g)	es sind Kleine Schnitte aber sie werden Größer Tiefer und sckvnerzvoller vllt Auch es ist Für euch eine erlösung aber es hält

(b) Texterkennung bei Texten mit unterlegten Bildern

Bild	Texterkennung
5.17(a)	flüstèrn hörèn äBèrdü Hörst mich nicht mäl schrèièn
5.17(c)	Ich habe _ge|c-:rnf ertrage Schmerzen zu wggibzrg« und mit Träncnifn d« rn Augen zu lach« um dch anderen zu geht dass es 'Guf ich zu und ück Um SIC mach«
5.17(e)	*unleserlich* Üfes sehen mu musst
5.17(g)	niemalä *unleserlich*

auch Narben und Selbstverstümmelung. Manche Schnitte gingen so tief, dass Fettgewebe zu sehen war. In diesem Abschnitt soll der Frage nachgegangen werden, inwieweit diese Bilder durch einen Konzepterkenner erkennbar gemacht werden können. Dafür greifen wir abermals auf Technologie des griechischen Forschungsinstituts *Information Technology Institut* (CERTH-ITI) zurück. Um einen Konzepterkenner zu trainieren, braucht man zusätzlich Material, von dem der Inhalt unterschieden werden soll. Als Vergleichsmaterial verwenden wir hier Tattoos, die oft eine überraschend große Ähnlichkeit zu Schnittwunden aufweisen können (Abbildung 5.18).

Um die Auswahl möglichst homogen zu machen und damit die Schwierigkeit für den Klassifikator zu erhöhen, haben wir das Material auf Bilder des Unterarms beschränkt. Auch haben wir nur Tattoos gewählt, die in den Farben Rot oder Schwarz gemalt sind, da farbige Tattoos wesentlich einfacher, etwa über Farbhistogramme, zu erfassen sind. Das Material bestand aus 2mal ungefähr 460 Bildern, wovon jeweils 400 für das Training des Algorithmus verwendet wurden. Die übrigen Bilder wurden als Testanfrage für die Klassifikationsgüte genutzt. Der Konzepterkenner liefert dann zu jedem Bild eine Aussage darüber, für wie wahrscheinlich er es für ein Bild mit Ritzwunden hält. Das Training der Verfahren benötigt einen vollen Tag, die Anfrage selbst dauert wenige Sekunden pro Bild.

Die Ergebnisse für Tattoos sind in Abbildung 5.19(b), die Ergebnisse für Ritzbilder in Abbildung 5.19(a) dargestellt. Bei einem Schwellwert von 50 % Wahrscheinlichkeit gibt es insgesamt 8 von 151 Bildern, die fehlklassifiziert wurden, davon 6 aus dem Bereich Ritzbilder, die für Tattoos gehalten wurden. An der rechten Seite der Abbildungen sind exemplarisch einige Beispielbilder gezeigt. So wird bei den Tattoos eine stilisierte Sonne mit ziemlicher Sicherheit (fast 0 %) nicht als Ritzbild angesehen. Insbesondere bei geraden Linien, etwa in einem tätowierten Notensatz oder bei parallel verlaufenden, bereits heilenden Striemen, ist sich das Verfahren eher unsicher. In Abbildungen, wo teilweise das Fettgewebe zu erkennen ist, schlägt das Verfahren üblicherweise mit der Aussage »absolut sicher« an.



(a) Gradlinige Narben bei ritzendem Verhalten



(b) Gradlinige Muster bei Tattoos von chinesischem Schriftzeichen



(c) Muster durch zugefügte Wunden



(d) Muster bei komplexen Tattoos



(e) Geritzte Symbole



(f) Tätowierte Symbole

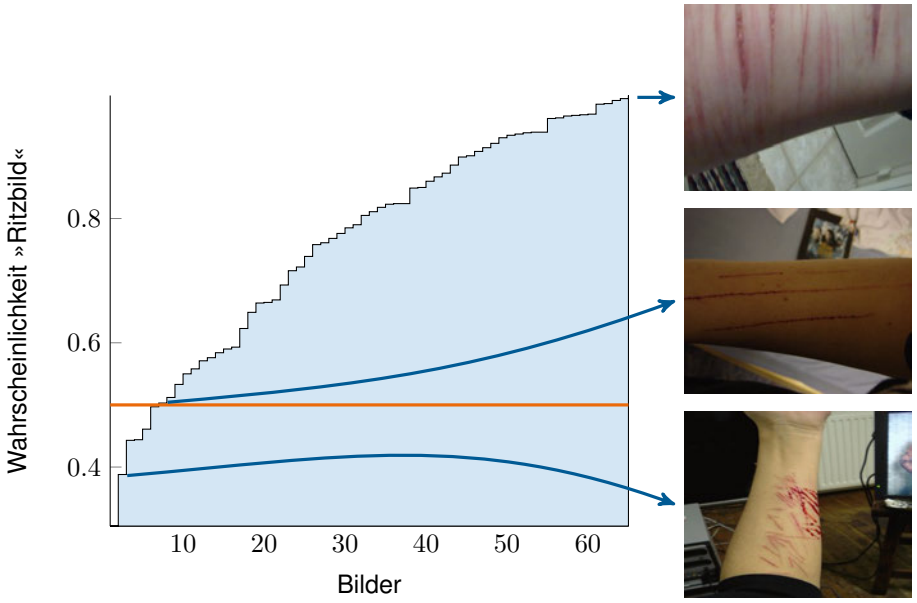


(g) Eingeritzte Schrift

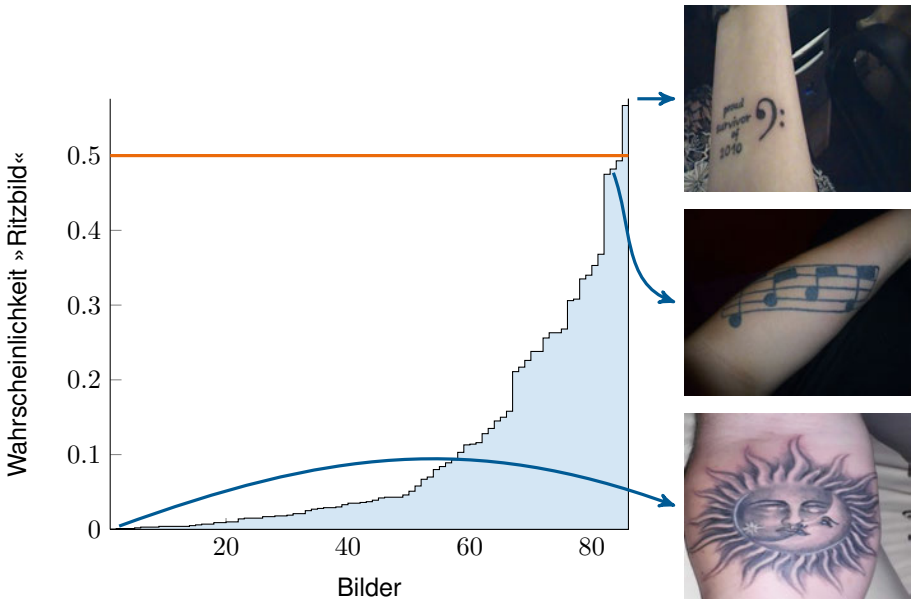


(h) Tätowierte Schrift

Abbildung 5.18
Vergleich von Bildern aus der Kategorie Selbstverletzung und Tattoos, jeweils am Unterarm.



(a) Ergebnisse des Konzepterkenners »Ritzbilder«, bei tatsächlich vorliegenden Verletzungen des Unterarms.

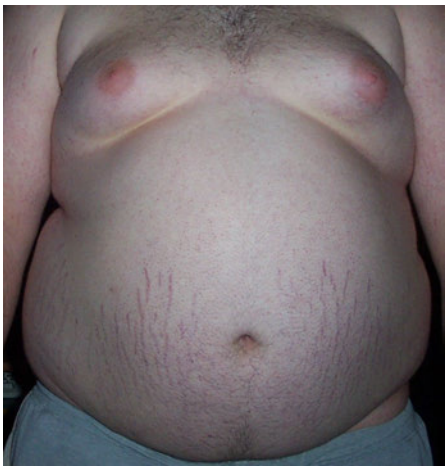


(b) Ergebnisse des Konzepterkenners »Ritzbilder«, bei Tattoos auf dem Unterarm.

Abbildung 5.19
Übersicht über die ausgegebene Wahrscheinlichkeit des Konzepterkenners »Ritzen / Tattoo« nach Kategorie. Angenommener Schwellwert bei 50 %.



(a) Am Häufigsten: berühmte Persönlichkeiten oder Models auf dem Laufsteg, hier von einem vergleichsweise normal gebauten Model. Bildquelle: http://upload.wikimedia.org/wikipedia/commons/thumb/e/ed/Romantic_fashion_model.JPG, abgerufen am 29.11.2012 im englischsprachigen Artikel »Catwalk«.



(b) Selten: »Anti-Thinspos«, also als abschreckend verwendete Bilder. Bildquelle: http://upload.wikimedia.org/wikipedia/commons/thumb/b/bb/Central_Obesity_008.jpg, abgerufen am 29.11.2012 im deutschsprachigen Artikel »Adipositas«.



(c) Selten: Verherrlichte Bilder von abgemagerten Personen, hier am Beispiel von Christian Bale in dem Film »Der Maschinist«.

Abbildung 5.20
Archetypische Beispiele von verschiedenen Thinspo-Bildern.

Zusammenfassend kann man sagen, dass die Konzepterkennung bei Ritzbildern selbst unter schwierigen Bedingungen sehr gut funktioniert und ein Einsatz in der Praxis denkbar erscheint, etwa bei der Durchforstung größerer unbekannter Bildsammlungen.

5.5.5 Fingerprinting auf Bildsammlungen im Anorexie-Kontext

Auf einem frei zugänglichen Server zum Thema »thinspos« lagen 84 163 Dateien, vornehmlich Bilder, von insgesamt 5,3 Gigabyte Datenumfang. Nach einer Sichtung der Daten ziehen wir die folgenden Schlüsse.

Zunächst ist auffallend, dass die Häufigkeitsverteilung der Bilder insgesamt stark von der Quelle abhängt. So ist es relativ einfach, etwa aus Modeforen oder aus Illustrierten, an Bilder von dünnen Models auf Laufstegen oder an Bilder von berühmten Persönlichkeiten heranzukommen, was sich in den Daten widerspiegelt. Bilder mit extrem anorektischen Menschen finden sich jedoch eher selten. Insgesamt lassen sich die Bilder grob in 6 Kategorien aufteilen (vgl. Abbildung 5.20):

- Fashion-Models auf dem Laufsteg, sehr viele
- Berühmte Persönlichkeiten, extrem viele
- Frauen am Strand, sehr viele
- Extrem adipöse Personen, extrem wenige
- Vorher-Nachher Bilder, einige
- Extrem anorektische Personen, sehr wenige

Tabelle 5.18
Ergebnisse von pHash Bilder Fingerprinting auf Christian Bale »Der Maschinist«, als Distanzwert wiedergegeben.

Artefakte			
Verwaschen: 0	Schlechte Qualität: 2	Collage: 20	Text: 2
Verkleinerung			
50 %: 0	25 %: 2	10 %: 4	5 %: 8
Farbveränderung			
Glossy: 2	Verfremdet: 2	Sepia: 0	Schwarz-Weiß: 10
Rotation			
Durchgehend: 32			

Auch wenn insbesondere die hier ausgewählten Repräsentanten von Laufsteg-Models oft ein pathologisch erscheinendes Körpergewicht aufwiesen, ist eine eindeutige Aussage hierzu eher schwierig. Ähnliches gilt für die Bilder von berühmten Persönlichkeiten oder die Strandbilder. Wesentlich aussagekräftiger sind die Bilder mit extremer Anorexie; ihr geringes Vorkommen auf dem Server spricht dafür, dass es sich hierbei vermutlich nur um begrenzt viele Bilder handelt, die im Internet verbreitet sind, und dass hier Fingerprinting abermals eine vielversprechende Technik bietet.

Wir wollen uns daher nun der Frage widmen, wie stabil allgemein einsetzbare Bilder-Fingerprints auf robuste Veränderung reagieren. Dazu verwenden wir den pHash Algorithmus [Zauner, 2010] sowie mehrere, von uns veränderte Bilder einer Anorexie-Darstellung. Als Ausgangsbild verwenden wir einen ebenfalls auf dem Server liegenden Ausschnitt aus dem Film »Der Maschinist« (Spanien, 2004, R: Brad Anderson, P: Castelao Producciones/Fantasy Factory/Filmax Group), in der die Hauptfigur, gespielt durch Christian Bale, mehr und mehr dem Wahnsinn verfällt und dabei extrem abmagert. Der Schauspieler hat speziell für diese Rolle einen enormen Anteil seines Körpergewichtes abgenommen.

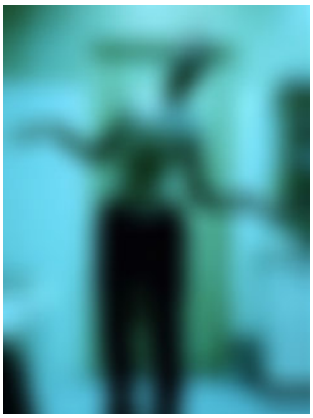
Wir hatten bei der Betrachtung der Videobilder in Abschnitt 3.4.3 festgestellt, dass die Bilder oft in schlechter Auflösung, verfremdet, farblich variiert oder mit überblendetem Text verändert worden sind. Rotation haben wir keine gesehen, sie ist aber natürlich denkbar. Davon ausgehend haben wir das Originalbild auf passende Weise verändert (Abbildung 5.21).

Die Ergebnisse des Fingerprints mit pHash sind in Tabelle 5.18 dargestellt. Das Verfahren reagiert auf die häufigsten Verfremdungen, also schlechte Qualität, leichte Farbveränderung oder übergelegter Text, ausgesprochen robust (Distanz 0 oder 2, von maximal 64 möglich). Allerdings reagiert die Technik sehr empfindlich auf Rotation oder Collagen.

Zusammenfassend lässt sich festhalten, dass Fingerprinting zur Identifikation von bekannten Bildern sehr stabil auf die meisten von den Videoautoren vorgenommenen Veränderungen reagiert und für die Praxis einsetzbar ist. Abzuwägen ist der Einsatz eines zum Beispiel gegenüber Rotation robusten und damit aufwändigeren Verfahren gegenüber der erforderlichen Schnelligkeit beziehungsweise der zur Verfügung stehenden Ressourcen.

5.5.6 Textklassifikation Pro-Ana

Wir haben uns bislang mit Videos und Bildern aus dem Bereich Selbstgefährdung gewidmet. In diesem Abschnitt wollen wir die Textinhalte der Anorexie verherrlichenden Webseiten etwas genauer untersuchen. Diese sind bei magersuchts- und bulimieverherrlichenden Foren zunächst widersprüch-



(a) Verwaschen



(b) Schlechte Bildqualität



(c) Ausschnitt rotiert



(d) Darübergelegter Text



(e) 50 % Originalgröße



(f) 25 % Originalgröße



(g) 10 % Originalgröße



(h) 5 % Originalgröße



(i) Glossy



(j) Verfremdet



(k) Sepia



(l) Schwarz-weiß

Abbildung 5.21
Verschiedene Veränderungen am Bild »Der Maschinist«.

lich in ihrer Darstellung. Es gibt einen großen Teil an vorgefertigten, fast standardisierten Textbausteinen wie etwa »Anas Brief«:[19]

> »Erlaube mir mich vorzustellen… Mein Name, oder wie ich von sogenannten ›Ärzten‹ genannt werde, ist Anorexie. Mein vollständiger Name ist Anorexia Nervosa, aber du kannst mich Ana nennen. Ich hoffe, wir werden gute Freunde.«

Auch Sinnsprüche wie »Engel haben keinen Hunger« oder »Dünn sein ist wichtiger als gesund sein« sind leicht erkennbar. Andererseits wählen die Betroffenen viele Strategien, um ihr Vorgehen zu verschleiern oder zu verheimlichen. Die Foren sind daher zum Teil geschlossen und nur über Anmeldung zugänglich; hier wäre zu prüfen, inwieweit die Startseite schon genug Aufschluss über ihren Inhalt bietet. Tipps zur Geheimhaltung werden oft auf den Foren angeboten. Auch werden bestimmte Wörter verwendet, die nur Eingeweihten vertraut sind (Beispiel »ATTE« = »Ana till the end«, also Abnehmen bis zum Hungertod).

So sind auch vorgeschaltete Disclaimer eindeutig, wie etwa:

> »Dies ist eine Pro-Ana Homepage. Wer gesund ist oder zur Zeit in therapeutischer Behandlung ist, sollte diese Seite verlassen.«

Ebenfalls gut erkannt werden können beispielsweise Body-Mass-Index Rechner mit entsprechend der Krankheit angepasster Auslegung (etwa »BMI <15 – Medizinische Bewertung: Sehr stark untergewichtig! Suchen Sie sofort einen Arzt auf! Ana's [sic] Bewertung: Gut, du bist perfekt! Wow, deine Knochen! Aber pass auf, viel weiter runter geht es nicht!«). Diese Eigenschaft der sich häufig wiederholenden Textmuster wird im folgenden Abschnitt zum automatischen Auffinden entsprechender Webpräsenzen genutzt.

5.5.6.1 Entdecken von Pro-Ana-Webseiten

Da keine Liste von manuell als jugendschutzrelevant markierten Webadressen vorlag, wird ein Verfahren vorgestellt, um entsprechende Listen mittels einer Heuristik – möglicherweise nicht vollständig korrekt, dafür aber weitgehend automatisch – zu erstellen.

Zunächst wird eine Möglichkeit zur automatischen Anfrage einer Websuche benötigt. In diesem Fall wurde die »Yahoo! BOSS Search API«[20] verwendet. Zum Auffinden des entsprechenden Materials werden drei Suchanfragen formuliert: »Anas Brief«, »Anas Gesetze« und »Dies ist eine Pro-Ana-Seite«.

Diese drei Anfragen liefern insgesamt 252 Treffer. Darunter finden sich allerdings auch Verweise zu »spiegel.de« oder »bmg.bund.de«, also Nachrichten- und Aufklärungsseiten. Hier hilft eine einfache Heuristik: Zur automatischen Filterung dieses zumeist professionellen Inhalts reicht es aus, auf die Wörter »Impressum«, »Nutzungsbedingungen« und »Terms of use« zu prüfen. Kommt mindestens einer dieser Begriffe auf der entsprechenden Webseite vor, gilt sie als professionell und wird dem Gegenkorpus zugeordnet.

[19] Im Internet gebräuchlich ist auch die falsche Schreibweise »Ana's Brief«.
[20] siehe http://developer.yahoo.com/boss/search/, Stand: November 2012

Eine oberflächliche Prüfung der URL-Listen zeigt tatsächlich sehr deutlich die hohe Qualität dieser einfachen Heuristik. Einzig der kostenlose Webauftritt-Anbieter »Jimdo«[21] scheint durch das Raster zu fallen, da die Software automatisch ein Impressum auf jeder Seite anlegt. Diese eine Ausnahme wird der Heuristik hinzugefügt.

Als Nächstes werden auf jeder Webseite die ersten 10 ausgehenden Links zum eigenen Host mit Tiefe eins verfolgt und zusätzlich heruntergeladen. Die Tiefe bestimmt beim Crawling die maximale Entfernung von der Ursprungsseite in der Anzahl der Links. Eine Tiefe von eins bedeutet also, dass nur die Links auf der Ursprungsseite verfolgt werden. Die in den referenzierten Webseiten enthaltenen Links werden nicht weiterverfolgt.

Auf diese Weise lässt sich auch das inhaltliche Umfeld erschließen, ohne zu viel irrelevanten Kontext zu generieren. Aus den ursprünglich 252 Suchtreffern, die sich in 49 Seiten mit und 202 ohne Impressumsverweis aufteilen, werden so direkt 1 350 Pro-Ana-Webseiten und 486 Webseiten aus professionellen Angeboten ermittelt.

5.5.6.2 Analyse der Webseiten

Eine Merkmalsselektion auf diesen Daten liefert hauptsächlich Auszüge aus Anas Brief, weil diese Merkmale sich in der positiven Klasse am stärksten wiederholen. Daher sind in der folgenden Liste alle in den verschiedenen Versionen von Anas Brief enthaltenen Merkmale herausgefiltert:

- dein gewicht
- stay strong
- nicht essen ohne mich schuldig
- esslöffel
- der geschmack
- xoxo
- apfelessig
- gewürztes essen
- das fett aus
- jede mahlzeit
- du weisst wofür
- hängen lassen
- ein schluck wasser
- bestrafe
- thinspo
- abnehmen willst
- abführmittel
- deine kalorienaufnahme
- die kalorien
- diäten
- bmi
- mia
- aussehen
- dünn sein
- …

[21] siehe http://de.jimdo.com/, Stand: November 2012

Tabelle 5.19
Die Klassifikationsergebnisse für die Zielkategorie *Pro-Ana* gegenüber *Allgemein ab 6* und *Pro-Ana-Information* getrennt nach Merkmalsgruppen.

Merkmale	F-Maß	Präzision	Trefferquote	Prä@90
Pro-Ana				
nur URL	58,7 %	64,4 %	53,9 %	61,0 %
nur Text	85,7 %	92,2 %	80,0 %	63,4 %
Text+URL	86,8 %	92,4 %	81,8 %	72,2 %
nur Haupttext	71,3 %	79,7 %	64,5 %	46,3 %
Haupttext+URL	76,3 %	80,7 %	72,3 %	58,4 %

Bei der Trennung der professionellen von den privaten Pro-Ana-Angeboten helfen offenbar direkte Ansprachen wie »*dein* Gewicht« oder »*deine* Kalorienaufnahme«. Diese kommen zum Beispiel in den eher beratenden Zeitungsartikeln sehr viel seltener vor. Natürlich sind einige der ermittelten starken Merkmale für sich gesehen harmlose Begriffe wie »Esslöffel« oder »Apfelessig«, die auch auf vielen anderen Webseiten zu finden sind. Einzelne Begriffe sagen nicht viel über die Kategorie einer Webseite aus. Ganz generell gilt auch hier, dass es die Masse an Merkmalen ist, die den Ausschlag gibt.

Bei diesem Klassifikations-Experiment gelten wieder die Rahmenbedingungen der Experimente der vorangegangenen Abschnitte (5.4.3, 5.3.6). Die Zielkategorie ist definiert als:

- *Pro-Ana*: Eine nach der beschriebenen Heuristik automatisch erstellte Sammlung von 1 350 Webseiten aus 153 unterschiedlichen Webangeboten.

Als negative Klasse werden folgende Kollektionen kombiniert:

- *Allgemein ab 6*: Eine Sammlung von 2 877 manuell als nicht jugendschutzrelevant annotierten Webseiten ohne thematischen Fokus.
- *Pro-Ana-Information*: Eine nach der beschriebenen Heuristik automatisch erstellte Sammlung von 486 als professionell und nicht jugendschutzrelevant eingestuften Webseiten aus 42 unterschiedlichen Webangeboten.

Tabelle 5.19 zeigt, dass die Präzision trotz der teilweise allgemeinen Merkmale mit 92,4 % recht hoch, das Overblocking also eher gering ist, und das bei einer Trefferquote von 81,8 %. Demgegenüber fällt die Präzision bei einer Anhebung der Trefferquote auf 90 % direkt auf 72,2 % ab. Wenn also, um wenige relevante Webseiten zu übersehen, der Underblocking-Fehler auf 10 % reduziert wird, steigt der Overblocking-Fehler relativ stark. Dies zeigt, dass einige der für diesen Versuch nach der beschriebenen Vorgehensweise automatisch als jugendschutzrelevant eingestuften Webseiten nur sehr wenige Hinweise auf die Zielkategorie liefern. Ob dies ein Problem der vorangegangenen Kategorisierungsheuristik ist oder ein generelles Problem der automatischen Klassifikation lässt sich ohne manuell annotierte Daten nicht überprüfen.

Allerdings lässt sich sagen, dass ein großer Vorteil dieses automatischen Klassifizierers gegenüber der simplen Heuristik, mit der die Trainingsdaten ausgewählt wurden, seine Flexibilität ist. Bei einem banalen Tippfehler wie »Annas Brief« kann es schon sein, dass eine Regel nicht mehr greift. Das maschinell erlernte Modell ist abstrakter und daher robuster gegen solche leichten Veränderungen am Inhalt und erkennt auch Merkmale aus dem Kontext der bekannten Zielwörter. So würde die beschriebene Heuristik auch sofort scheitern, sobald eine Pro-Ana-Webseite ein Impressum enthält. Für den Klassifizierer ist der Term »Impressum« zwar ein wichtiges Merkmal für nicht-relevanten Inhalt, es ist aber nur eins von sehr vielen Kriterien.

Zusammenfassend lässt sich sagen, dass das Aufspüren und Erkennen von Pro-Ana Seiten relativ leicht fällt. Durch technische Verfahren können bekannte Textbausteine auch dann gut wiedererkannt werden, wenn sie verändert sind. Abgewogen werden muss jedoch, inwieweit der Aufwand zum Trainieren der Klassifizierer im Verhältnis zur Anzahl der zu erwartenden Seiten steht. Auch ist zu beachten, dass passwortgeschützte Foren für uns in den Experimenten nicht zugänglich waren, so dass sich hierüber keine Aussagen treffen lassen.

5.5.7 Zusammenfassung: Betrachtung von Verfahren im Bereich Selbstgefährdung

Wir haben uns in diesem Abschnitt mit mehreren technischen Verfahren zur automatischen Erkennung von selbstgefährdenden Inhalten beschäftigt. Der Bereich der Selbstgefährdung ist im Vergleich zu Gewalt oder Extremismus am stärksten durch Inhalte geprägt, die die Betroffenen jeweils selbst erstellen. Das führt dazu, dass sich die Inhalte zwar thematisch und vom Aufbau her sehr ähnlich verhalten, dass man aber selten unverändert gebliebene Bausteine wiederfinden kann: Bilder werden neu betextet, Musik wird neu unterlegt, Texte werden umformuliert.

Prinzipiell lässt sich festhalten, dass die Verfahren entweder bereits gut funktionieren (etwa bei Fingerprinting oder Konzepterkennung) oder mit zumutbarem Aufwand trainiert werden können (etwa bei der automatischen Texterkennung). Es ist aber auch deutlich geworden, dass die Verfahren immer nur bei einem Teil des vorhandenen Materials die Möglichkeit haben, entsprechende Merkmale zu extrahieren. Es bietet sich an, die Ergebnisse der jeweiligen Verfahren zu kombinieren und in einen übergreifenden Klassifikator als Teilmerkmale einfließen zu lassen, was zwar den technischen Aufwand erhöht, aber gleichzeitig die Präzision deutlich verbessern kann.

5.6 Abschließende Bewertung: Fragen und Antworten

In diesem Abschnitt werden aus den Experimenten generelle Aussagen für den technischen Jugendschutz abgeleitet. Wir gehen dabei auch auf Aspekte der absehbaren Praxistauglichkeit ein und verweisen auf bereits existierende Industriezweige. Hierzu kommen wir zurück auf die bereits eingangs gestellten Fragen in diesem Kapitel.

5.6.1 Fragestellungen aus dem Bereich Text

Zu den Verfahren aus dem Bereich Text können wir die folgenden Antworten aus den Experimenten ableiten:

Inwiefern kann man durch Methoden der Textklassifikation die manuelle Erstellung und Wartung von Keyword-Listen unterstützen oder ersetzen? Ganz eindeutig kann festgestellt werden, dass automatische Verfahren der Keyword-Extraktion, der Themenmodellierung und der Textklassifikation die Erstellung und Wartung von Keyword-Listen unterstützen können. Dies haben die Experimente in Abschnitt 5.2 klar gezeigt. Dort haben die automatischen Verfahren bessere Ergebnisse geliefert als die Verwendung einer manuellen Keyword-Liste. Ob die manuelle Erstellung und Wartung von Keyword-Listen komplett ersetzt werden kann, ist schwieriger zu beantworten. Zu bedenken ist, dass jedes automatisch erzeugte Modell nur so gut sein kann wie die Daten, aus denen es generiert wurde. Werden also zum Beispiel aktuelle Dokumente zum Generieren von Modellen verwendet, so könnten Wörter, die zu einem früheren Zeitpunkt bedeutsam waren, unter Umstän-

den nicht berücksichtigt sein. Insofern müsste ein automatisch erzeugtes Modell über einen längeren Zeitraum gebaut werden. Bei einer Aktualisierung des Modells müssten zudem bisher verwendete, ältere Dokumente zumindest zum Teil mit berücksichtigt werden. Generell sind automatische Verfahren wesentlich weniger aufwändig in der Wartung.

Ist eine automatische Klassifikation von jugendschutzrelevanten Inhalten auch detailliert, also mit konkreter Altersklassifikation, möglich? Wie verhält es sich konkret bei den Unterkategorien? Die in Abschnitt 5.3.6 beschriebenen Ergebnisse zeigen, dass durch eine zweistufige Klassifikation die Gewalt-Inhalte auch feiner in ihre jeweiligen Altersstufen getrennt werden können. Da in den Experimenten nur der Datenbestand für die ab 18 freigegebenen Inhalte eine ausreichende Menge umfasst, ist eine Verallgemeinerung diesbezüglich zwar schwierig, die Experimente legen jedoch nahe, dass sich die Erkennungsraten mit wachsender Anzahl verfügbarer Daten noch wesentlich steigern lassen.

Insgesamt waren die für die Studie verfügbaren Trainingsdaten weder hinsichtlich ihres Umfangs noch hinsichtlich ihrer Repräsentativität ausreichend. Zudem gibt es eine Reihe zusätzlicher Merkmalsgruppen, welche für die Studie noch nicht evaluiert werden konnten, zum Beispiel Hyperlinks und Sprachmodelle (siehe Abschnitte 4.4.5 und 5.4.4). Daher erscheint eine wesentliche Verbesserung der Klassifikationsgenauigkeiten möglich. Schließlich ist auch ein halb-automatisches Vorgehen machbar (vgl. Abschnitt 4.2.1.1), bei welchem die »sicher« klassifizierten Dokumente bei einem sehr geringen Fehler automatisch in Klassen eingeordnet werden und die »unsicheren« Dokumente manuell klassifiziert werden.

Können automatisch bestimmte sprachliche Merkmale für Unterkategorien von jugendschutzrelevanten Inhalten gefunden werden? Kann man anhand dieser nahe beieinander liegende Inhalte, zum Beispiel Geschichtsrevisionismus und historische Online-Enzyklopädien, auseinanderhalten? Da im Bereich Geschichtsrevisionismus und Rechtsextremismus häufig auf die Verwendung fremdsprachlicher Lehnwörter verzichtet wird, ergibt sich in diesen Angeboten eine spezielle Terminologie, die sich mit statistischen Verfahren aufdecken lässt. Dies wurde in Abschnitt 5.4.2 exemplarisch dargestellt. Entsprechende Inhalte, die diese spezielle Terminologie nicht verwenden, sind wesentlich schwerer zu identifizieren. Das konnte man unter anderem an der schlechteren Präzision in der Erkennung von extremistischen Webseiten erkennen, die keine strafrechtlich relevanten Inhalte aufweisen (Abschnitt 5.4.3). Es ist in dem Fall also notwendig, den Sinn des Textinhalts zumindest annähernd zu erfassen. Statt nur auf Wortebene müssen charakteristische Muster hier beispielsweise in Form bestimmter Phrasen identifiziert werden. Dies ist mit statistischen Verfahren zwar in gleicher Weise möglich, nur werden hierzu erheblich mehr annotierte Dokumente benötigt, da sich Phrasen üblicherweise wesentlich seltener wiederholen als einzelne Wörter.

Um robuste Klassifikationsmodelle zu erhalten, ist es sinnvoll, möglichst ähnliche Beispieldokumente aus unterschiedlichen Kategorien zu betrachten. Die charakteristischen Merkmale der jeweiligen Kategorien lassen sich auf diese Weise deutlich schärfer herausstellen. Auch für beratende Angebote zu selbstverletzendem Verhalten oder für Webseiten, die Aussteiger aus der rechtsextremen Szene unterstützen, lassen sich die trennenden Merkmale hervorheben, und das trotz der inhaltlichen und terminologischen Nähe zum jugendschutzrelevanten Material. Für eine erfolgreiche Klassifikation ist vor allem Auswahl und Größe der Datengrundlage wichtig.

Sobald ein Klassifikator ausreichend trainiert ist, lässt er sich auch für große Daten gut anwenden. Dass dies möglich ist, belegt unter anderem die Vielzahl an eingesetzten Spamfiltern, die auch bei großem E-Mail-Aufkommen recht gut zum Beispiel offizielle Anschreiben von Köder-E-Mails auseinanderhalten können.

Kann automatisch ein typisches Themenspektrum in einer Menge verdächtiger Webseiten generiert werden? Auch hier legen die Experimente in Abschnitt 5.2 nahe, dass dies möglich ist. Generell kann zu einer Menge von Dokumenten ein Themenspektrum auf unterschiedlichem Detailniveau automatisch generiert werden. Dabei ist zu beachten, dass in diesem Spektrum natürlich auch allgemeine und irrelevante Themen, zum Beispiel solche, die ausschließlich aus Stoppwörtern (Artikeln, Pronomen etc.) bestehen, auftreten. Ein neues, vorher ungesehenes Dokument lässt sich dann anhand des Themenspektrums einordnen. Diese Einordnungen werden sich dann bei jugendschutzrelevanten Webseiten deutlich von denen unbedenklicher Webseiten unterscheiden. Dennoch sollten diese Techniken eher zum Beispiel zu Recherchezwecken und weniger direkt in Filtern eingesetzt werden. Zum Filtern eignen sich Klassifikationstechniken besser, da diese speziell auf das Ziel der Diskriminierung verschiedener Kategorien ausgerichtet sind.

Kann man anhand des Textes erkennen, inwieweit die dargestellten Bilder jugendschutzrelevant sind? Die Frage kann nicht abschließend beantwortet werden. Es ist eventuell machbar, kann aber letztlich nur anhand von sehr großen, bereits eingeordneten Datenmengen verifiziert werden. Für das hier gewählte Beispiel (Robbenjagd, Abschnitt 5.3.5) war es hingegen nicht möglich.

5.6.2 Fragestellungen aus dem Bereich Multimedia

Zu den Verfahren aus dem Bereich Multimedia können wir die folgenden Antworten aus den Experimenten ableiten:

Kann man ein neu hochgeladenes, aber bereits bekanntes Video erkennen, auch dann, wenn es leicht verändert wurde? Ja. Auch bei stark komprimiertem Material oder bei minimal veränderten Videos etwa durch vorgeschaltete oder überlagerte Webseitenlogos reagieren Video-Fingerprinting Techniken robust in der Erkennung von bereits bekanntem Material. Das Arbeiten mit Bildmaterial für Video-Fingerprints ist deutlich aufwändiger als beispielsweise für Audio-Fingerprints und muss bei diesen Verfahren als Aufwand entsprechend berücksichtigt werden. Werden die Fingerprints hingegen auf die Tonspur reduziert, arbeiten die Verfahren deutlich schneller und weniger speicherintensiv. Jedoch werden dann zum Beispiel Gewaltvideos mit neu unterlegter Musik nicht mehr erkannt. Für größtmögliche Abdeckung ist eine Kombination beider Modalitäten erforderlich.

Ein weiterer Faktor, der sich in Geschwindigkeit und Speicheranforderungen widerspiegelt, ist die Dichte der Fingerprints auf dem Material. Engmaschige Fingerprints sind sehr aufwändig, ermöglichen aber das Erkennen von Kompilationen, also Zusammenschnitte von mehreren Videos, selbst wenn von diesen Videos nur eine Auswahl bekannt ist.

Die Verfahren werden auch heute schon in der Filmindustrie für die sogenannte Copy Detection eingesetzt und sind gut erforscht. Dennoch erscheint bei der schieren Datenmenge an neu hochgeladenem Material auf den diversen Videoportalen nur eine Untersuchung im Verdachtsfall als sinnvoll. Das ist zum Beispiel dann der Fall, wenn die Titel der Videos relevante Keywords enthalten, die ihrerseits aber mehrdeutig sein können (vergleiche die große Anzahl von Bands mit dem Namen »Braindead« im Titel, Abschnitt 5.3.2).

Kann man neu hochgeladene, jugendschutzrelevante Musikstücke automatisch erkennen, auch dann, wenn sie etwa nachgesungen werden? Gerade für Musik eignen sich Audio-Fingerprinting Verfahren sehr gut. Wir haben sie erfolgreich für rechtsextremistische Musik (Abschnitt 5.4.5) einsetzen können und auch für selbstgefährdende Videos scheint diese Technik hilfreich (Abschnitt 5.5.2). Die Technologie findet heute auf vielen Endgeräten bereits ihren Einsatz, zum Beispiel zur Liederkennung auf Handys, was auch ihre Echtzeitfähigkeit unterstreicht.

Auch nachgesungene Lieder kann man, mit Einschränkungen, gut wiedererkennen (Abschnitt 5.4.6). Prinzipiell gilt: je länger die Passagen hierbei sind, in denen sich die Stücke gleichen, desto besser. Vergleichbare, wenn auch wesentlich einfacher gestrickte Ansätze findet sich beispielsweise in Karaoke-Spielen, in denen der Spieler eine positive Bewertung für besonders gut nachgesungene Lieder erhält.

Wie kann man die redaktionelle Klassifikation bei unbekannten Liedern möglichst effizient gestalten? Die Antwort zu dieser Frage ähnelt zum Teil der vorhergehenden Antwort, beleuchtet aber den Aspekt, dass man die Lieder im Voraus nicht kennt. Wir haben insbesondere für die Musikunterlegung von Videos im Bereich Selbstgefährdung (Abschnitt 5.5.2) gesehen, dass sich die Lieder automatisch gruppieren lassen und man schnell Auffälligkeiten wie etwa ein oft wiederkehrendes Lied daran erkennen kann. Dies kann man leicht für die anderen Kategorien übertragen. Bei Musikvideos mit jugendschutzrelevantem Inhalt auf Online-Portalen ist es häufig der Fall, dass kein von der Band offiziell produziertes Video hinterlegt ist, sondern dass ein Fan das Lied hochlädt und als Bildmaterial entweder Fotos der Band oder direkt das CD-Cover einstellt. Solche Hinweise können wiederum über Logoerkennung (Abschnitt 5.4.7) gut aggregiert werden. Denkbar ist also bei einer großen Menge an unbekanntem Liedgut, dass man Videos einander automatisch zuordnet, sobald sie das gleiche Lied verwenden oder die gleichen Bilder zeigen. Für eine redaktionelle Klassifikation ist dann eine Sichtung wesentlich vereinfacht, da durch diese Methode nicht nur ein Lied nur noch einmal angehört werden muss, sondern sich auch die Inhalte einer CD automatisch gruppieren würden.

Kann man bereits bekannte jugendschutzrelevante Bilder wiederfinden, auch dann, wenn sie verändert worden sind, zum Beispiel in Collagen? Unproblematisch sind häufig anzufindende Qualitätsverluste durch Kompression und Verkleinerung sowie darübergelegte Schrift (Abschnitt 5.5.5). Andere Veränderungen wie etwa einer Rotation oder einer Collage können, je nach Verfahren, schwieriger sein. Lösbar sind sie aber prinzipiell alle, wie man an dem wesentlich komplexeren Experiment für Gewaltfilme sehen kann (Abschnitt 5.3.2.3). Gerade für den Bereich Selbstgefährdung erscheint eine derartige Technik sinnvoll, da bei der manuellen Sichtung insbesondere die professioneller aufbereiteten Bilder wie etwa eine stilisierte Durchtrennung der Pulsader in sehr vielen Videos zu finden war.

Kann man unbekannte jugendschutzrelevante Bilder aufgrund ihres Inhaltes identifizieren? Das ist mit automatischen Verfahren nur sehr eingeschränkt möglich, wenn auch in Spezialfällen wie etwa der Verletzung des Unterarms sehr gute Ergebnisse erzielt werden können (Abschnitt 5.5.4). Auch wenn es dazu bereits Ansätze in aktuellen Forschungsprojekten gibt, erscheint eine allgemeine Erkennung von Gewalt oder ähnlich abstrakt gefassten Begriffen bis auf Weiteres nicht möglich.

Kann man bei Spielfilmen mit mehreren, altersdifferenzierten Fassungen eine Zuordnung vornehmen, auch wenn nur einzelne Szenen gezeigt werden? Ja – wenn die Original-Videos vorliegen (siehe Abschnitt 5.3.2).

Das Bestimmen der geschnittenen oder veränderten Szenen ist über die Tonspur mittels Audio-Fingerprinting weitestgehend unproblematisch (Abschnitt 5.3.2.2). Bei Spielfilmen mit einem geringen Anteil von Sprache ist selbst eine Zuordnung bei synchronisiertem Material denkbar. Mögliche Grenzen sind: extreme Frequenzstörungen (»leiern«) zum Beispiel bei alten VHS-Mitschnitten, mitsprechende Kommentatoren zum Beispiel bei Filmkritiken, oder zu kurze Szenen (3 Sekunden oder kürzer). Die zugrundeliegende Technologie ist, insbesondere durch großes Interesse der Musikindustrie, sehr gut erforscht und wird aktiv weiterentwickelt. Der Echtzeitfaktor der Verarbeitung ist zwar üblicherweise geringer als die Länge des zu betrachtenden Videos, aber nicht gänzlich zu vernachlässigen.

Genauer, aber erheblich aufwändiger in allen Aspekten der Ressourcennutzung, also Zeit, Festplattenanforderung und Speicher, ist eine Erkennung mittels Video-Fingerprinting (Abschnitt 5.3.2.3). Allerdings können so auch Trailer oder Kommentare zu jugendschutzrelevanten Videos gut gefunden werden, was insbesondere auch bei realen Gewaltvideos interessant ist, die zum Beispiel eine neu unterlegte Musikspur beinhalten.

Reicht die Qualität einer automatischen Spracherkennung auf Videos aus, um auf deren Inhalt schließen zu können? Ohne entsprechend angepasste Modelle nur sehr eingeschränkt, da oft gleich mehrere Faktoren eine ausreichende Analysequalität erschweren. Auf Seiten von Let's Play Videos (Abschnitt 5.3.4) sind dies vor allem Hintergrundgeräusche, schlechte Tonqualität und besonderes Vokabular. Günstiger stellt sich die Situation für die exemplarisch analysierten Videos von Holocaust-Leugnern dar (Abschnitt 5.4.8). Prinzipiell sind alle auftretenden Probleme gezielt lösbar. Die hierzu notwendigen Schritte des speziellen Trainings und der Adaption der Verfahren sind allerdings als aufwändig einzustufen.

Die Verfahren bedürfen eines speziellen Hintergrundwissens und vor allem besonderen Trainingsmaterials. Sie werden bereits für andere Zwecke erfolgreich auch auf großen Mediatheken, zum Beispiel von Fernsehanstalten, eingesetzt. Wenn auch viele frei zugängliche Verfahren existieren, so ist das Trainingsmaterial selbst aber in der Regel nur kommerziell verfügbar. Wenn auch viele interessante Aspekte aus einem Video gewonnen werden könnten, ist eine Umsetzung der Technik daher eher als aufwändig einzustufen.

Reicht die Qualität einer automatischen Texterkennung auf Videos oder Bildern aus, um auf deren Inhalt schließen zu können? Die Aspekte zur Bewertung der automatischen Texterkennung auf Videos ähneln in vielerlei Hinsicht denen zur automatischen Spracherkennung. Sie können ein sehr nützliches und mächtiges Werkzeug sein, funktionieren auf dem Material allerdings nicht ohne gezielte Anpassung (vgl. Abschnitt 5.5.3).

Bei der Erschließung von beispielsweise altem Zeitungsmaterial oder eingescannten Büchern lassen sich mittlerweile gute Ergebnisse auch bei ungewöhnlicher Schrift erzielen und auch handschriftliche Notizen stellen nicht mehr notwendigerweise eine Barriere dar. Dass die Technik gut ausgereift ist, zeigt sich auch an den sogenannten Captchas, also künstlich generierten Buchstabencodes, die auf Webseiten zur Verifikation eines menschlichen Nutzers eingesetzt werden. Da die heutigen automatischen Verfahren die Buchstaben mittlerweile ebenfalls gut erkennen können, sind heutige Captchas meist schon so schwierig ausgelegt, dass auch eine menschliche Erkennung nicht immer einwandfrei machbar ist.

5.7 Zusammenfassung

Wir haben uns in diesem Kapitel mit der Bewertung von technischen Verfahren zur Erkennung und Altersklassifikation von jugendschutzrelevantem Material aus den Bereichen Gewalt, Extremismus und Selbstgefährdung beschäftigt. Neben einer gesamtheitlichen Betrachtung ist ein maßgeschneidertes Vorgehen für die jeweiligen Unterkategorien als sehr sinnvoll zu betrachten, muss aber dem erhöhten Aufwand in der Erstellung und Pflege der einzelnen Verfahren kritisch gegenübergestellt werden.

Das Themenspektrum »Gewalt« ist inhaltlich am vielfältigsten, was eine automatische Erkennung erschwert. Das spiegelt sich auch in den Ergebnissen der Textklassifikation wieder, die sich jedoch durch ausreichend viele Beispiele für jede Unterkategorie deutlich zuverlässiger umsetzen ließe. Für aufwändig produziertes Multimedia-Material, wie etwa besondere Fotos oder Spielfilme, ist eine

Datenbank mit Fingerprints sehr hilfreich und kann auch zur Altersklassifikation von Ausschnitten gut eingesetzt werden. Ebenfalls ein wichtiger Baustein ist das Vorhalten von Fingerprints bei Filmen mit besonders drastischer Gewaltdarstellung, da die Anzahl der betroffenen Videos insgesamt als begrenzt angesehen werden kann.

Das Themenspektrum »Extremismus« ist leichter auf Basis des Inhalts zu erfassen. Das liegt insbesondere daran, dass in entsprechenden Kreisen bestimmte Keywords, Logos, Parolen oder Personen immer wieder auftauchen und entsprechend eindeutig von den Verfahren erfasst werden können. Interessant ist hierbei die Erkenntnis, dass auch Audio-Variationen, etwa durch Nachsingen von einschlägigen Liedern, gut zurückverfolgt werden können.

Das Themenspektrum »Selbstgefährdung« ist am stärksten durch nutzergenerierte Inhalte, etwa Foren, private Homepages und selbst hochgeladene Videos und Bilder geprägt. Die jeweiligen Merkmale in den Untergruppen, also etwa Anorexie oder selbstverletzendes Verhalten, sind sehr speziell und oft genug auch für sich genommen nicht eindeutig. Für einen Klassifikator mit hinreichender Präzision müssen hier die Merkmale der einzelnen Verfahren zusammengeführt und gemeinsam bewertet werden.

Eine wesentliche Erkenntnis der Textanalyseverfahren ist, dass sich eine Suche nach einschlägigem Material durch Keywords in erheblichem Maße durch technische Verfahren unterstützen lässt. Besonders Webseiten aus dem thematischen Umfeld der bekannten Begriffe aus Anas Brief o.ä. lassen sich nicht ganz so leicht manuell über Schlagwörter identifizieren. Ein modellbasiertes automatisches Verfahren, das auf diesem Umfeld trainiert wurde, kann also eine große Unterstützung auch bei der manuellen Recherche und Einstufung darstellen.

Es konnte nicht abschließend gezeigt werden, inwieweit Textklassifikation dazu in der Lage ist, Webseiten aus den betrachteten Teilbereichen altersklassifiziert einzustufen. Allerdings ist vor allem eine hohe Präzision der Verfahren für die Aussage »ab 18« zu beobachten, es wird also selten ein Overblocking aufgrund der extrahierten Merkmale stattfinden. So können in großem Maßstab und mit hinreichender Sicherheit besonders drastische Webinhalte automatisch blockiert werden. Die Experimente haben aber auch gezeigt, dass insbesondere bei heterogenem Material viele Webseiten nicht erkannt werden, wenn die Modelle nicht ausreichend trainiert sind. Im Allgemeinen gilt: je näher sich die Kategorien inhaltlich sind, desto mehr Beispiele werden benötigt.

Fingerprinting-Verfahren finden sich in allen Themen wieder. Ihr Einsatz ist für nahezu jegliche Art von Material sinnvoll. Die Altersklassifikation erfolgt hier implizit, indem bereits einmal gesichtete Daten in entsprechende Kategorien eingeteilt werden, um diese automatisch auf das leicht veränderte Material zu übertragen.

Logoerkennung ist insbesondere für die Kategorien »Gewalt« und »Extremismus« sinnvoll. So ist eine Datenbank bestehend aus Symbolen, die extremistischen Strömungen zugeordnet werden können, problemlos aufsetzbar. Die Symbole können dann auch verfremdet oder überdeckt auf Fotos und in Videos wiedererkannt werden. Erweitert durch Logos, die im Gewaltkontext von einschlägigen Foren und Sammelseiten verwendet werden, stellt die Logoerkennung ein mächtiges und praxistaugliches Werkzeug dar.

Differenziert zu betrachten sind die Verfahren der Textextraktion sowohl auf Bild- als auch auf Audioebene. Zwar böte bei den vorliegenden Inhalten eine gute Erkennung auch hier wichtige Hinweise auf die Natur der Daten, doch müssen die Verfahren dafür gezielt und verhältnismäßig aufwändig angepasst werden.

Generell lassen die Ergebnisse aber vor allem einen Schluss zu: heutige technische Verfahren können in wesentlichen Aspekten die redaktionelle Klassifikation ergänzen und unterstützen. Hierzu zählen insbesondere folgende Ansätze:

- das Aufspüren problematischer Keywords und die Wartung existierender Verfahren kann automatisch erfolgen;
- bekannte jugendschutzrelevante Inhalte können in allen Modalitäten (Text, Audio, Bild) wiedererkannt werden und müssen nicht jeweils erneut redaktionell gesichtet werden;
- auch für spezielle Themenbereiche können Textklassifikatoren so auf sprachliche Charakteristika des Genres eingestellt werden, dass auch unbekannte, textuelle Inhalte mit geringem Overblocking automatisch erkannt und ggf. getrennt nach Altersgruppen eingeordnet werden;
- die manuelle Klassifikation kann durch Vorfilterung und Priorisierung unbekannter Inhalte wesentlich effizienter gestaltet werden.

Die Genauigkeit von vielen Verfahren ist für eine Vielzahl an Einsatzgebieten mehr als ausreichend und die meisten Verfahren können auch für große Datenbanken und Trainingsdatenmengen hochskaliert werden. Bei der stetig wachsenden Menge an verfügbaren Web-Inhalten stellen sie so ein wichtiges technisches Werkzeug zur Unterstützung des Jugendmedienschutzes dar.

TEIL III

WEGE IN DIE PRAKTISCHE UMSETZUNG

KAPITEL 6

INTEGRATION TECHNISCHER VERFAHREN

Nachdem in den vorangegangenen Kapiteln Möglichkeiten technischer Verfahren für die Klassifizierung von Web-Inhalten ausführlich vorgestellt und evaluiert worden sind, befasst sich dieses Kapitel mit Aspekten der technischen Integration in Jugendschutzprogramme beziehungsweise in technische Infrastrukturen, die für die Bewertung von Web-Inhalten durch den Einsatz solcher Verfahren notwendig sind. Das Kapitel ist wie folgt gegliedert:

In Abschnitt 6.1 wird die Integration technischer Verfahren mit Hinblick auf Effizienz, Effektivität und Potenzial für Echtzeitanalysen untersucht. Konkret werden mögliche Betriebsumgebungen betrachtet und diskutiert.

Im darauf folgenden Abschnitt 6.2 werden die gängige Praxis bei der konzeptionellen Gestaltung und Implementierung von Filterlisten, empfehlenswerte Erweiterungen beim Einsatz von technischen Verfahren und Anwendungsumgebungen für Filterlisten besprochen.

Anschließend werden in Abschnitt 6.3 Ansätze zur kollaborativen Bewertung von Webangeboten durch Eltern- und Expertennetzwerke mittels Crowdsourcing dargestellt. Chancen und Risiken sowie mögliche Varianten der Integration werden dargestellt.

Der abschließende Abschnitt 6.4 befasst sich mit weiteren Möglichkeiten für eine teilautomatisierte Qualitätssicherung. Dabei werden insbesondere Auswahlstrategien für manuell zu bewertende Webangebote untersucht.

6.1 Betriebsumgebungen für technische Verfahren

Der folgende Abschnitt betrachtet mögliche Betriebsumgebungen für den Einsatz technischer Verfahren unter besonderer Berücksichtigung der Faktoren Effektivität, Effizienz und Echtzeitfähigkeit. Im Fokus liegen dabei verschiedene konzeptionelle Varianten des Einsatzes der in den vorherigen Kapiteln vorgestellten technischen Verfahren. Diese reichen vom reinen nutzerseitigen Einsatz über Serverlösungen bis hin zur Offline-Analyse bei rechenintensiven Analyseverfahren. Dazu werden zunächst die zentralen Begriffe *Effektivität, Effizienz und Echtzeitfähigkeit* im Kontext der Integration technischer Verfahren im Jugendmedienschutz definiert.

Im Folgenden wird immer dann von einem *Gesamtsystem* gesprochen, wenn nicht einzelne technische Verfahren betrachtet werden, sondern der vollständige Prozess der Bewertung und Klassifikation von Web-Inhalten einschließlich aller nötigen technischen und manuellen Schritte. Zu diesen Schritten zählen beispielsweise

- das Zusammentragen potentiell zu bewertender Webseiten (etwa durch einen Webcrawler oder durch Sammeln von Nutzerfeedback),
- die Auswahl und Priorisierung der Kandidaten,
- die manuelle Bewertung durch Experten oder die automatische Bewertung durch technische Verfahren,
- die Aggregierung der Bewertung in Filterlisten,
- die Filterung der Inhalte vor der Auslieferung an den Nutzer.

Effektivität: Zunächst lässt sich ein Gesamtsystem zur Bewertung von Web-Inhalten hinsichtlich der Effektivität der Gesamtlösung bewerten. Eine höhere Effektivität bedeutet hierbei, dass das übergeordnete Ziel des Systems zu einem höheren Grad erreicht wird: Kinder werden seltener mit jugendschutzrelevanten Web-Inhalten konfrontiert. Systeme mit geringer Effektivität können trotzdem interessante Lösungen darstellen (wenn sie zum Beispiel geringe Kosten verursachen und trotzdem über den Status Quo hinausgehen), Systeme mit hoher Effektivität sind nicht zwangsläufig attraktiver (wenn sie zum Beispiel zu hohem personellen Aufwand oder hohen Betriebskosten führen). In Kapitel 5 wurden verschiedene technische Verfahren bereits hinsichtlich ihrer individuellen Effektivität betrachtet.

Effizienz: Zunächst kann für die technischen Verfahren eine informationstechnische Effizienz im Sinne der Informatik bestimmt werden: effiziente algorithmische Verfahren lösen die gleiche Aufgabe schneller (oder mit weniger Speicherbedarf) als ineffiziente Verfahren (vgl. Abschnitt 2.2.4). Diese informationstechnische Effizienz eines Verfahrens hat somit einen direkten Einfluss darauf, wie viel Rechenleistung für eine definierte Aufgabe benötigt wird – und zwangsläufig, wie hoch die Betriebskosten sind. Darüber hinaus muss auch die Effizienz im Sinne der Wirtschaftlichkeit eines Gesamtsystems betrachtet werden. Bei der Kombination verschiedener Klassifikationsverfahren muss die Effizienz einzelner Algorithmen von der Effizienz des Gesamtsystems unterschieden werden. Ein Algorithmus kann beispielsweise eine sehr hohe informationstechnische Effizienz besitzen, aber aufgrund eines geringen Beitrags zur Leistung des Gesamtsystems im Vergleich zu hohen Lizenzkosten unwirtschaftlich sein. Eine Effizienzsteigerung eines Gesamtsystems wird dann erreicht, wenn unter Zuhilfenahme eines technischen Teilsystems bei mindestens gleichbleibender Bewertungsqualität und bei höchstens gleichen Kosten verglichen mit einer rein manuellen Bewertung mehr Web-Inhalte klassifiziert werden können und damit Kinder und Jugendliche weniger häufig für sie ungeeignete Webseiten aufrufen.

Echtzeitfähigkeit: Die Echtzeitfähigkeit beschreibt ebenfalls unterschiedliche Eigenschaften technischer Systeme. Bei der Vorstellung von Multimedia-Klassifikationsverfahren wird oft davon gesprochen, dass die *Analyse in Echtzeit durchgeführt werden kann* (vgl. Kapitel 5). In diesem Fall wird damit der Umstand umschrieben, dass zum Beispiel ein Analyse-Algorithmus für die Verarbeitung eines Videos mit der Länge von einer Minute auch eine Minute Analysezeit benötigt. Somit wird eigentlich die algorithmische Effizienz der Verfahren in Echtzeitfaktoren angegeben wie etwa *halbe Echtzeit* oder *dreifache Echtzeit*. Bei informationstechnischen Systemen im Allgemeinen sind Echtzeitsysteme allerdings solche Systeme, die bestimmte quantitativ festgelegte Echtzeitanforderungen erfüllen. Dabei werden harte fehlersichere, harte fehleroperationelle und weiche Echtzeitsysteme unterschieden. Bei harten fehlersicheren Echtzeitsystemen gilt das Überschreiten einer definierten Antwortzeit als Fehler (Beispiel: Ampelsteuerung), bei harten fehleroperationellen Echtzeitsystemen ist ein Überschreiten nicht unmittelbar ein Fehler und wird von Fall zu Fall ausgewertet (Beispiel: Autopiloten-Systeme) und bei weichen Echtzeitanforderungen gelten die quantitativen Restriktionen eher als Qualitätskriterium (Beispiel: Suchmaschinen). Eine typische weiche Echtzeitanforderung an eine Suchmaschine wäre zum Beispiel »75 Prozent aller Suchanfragen müssen in weniger als 500 Millisekunden und 95 Prozent in weniger als einer Sekunde beantwortet werden«. Diese Begrifflichkeiten sind durch die ISO/IEC-Norm 2382[1] definiert. Insofern gibt es im Bereich des technischen

[1] siehe http://www.iso.org/iso/home/search.htm?qt=2382&published=on, Stand 1.12.2012

Tabelle 6.1
Technische Verfahren und ihr Potenzial für Echtzeitanalyse (vgl. Kapitel 5).

Betriebsmodus	Technische Verfahren
Echtzeitfähig bei Betrieb beim Nutzer	Textklassifikation
Echtzeitfähig als Cloud-Dienst	Fingerprinting, Logoerkennung
Nicht für Echtzeitbetrieb geeignet (Textverfahren)	Topic Clustering, Keyword Generierung
Nicht für Echtzeitbetrieb geeignet (Multimediaverfahren)	Spracherkennung, Texterkennung, Konzepterkennung

Jugendmedienschutzes nur weiche Echtzeitkriterien. Weder bei der Analyse von Web-Inhalten auf Servern der Jugendschutzprogramm-Hersteller sind harte Echtzeitsysteme – allerdings algorithmisch effiziente Systeme – notwendig noch bei der Filterung von Web-Inhalten in den Jugendschutzprogrammen, wenn sie beim Nutzer ausgeführt werden. Das bedeutet nicht, dass diese Vorgänge nicht dennoch äußerst zeitkritisch sein können. Geeignete Werte, wie etwa die maximal tolerierbare Verzögerung zwischen Anfrage und tatsächlicher Darstellung eines Webangebots im Browser, wenn ein Jugendschutzprogramm aktiviert ist, müssen durch Usability-Experten und Hersteller im konkreten Anwendungsfall bestimmt werden. Möglicherweise können zur Maximierung des Schutzniveaus größere Verzögerungen in Kauf genommen werden als etwa bei modernen Web-Suchmaschinen.

Für den Betrieb technischer Verfahren werden in diesem Abschnitt drei verschiedene Konfigurationen betrachtet: Betrieb beim Nutzer, Betrieb als Proxy- oder Cloud-Dienst in einer skalierbaren Serverumgebung und der reine Offline-Betrieb für nicht echtzeitfähige Verfahren. Diese Betriebsmodi werden im Folgenden hinsichtlich Effizienz und Echtzeitanforderungen untersucht, die Vor- und Nachteile diskutiert und die in Kapitel 4 vorgestellten Verfahren eingeordnet. Tabelle 6.1 gibt einen Überblick über die Echtzeitfähigkeit der Verfahren.

Verfahren, die für einen der oben genannten Betriebsmodi ungeeignet sind, sind dies aus einem oder beiden der folgenden Gründe: Entweder ist die Verarbeitungzeit zu hoch, um akzeptable Antwortzeiten zu ermöglichen, oder das entsprechende Verfahren ermöglicht keine unmittelbare Entscheidung, sondern liefert nur Indizien zur Unterstützung der eigentlichen redaktionellen Klassifikation. Während Verfahren wie Fingerprinting Aussagen treffen wie »dieses Video ist zu 95 Prozent identisch mit einem bereits bekannten als ab 18 Jahre klassifizierten Video«, müssen anderen Verfahren, etwa der Sprach- oder Logoerkennung, weitere Bewertungsstufen folgen, um aus den gewonnenen Indizien Entscheidungen abzuleiten. Beispielsweise kann ein Logoerkennungsverfahren nur den Hinweis liefern, dass ein Video rechtsradikale Symbole enthält, nicht aber, ob es sich um ein rechtsradikales Propagandavideo handelt.

6.1.1 Betrieb beim Nutzer

Ein technisches Verfahren kann direkt beim Nutzer als Teil eines Jugendschutzprogramms ausgeführt werden. Von den in Kapitel 4 vorgestellten Verfahren ist dabei die Textklassifikation ein möglicher Kandidat. Je nach Auswahl der zu vergleichenden Webseiten-Merkmale können viele Dokumente pro Sekunde analysiert und in Kategorien einsortiert werden (vgl. Abschnitt 4.4.2). Allerdings ist der Betrieb der Klassifikation beim Nutzer aus mehreren Gründen problematisch.

Zunächst besteht das Problem, dass neben dem Jugendschutzprogramm weitere Software-Module eingeführt werden müssten, die gegebenenfalls vom Betriebssystem abhängig und womöglich gar

nicht für die gewünschten Plattformen verfügbar sind. Selbst Versionsunterschiede innerhalb eines Betriebssystems (etwa von Windows Vista zu Windows 7) haben oft Folgen für die Kompatibilität der Software. Eine Portierung der Verfahren auf andere Betriebssysteme oder Versionen ist gegebenenfalls mit hohem Aufwand verbunden und steht eventuell nicht im Verhältnis zum erwarteten Nutzen. Dies setzt zudem voraus, dass die Lizenzbestimmungen oder die Hersteller der Verfahren eine Portierung auch zulassen.

Gegen einen solchen Betrieb spricht außerdem, dass die genannten Klassifikationsverfahren häufig auf statistischen Modellen beruhen, die ebenfalls für den Betrieb im Jugendschutzprogramm ausgeliefert werden müssen. Je nach Verfahren können diese Modelle allerdings mehrere Gigabyte groß werden – die Installationspakete würden somit für den Nutzer abschreckende Größen annehmen und den Datenverkehr auf den Servern des Jugendschutzprogramm-Herstellers unnötig erhöhen.

Weiterhin erschwert die Ausführung beim Nutzer das Sammeln von Feedback für die Verbesserung der technischen Verfahren (Active Learning) und die Vermeidung redundanter Analyse. Jede angestoßene Analyse liefert nützliche Informationen, zum Beispiel:

- Ob das Webangebot bereits durch ein technisches Verfahren klassifiziert worden ist,
- wie oft das Angebot über ein Jugendschutzprogramm aufgerufen wird,
- wann und mit welcher Version des Verfahrens es zuletzt klassifiziert worden ist und
- das Ergebnis der Analyse selbst.

Diese Informationen erfordern im Fall des Betriebs direkt beim Nutzer zusätzliche Kommunikation mit einem Server, der diese Daten entgegen nimmt und speichert – sofern keine »anonyme Analyse beim Nutzer« zugelassen werden soll. Das Abrufen dieser Informationen durch das Jugendschutzprogramm erfordert dann auch eine Kommunikation mit dem Server, alternativ können die Jugendschutzprogramme in kurzen regelmäßigen Abständen aktualisiert werden (vgl. Signaturdateien bei Virenscannern).

Unter den vorgestellten technischen Verfahren können Textklassifikationsverfahren auch beim Betrieb auf dem Nutzerrechner weiche Echtzeitanforderungen im Millisekunden-Bereich erfüllen. Dieses Vorgehen ist allerdings nicht sehr effizient, da es möglicherweise hohe Entwicklungskosten für die Bereitstellung und Aktualisierung der Verfahren im Jugendschutzprogramm verursacht und die Analysen (außer über die beschriebene zusätzliche Kommunikation mit einem Server) anonym und eventuell redundant beim Nutzer durchgeführt werden.

6.1.2 Betrieb als Proxy- oder Cloud-Dienst in einer skalierbaren Serverumgebung

Da Jugendschutzprogramme die aktuell ausgeführten Webzugriffe auf nicht altersgerechte Inhalte filtern, kann eine permanente aktive Internetverbindung vorausgesetzt werden. Eine Analyse von lokalen Inhalten, wie beispielsweise bei Virenscannern, ist nicht erforderlich. Daher kann auch die Möglichkeit erwogen werden, einzelne Klassifikationsverfahren nicht beim Nutzer selbst, sondern über

1. einen im Netz bereitgestellten Cloud-Dienst oder
2. über einen Proxyserver zu betreiben.

Mit einem Cloud-Dienst ist ein Dienst gemeint, der nicht lokal, zum Beispiel als Computerprogramm, sondern über ein beliebig skalierbares verteiltes Rechnernetz bereitgestellt wird. Ein Proxyserver ist eine Kommunikationsschnittstelle, die bei Bedarf zwischen zwei Rechnern im Internet geschaltet

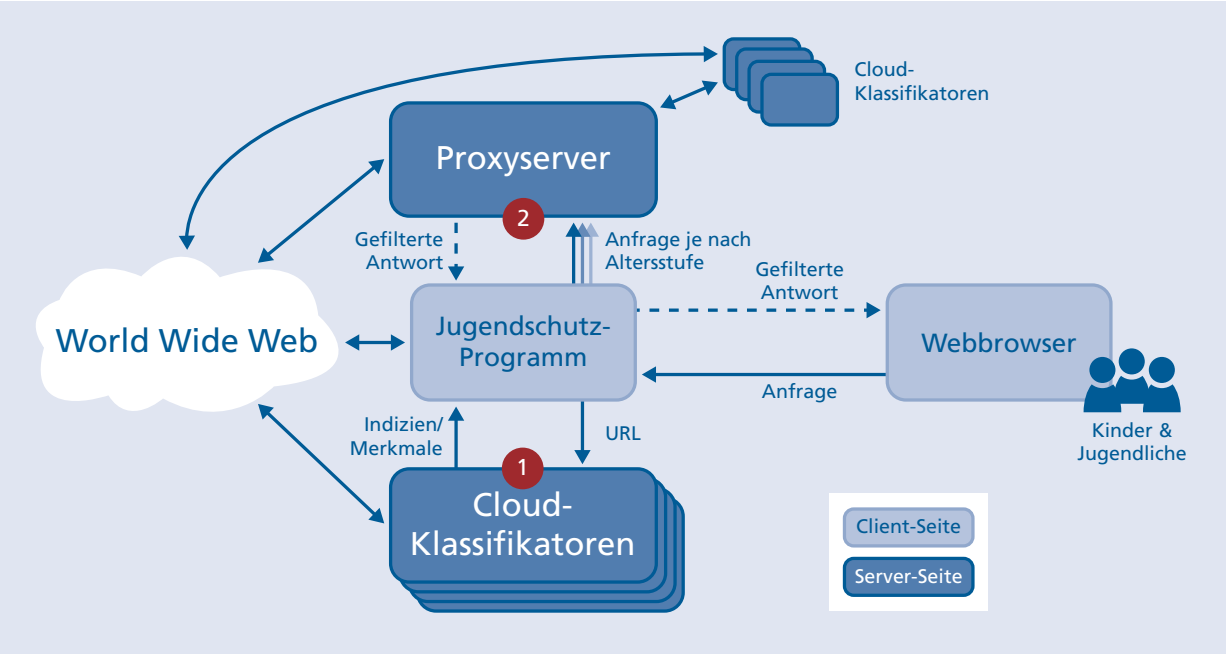


Abbildung 6.1
Varianten für den Einsatz von technischen Verfahren über Proxy- oder Cloud-Lösungen.

wird. Typischerweise werden Proxies für die Anonymisierung, für Lastverteilung oder für Filterung von HTTP-Aufrufen verwendet. Im einfachsten Fall (Anonymisierung) dient der Proxyserver nur dazu, die Adressen der Kommunikationspartner voreinander zu verbergen, das heißt, Partner A und Partner B kommunizieren nur über den Proxy, der als Vermittler dient, wobei Partner A und B nichts voneinander wissen. Proxyserver übernehmen je nach Anwendungsfall allerdings auch aktiv Filter- oder Weiterleitungsentscheidungen. Vergleichbar zu dem hier vorgestellten Proxyserver wäre ein Unternehmensproxy so konfiguriert, dass er Zugriffe auf soziale Netzwerke durch die Mitarbeiter unterbindet. Alle gängigen Webbrowser und auch viele Netzwerk-Router unterstützen die Konfiguration von Proxyumleitungen. Außerdem basiert das Funktionsprinzip von Jugendschutzprogrammen darauf, den Datenverkehr vor dem Versand über die Netzwerkschnittstelle des Geräts zu untersuchen und, je nach Ziel des HTTP-Aufrufs, diesen zu unterbrechen und eine kindgerechte Warnung zurückzuliefern (bei Aufruf eines nicht altersgerechten Webangebots) oder den Aufruf passieren zu lassen (bei unbedenklichen Webangeboten). Das Jugendschutzprogramm fungiert dabei als lokaler Proxyserver. Die eigentliche Auswertung des Aufrufs muss dabei nicht zwangsläufig in einem entfernten, durch den Hersteller des Jugendschutzprogramm betriebenen, Proxyserver geschehen. Die beiden hier betrachteten Varianten sind in Abbildung 6.1 dargestellt.

Die Abwicklung der lokalen Webbrowser-Anfragen erfolgt in beiden Fällen über ein lokal installiertes Jugendschutzprogramm. Das Programm prüft für das Ziel der Anfrage zunächst immer

- ob es auf der Filterliste des BPjM-Moduls (vgl. Abschnitt 1.2.1.3) enthalten ist,
- ob es auf der individuellen Eltern-Filterliste enthalten ist,
- ob eine vorhandene Auszeichnung durch das age-de.xml Label existiert,
- ob es auf einer lokal oder auf einem Webserver vorgehaltenen Hersteller-Filterliste enthalten ist.

Wenn keiner der soeben gelisteten Fälle eingetreten ist, wäre es eine Möglichkeit (Variante 1), die aufgerufene URL an einen oder mehrere Klassifikatoren zu senden, die als Cloud-Dienste bereit stehen. Die einzelnen Klassifikatoren würden das Webangebot der zugehörigen URL untersuchen, Indizien und Merkmale für eine Altersklassifikation einsammeln und diese an das Jugendschutzprogramm zurückliefern. Die eigentliche Auswertung würde dann durch das Jugendschutzprogramm

erfolgen. Je nach Einstufung wird die Anfrage entweder normal weitergeleitet oder eine Warnseite angezeigt. Eine andere Möglichkeit (Variante 2) wäre die Delegation der gesamten Anfrage an einen Proxyserver.[2] Dem Proxyserver steht so der gesamte HTTP-Request zur Verfügung, inklusive aller Parameter und Cookies (zur Bedeutung dieser Informationen siehe Abschnitt 6.2). Der Proxyserver würde die Anfrage auswerten, gegebenenfalls einzelne Web-Inhalte an technische Klassifikatoren delegieren, die gesammelten Indizien und Merkmale sammeln, eine Altersklassifikation vornehmen und die Anfrage je nach Entscheidung passieren lassen oder eine Warnseite zurückliefern. Der Unterschied zwischen den beiden Varianten liegt nur in der Auswertung der Indizien und der existierenden Filterlisten. Bei Variante 1 findet die Auswertung im Jugendschutzprogramm, bei Variante 2 im Proxyserver statt. Bei der Variante 2 wäre außerdem denkbar, die aufgelisteten Vorprüfungen (BPjM-Modul, Elternlisten, age-de.xml Auswertung, Hersteller-Listen) teilweise oder vollständig im Proxyserver durchzuführen.

Variante 1 hat den Vorteil, dass kein Proxyserver betrieben werden muss und die Auswertung auf der Hardware des Nutzers durchgeführt wird. Die Rechenleistung des Nutzercomputers ist ausschlaggebend für die Verzögerung zwischen Anfrage und Auslieferung der Webseite, aber es besteht nicht wie bei Variante 2 das Problem, dass die effiziente Beantwortung von Anfragen für viele parallele Nutzer sichergestellt werden muss. Dennoch besteht die Möglichkeit, ausgelagerte Klassifikations-Dienste für die Bewertung von Webangeboten zu nutzen. Abgesehen von der Nutzung von Klassifikationsdiensten entspricht dies, also die vollständige Durchführung der Bewertung beim Nutzer, auch der aktuellen Praxis in den Jugendschutzprogrammen. Möchte der Hersteller für die Qualitätssicherung und Erweiterung der Filterlisten Nutzungsstatistiken sammeln (zum Beispiel, welche Seiten durch die Nutzer aufgerufen worden sind), so müssen diese Daten zusätzlich über eine Rückschnittstelle an einen Server geschickt werden.

Der Vorteil von Variante 2 liegt in der Auslagerung der Klassifikation auf einen zentralen Webserver und den damit sehr kurzfristig möglichen Updates für Neuklassifikationen. Zum Beispiel könnte ein Experte ein neues Webangebot als »ab 18 Jahre« kennzeichnen und dies in der im Proxyserver verfügbaren Filterliste vermerken. Alle bei den Nutzern installierten Jugendschutzprogramme würden somit ab dem nächsten Aufruf des Webangebots auf diese Klassifikation reagieren. Werden die Filterlisten lokal vorgehalten, ist zunächst ein Update des Jugendschutzprogramms (oder genauer: der enthaltenen Filterliste) notwendig. Da in dieser Variante jeder Internet-Zugriff von einem Jugendschutzprogramm über den Proxyserver geschieht, können die Hersteller an dieser Stelle auch entscheiden, ob sie neben dem reinen Filterlisten-Abgleich der aufgerufenen URL auch inhaltsbasierte Filterung über technische Verfahren integrieren. Somit könnte etwa ein Prozess implementiert werden, der einen Aufruf zuerst mit einer Filterliste abgleicht und, wenn das Webangebot dort nicht aufgeführt ist, anschließend eine Textklassifikation zur unmittelbaren Filterentscheidung anstößt.

Da der Proxyserver autonom auf Anfragen reagieren kann, könnte diese Variante auch in Netzwerkgeräten betrieben werden – zum Beispiel, indem der Router so konfiguriert wird, dass er Anfragen zunächst immer über den Proxyserver leitet. Hierbei sind in jedem Fall die Interessen der Nutzer zu berücksichtigen. Beispielsweise wird im JusProg Jugendschutzprogramm der Nutzer explizit gefragt, ob er zur Unterstützung von Produktverbesserungen Nutzungsstatistiken an das JusProg-System schicken will. Da diese Daten über einen derartigen Proxyserver automatisch erhoben werden können, muss der Nutzer darüber explizit in Kenntnis gesetzt werden.

Beide Varianten haben gegenüber dem Betrieb von technischen Verfahren beim Nutzer den Vorteil, dass der Hersteller volle Kontrolle über die nötige Infrastruktur hat, in deren Rahmen die Verfahren zum Einsatz kommen. Dazu zählt in erster Linie die freie Wahl des Betriebssystems, aber auch die Skalierung des Systems und die Tatsache, dass die technischen Verfahren nicht mit dem Jugend-

[2] Dies beinhaltet die im Jugendschutzprogramm aktuell eingestellte Altersstufe des Nutzers. Die Weiterleitung an den Proxyserver könnte beispielsweise über unterschiedliche Ports realisiert werden.

schutzprogramm ausgeliefert werden müssen. Dies erleichtert die Identifikation von Verfahren mit geeigneten Lizenzmodellen und erlaubt schlanke Endnutzer-Anwendungen.

Auf der anderen Seite ist aber zu beachten, dass zusätzliche technische Maßnahmen nötig sind, um auch bei entfernten Filterentscheidungen die Nutzerautonomie zu gewährleisten. Je nach Anwendungsfall müssen entweder in Anwendungen oder Netzwerkgeräten neue Funktionen geschaffen werden, um die Filterung für unterschiedliche Altersstufen zu konfigurieren (Kinder oder Jugendliche unterschiedlichen Alters), temporär zu deaktivieren (Surfen für Eltern) oder um eigene Regeln zu erweitern (Elternlisten).

In Abhängigkeit von den durch die Hersteller unter Usability-Aspekten festzulegenden Echtzeitanforderungen haben Textklassifikations-, Fingerprinting- und Logoerkennungs-Verfahren prinzipiell das Potenzial, eine On-The-Fly Analyse von Web-Inhalten zu ermöglichen. Textklassifikationsverfahren werden zum Beispiel bereits seit mehreren Jahren in Spamfiltern eingesetzt. Die Fingerprinting- und Logoerkennungsverfahren können Aussagen über die Identität von Web-Inhalten treffen und eignen sich daher, um direkte Filterentscheidungen zu treffen. Die algorithmische Eignung solcher Verfahren für den Betrieb unter weichen Echtzeitbedingungen kann bereits an den verschiedensten kommerziellen Projekten beobachtet werden, wie etwa Shazam[3] (Audio-Fingerprinting) oder Google Search-By-Image[4] (Bild-Fingerprinting, Logoerkennung).

6.1.3 Offline-Betrieb

Die verbleibenden Verfahren (siehe Tabelle 6.1) besitzen entweder eine sehr hohe Latenz (lange Dauer zwischen Start des Analysevorgangs und Ergebnis der Analyse) oder sie können keine unmittelbaren Klassifikations-Entscheidungen treffen. Sie können allerdings zur Vorauswahl der manuell zu bewertenden Web-Inhalte dienen, da sie Indizien liefern, um im Gesamtklassifikationsprozess zur Erstellung von Filterlisten beizutragen. Für die Integration in das bisher skizzierte Gesamtsystem sind zwei Varianten denkbar, die in Abbildung 6.2 dargestellt sind.

Wie in den weiter oben skizzierten Betriebsumgebungen gibt es ein beim Nutzer installiertes Jugendschutzprogramm, dass die Anfragen der Nutzer entgegennimmt und entweder

1. direkt anhand einer lokal oder online vorgehaltenen Filterliste bewertet (gegenwärtige Praxis) oder
2. die Bewertung an einen Proxyserver delegiert.

Die Filterlisten werden dabei offline aus einer Kombination von technischen Verfahren und manuellen Experten-Bewertungen erzeugt. Die beiden dargestellten Varianten unterscheiden sich dabei in der Auslieferung der aktualisierten Filterlisten an die Jugendschutzprogramme. Bei der Nutzung der lokalen Filterlisten (Variante 1) wird die Aktualisierung über ein Update für alle beim Nutzer betriebenen Jugendschutzprogramme ausgerollt. Bei Betrieb einer zentralen Filterliste (Variante 2) kann das Update direkt auf der zentralen Liste durchgeführt werden und steht somit allen Jugendschutzprogrammen direkt zur Verfügung. Die generellen Vor- und Nachteile beider Betriebsvarianten wurden im vorherigen Unterabschnitt bereits diskutiert.

[3] siehe http://www.shazam.com/, Stand 1.12.2012
[4] siehe http://www.google.com/insidesearch/features/images/searchbyimage.html, Stand 1.12.2012

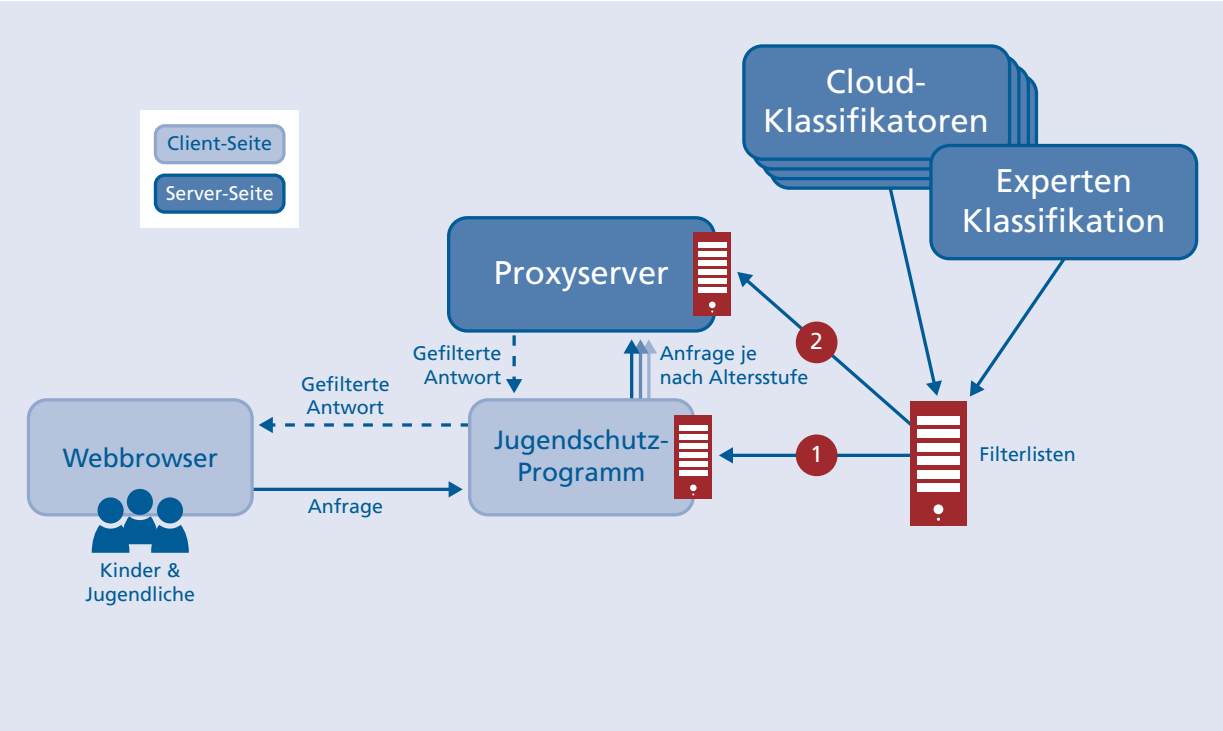


Abbildung 6.2
Varianten für den Einsatz von offline betriebenen technischen Verfahren.

6.2 Konzeption und Implementierung von Filterlisten

Dieser Abschnitt befasst sich mit Aufbau und Implementierung von Listen, die für das Filtern von Web-Inhalten eingesetzt werden. Nach einer kurzen Betrachtung der gegenwärtig verbreiteten Praxis in der Gestaltung dieser Listen (Abschnitt 6.2.1) werden konkrete Vorschläge zu einer Erhöhung der bestehenden Granularität (Abschnitt 6.2.2) beim Einsatz technischer Verfahren getroffen. Dann wird auf die Problematik eingegangen, die durch das zugrunde liegende Web-Protokoll entstehen kann (Abschnitt 6.2.3). Schließlich widmet sich Abschnitt 6.2.4 der Frage, welche Stellen in der Datenübertragung vom Inhalteanbieter zum Endnutzer sich für die Filterung von Inhalten anbieten.

6.2.1 Status Quo

Die aktuell anerkannten Jugendschutzprogramme verwenden Filterlisten zur Filterung der Web-Inhalte. Die Listen orientieren sich in der Regel an ganzen Web-Angeboten. Jede Filterliste besteht aus einer Liste von IP-Adressen und Domain-Namen, die für die jeweilige Alterskategorie (inklusive aller niedrigeren Altersstufen) blockiert werden sollen. Wenn eine Webseite aufgerufen wird, wird der Domain-Name beziehungsweise die IP-Adresse des Serverhosts mit den Einträgen in den Filterlisten verglichen und der Aufruf gegebenenfalls blockiert.

Da Webauftritte meist durch Domains abgegrenzt werden, bedeutet Filterung nach Domains und IP-Adressen in der Praxis eine Filterung nach Webauftritten. Für diejenigen Webauftritte, die thematisch homogen sind und daher zumeist die gleiche Jugendschutzrelevanz aufweisen, stellt die Filterung nach Domains eine geeignete Granularität dar. Die Vorteile liegen in einer schnellen und einfachen Erstellung der Listen sowie einer einfachen Überprüfung entsprechender Aufrufe von Webseiten.

An seine Grenzen stößt diese Vorgehensweise allerdings für entsprechend heterogene Webauftritte, wo sich die Inhalte zum Teil wesentlich voneinander unterscheiden, wie zum Beispiel größere Online-Videoportale. So ist zum Beispiel YouTube in einer der derzeitigen Filterlisten pauschal »ab 12« klassifiziert. Neben Inhalten, die auch für jüngere Konsumenten geeignet erscheinen, zum Beispiel Kinderliedern, finden sich auf YouTube jedoch auch Inhalte, die erst ab 16 oder 18 geeignet sind, also etwa Gewaltdarstellungen oder erotisches Material, so dass die »ab 12« Klassifikation in erheblichem Maße zu Over- und auch Underblocking führt. Anders betrachtet gibt es mit solchen Filterlisten für sehr viele Webauftritte keinen Weg, um den gezeigten Inhalt eindeutig zu identifizieren.

Dies betrifft Webauftritte ohne age-de.xml Auszeichnung durch die Anbieter. Mit einer solchen Auszeichnung könnten durchaus unterschiedliche Altersstufen für einzelne Bereiche eines Webangebots vergeben werden. Für nicht ausgezeichnete Inhalte müssen die Jugendschutzprogramme aber auf eigene Filterlisten zurückgreifen.

Zusammenfassend lässt sich sagen, dass eine Liste von IP-Adressen und Domain-Namen für thematisch heterogene Webangebote eine zu ungenaue Granularität aufweist.

6.2.2 Erhöhung der Filtergranularität

Wie bereits aufgezeigt wurde, stellen thematisch heterogene Webauftritte eine besondere Herausforderung bei der Definition von Filterlisten dar, da mit Domains oft nur ganze Webauftritte gefiltert werden, jede Webseite des Webauftritts jedoch eine unterschiedliche Jugendschutzrelevanz innehaben kann. Für solche Webauftritte könnte mit einer höheren Filtergranularität wesentlich flexibler reagiert werden. Empfohlen wird daher, in Filterregeln nicht nur die Domain einer Webseite zu betrachten, sondern die Möglichkeit zu schaffen, alle Teile einer URL in Filterregeln mit aufzunehmen.

Eine Einführung von URLs in den Filterlisten bringt allerdings auch einige Herausforderungen mit sich, auf die gesondert eingegangen werden muss. So haben nicht zwangsweise alle Informationen in einer URL einen direkten Einfluss auf den dargestellten Inhalt. Beispielsweise können auf populären Videoportalen Informationen für die Darstellungsart und den Verlauf der Session in der URL enthalten sein, etwa in der Form &hd=1&feature=related&t=33s, das gezeigte Video ändert sich dadurch jedoch nicht. Das bedeutet aber auch, dass inhaltsunterscheidende Teile der URL jeweils an den Anbieter angepasst werden müssen, was nicht für alle Webauftritte umsetzbar erscheint.

Zu beachten ist aber, dass die Möglichkeit zur Filterung nach den weiteren URL-Teilen nicht für jeden Webauftritt genutzt werden muss. Thematisch homogene Webauftritte können weiterhin nach ihren Domains blockiert werden. Für Webauftritte, bei denen wie bei typischen Videoportalen mit User Generated Content die Vielzahl der Inhalte eine vollständige Liste kaum zulässt, kann eine generelle Altersklassifikation festgelegt werden, die immer dann greift, wenn ein einzelnes Video nicht mit einer gesonderten Altersklassifikation versehen ist. So kann gezielt das Over- und Underblocking reduziert werden.

Technische Umsetzung Zur Erhöhung der Filtergranularität wird empfohlen, dass die Filterlisten sowohl die Domain, den Pfad als auch die Query Parameter beinhalten können.[5] Für eine weiter erhöhte Flexibilität und Genauigkeit können zudem reguläre Ausdrücke verwendet werden, um die URL auf die eigentlich inhaltstragenden Merkmale zu reduzieren. So kann zum Beispiel ein Video unter der URL

[5] Die URL http://www.beispielportal.de/sport/?lang=de enthält zum Beispiel die Domain »beispielportal.de«, den Pfad »sport« und die Query Parameter »lang=de«.

(1) http://www.beispielVideoPortal.com/watch?v=lAl28d6tbko
(2) http://www.beispielVideoPortal.com/watch?v=lAl28d6tbko&hd=1&feature=related&t=33s

unter der Annahme, dass nur die Information v=lAl28d6tbko über den Inhalt des dargestellten Videos entscheidet, durch folgende Filterregel[6] erfasst werden:

(3) (.+\.)?beispielVideoPortal.com/watch?(.*&)?v=lAl28d6tbko(&.*)?

Die so gewonnene Flexibilität bringt allerdings auch neue Aspekte mit sich, die in der Verarbeitung der Filterliste zu berücksichtigen sind. Zunächst kann bei dieser Methode prinzipiell eine Ambiguität dadurch auftreten, dass eine Anfrage auf mehrere Filtereinträge zutrifft. Im obigen Beispiel wäre also die URL (2) sowohl durch den regulären Ausdruck (3) als auch durch die Aufnahme der Domain auf die Filterliste erfasst. Das bedeutet für die Praxis, dass den Einträgen in der Filterliste Prioritäten zugeordnet werden müssen. Sinnvoll erscheint in diesem Fall, dass der Filtereintrag mit der höchsten Spezifität, das heißt derjenige, der die meisten Zeichen trifft, andere Filterregeln überschreibt.

Daneben ist zu beachten, dass insbesondere bei Webauftritten mit einer großen Anzahl zu erwartender Einzelregelungen, wie eben bei Online-Videoportalen der Fall, auch eine effiziente Datenstruktur zur Überprüfung der Regeln von Nöten sein kann.

Zusammenfassend kann gesagt werden, dass heutige Filterlisten durch eine erhöhte Granularität entscheidend an Flexibilität gewinnen können. Die hier vorgeschlagenen Lösungsansätze sind zudem mit der heute gebräuchlichen Technik abwärtskompatibel, was bei einem Umstieg einen niedrigen initialen Aufwand bedeutet. Da die manuelle Klassifikation von Inhalten (das heißt einzelnen Webseiten) erheblich mehr Aufwand bedeutet als die manuelle Klassifikation von ganzen Webauftritten, ist die hier aufgezeigte Lösung als Ergänzung und nicht als Ersatz zu verstehen. Eine gezielte Erweiterung bestehender Filtereinträge für Webauftritte muss daher entsprechend priorisiert werden. Ein Szenario wäre die Altersklassifizierung eines Angebots, zum Beispiel eines Videoportals, auf Domainebene und eine zusätzliche optionale und priorisierte Bewertung von Einzelinhalten, wenn diese in eine andere Altersstufe eingeordnet werden müssen. Entscheidend ist hierbei auch, dass der Einsatz von automatischen Verfahren die Klassifikation von Webseiten unterstützen kann. Insbesondere durch die Fingerprinting-Verfahren (vgl. Abschnitt 4.5.1) können zum Beispiel einmal klassifizierte oder geblockte Videos auch auf anderen Videoportalen oder an anderen Stellen im selben Videoportal automatisch wiedergefunden werden und müssen nicht für jeden gespiegelten Inhalt erneut redaktionell klassifiziert werden.

6.2.3 Filterung von sicheren Verbindungen (HTTPS)

In den vorangegangenen Betrachtungen wurde davon ausgegangen, dass die Anfragen über das Protokoll HTTP durchgeführt werden. Eine steigende Anzahl an Webseiten wird heutzutage allerdings über das Protokoll HTTPS übertragen, zumal viele Webauftritte ihre Besucher bereits automatisch auf dieses Protokoll umleiten. Gegenwärtig verwenden weltweit etwa 10 Prozent der eine Million am häufigsten besuchten Webangebote das HTTPS Protokoll. Von den in dieser Liste enthaltenen de-Domains unterstützen es etwa 15 Prozent [Schleier und Schreiber, 2012]. HTTPS wird teilweise auch optional angeboten. Beispielsweise kann das Videoportal YouTube sowohl über HTTP als auch HTTPS genutzt werden. Gerade im Anbetracht der wachsenden Zahl mobiler Internetzugriffe steigt die Bedeutung des HTTPS-Protokolls als sicherer Kommunikationsweg. Auch große Anbieter wie Facebook haben erst kürzlich entsprechend nachgerüstet [Asthana, 2012].

[6] Der reguläre Ausdruck in diesem Beispiel verwendet die Perl-/Java-/Python-Syntax.

Bei einer HTTPS-Verbindung sind nur wenige Informationen nach außen sichtbar. Für eine höhere Granularität der Filterung auch für HTTPS-Verbindungen sind daher besondere entweder technische oder organisatorische Maßnahmen nötig, auf die in diesem Abschnitt eingegangen wird.

Der wesentliche Unterschied zwischen HTTP-Verkehr und HTTPS-Verkehr ist die vorgenommene Verschlüsselung, die lediglich vom Sender und vom Empfänger wieder entschlüsselt werden kann. Somit können alle anderen Rechner, durch die der Verkehr geleitet wird, lediglich die IP-Adresse und bestenfalls den FQDN des Servers, zu dem der Browser verbunden ist, einsehen. Wenn der Browser eine Verbindung zu einem Server aufbaut, weist der Server seine Identität mittels eines Zertifikats nach, in dem die Zugehörigkeit der IP-Adresse zu dem FQDN festgelegt ist. Falls ein Rechner sich als ein angefragter Server ausgibt, aber kein vertrautes Zertifikat vorweisen kann, wird der Browser die Verbindung zum Rechner üblicherweise abbrechen. Die Liste der Zertifikate wird dabei entweder vom Browser, vom Betriebssystem oder vom Benutzer vorgegeben und jedes Server-Zertifikat muss darauf letztendlich zurückzuführen sein.

Ein rein technischer Ansatz zur Filterung von HTTPS-Verbindungen wäre das Hinzufügen eines Jugendschutz-Servers als anerkannte Zertifizierungsstelle. Somit kann der Server entsprechende Verbindungen zunächst zurückhalten und filtern, bevor er die Verbindungen zu den ursprünglichen Servern aufbaut. Dieser Ansatz hat allerdings einige zusätzliche Konsequenzen:

- Das Jugendschutzprogramm übernimmt damit faktisch die Verantwortung für die Kontrolle der Verbindungssicherung. Das schließt das Melden von gegebenenfalls auftretenden Problemen sowie den Abbruch von Verbindungen bei unsicheren Servern mit ein.
- Die Filterung von HTTPS-Verbindungen ist ein zusätzliches Sicherheitsrisiko, da hier ein zusätzlich angreifbarer Server in den Kommunikationskanal eingebracht wird. Ein erfolgreicher Angriff auf diesen Server, etwa durch
  - die Entschlüsselung des privaten Schlüssels des Jugendschutzprogramms
  - oder durch Laufzeitzugriffe, um die Daten im Speicher zu überwachen,

  kann den Server kompromittieren. Übertragene Daten bei vermeintlich sicheren HTTPS-Verbindungen könnten so durch Viren oder Phisher gelesen und gegebenenfalls geändert werden.
- Jugendschutzprogrammhersteller übernehmen die Verantwortung für die Liste von vertrauten Zertifikaten von Browserherstellern. Falls das Jugendschutzprogramm sich nicht auf die Betriebssystemliste von vertrauten Zertifikaten verlässt, sondern eine eigene Liste verwaltet, muss die Liste bei Problemen zügig aktualisiert werden. In den letzten Jahren ist es mehrmals passiert, dass international anerkannte Zertifizierungsstellen erfolgreich angegriffen worden sind und somit die Zertifikate für bekannte Webauftritte gefälscht wurden. Browserhersteller haben innerhalb von einigen Tagen die Zertifikate der kompromittierten Zertifizierungsstellen aus ihren Listen genommen, Betriebssystemhersteller kurze Zeit später [Ars Technica, 2011].
- Benutzer müssen sich an ein neues Benutzerschnittstellen-Konzept zur Kontrolle des Verbindungssicherungsstatus gewöhnen. Da das Jugendschutzprogramm die Kontrolle der Verbindungssicherung übernimmt, sollte es ggf. auftretende Probleme dem Benutzer melden. Dazu stehen ihm, wenn keine Browserplugins oder ähnliches verwendet werden, die Browsersignale nicht zur Verfügung. Viele Browser benutzen das Anzeigen eines Vorhängeschloss-Symbols und das Färben des Icons der Webseite, der Domains oder ähnliches als Zeichen für eine sichere Verbindung. Teilweise sind diese Browser-Signale standardisiert [W3C, 2010]. Wenn das Jugendschutzprogramm den Verbindungssicherungsstatus auf eine andere Weise übermitteln muss, muss der Benutzer sich an die neuen Signale gewöhnen.

Zusammenfassend kann gesagt werden, dass das Einbinden eines Jugendschutzservers zum technischen Umgang mit HTTPS Verbindungen prinzipiell möglich wäre, in der Praxis allerdings eine Vielzahl an Herausforderungen aufwirft, da in eine höchst sicherheitsrelevante Umgebung eingegriffen wird. An dieser Stelle sollten daher eher organisatorische Maßnahmen wie die stärkere Ein-

bindung der Inhalte-Anbieter untersucht werden. In Jugendschutzprogrammen können alternativ HTTPS-Verbindungen blockiert oder zumindest auf FQDN-Ebene gefiltert werden.

6.2.4 Anwendungsumgebungen für Filterlisten

Neben dem Aufbau von Filterlisten sowie den Herausforderungen durch das gewählte Verbindungsprotokoll ist auch die Wahl der anvisierten Anwendungsumgebung ein entscheidender Faktor für die Integration von Filterlisten. Prinzipiell können Filterlisten an beliebigen Punkten zwischen Quelle (Anbieter der Inhalte) und Ziel (Endgerät) installiert werden, mit unterschiedlichen Vor- und Nachteilen. In diesem Abschnitt werden diese jeweils gegenübergestellt.

In der Praxis kann der Einsatz von Filterlisten an verschiedenen Stellen des Datenübertragungsweges vom Inhalteanbieter zum Nutzer, also den Kindern und Jugendlichen, erfolgen. Dabei kommen vor allem vier Möglichkeiten in Frage, die hier in zunehmender Entfernung zum Endnutzer aufgelistet und erläutert werden:

- Filterung im Endgerät: In diesem Fall läuft auf dem Endgerät des Nutzers, das heißt dem Computer, Smartphone oder sonstigen internetfähigen Gerät, eine Software, die den Internetverkehr filtert. In diese Kategorie fallen die derzeit gängigen Jugendschutzprogramme.
- Filterung im Netzwerkgerät: Der Internetzugang im Haushalt wird heutzutage für gewöhnlich über Router geregelt, die entweder eigenständig oder als Komponente des Modems den Internetverkehr im Haushalt verteilen. Auf diesen Geräten sind meist bereits Module zur Filterung des Datenverkehrs vorhanden, die auch die Filterlisten für den Jugendschutz nutzen könnten.
- Filterung über einen Proxy-Server: Es ist möglich, den gesamten Datenverkehr zu einem Endgerät über einen festen Server umzuleiten. Auf diesem Server kann dann auch die Filterung erfolgen.
- Filterung beim Netzanbieter: Da der gesamte Datenverkehr zum Endnutzer über den Netzanbieter erfolgt, ist es auch möglich, eine Filterung schon beim Netzanbieter selbst durchzuführen.

Für eine anschauliche Verdeutlichung dieser Punkte sei folgender Vergleich herangezogen: Um in einem Haushalt das gelieferte Trinkwasser zu filtern, könnte analog zu den oben genannten Möglichkeiten das Wasser entweder an jedem einzelnen Hahn, am Wasseranschluss der Wohnung, an einem Trinkwasserverteiler oder direkt beim Wasserwerk gefiltert werden. Im Folgenden werden die verschiedenen Einsatzorte einzeln diskutiert. Das Bild der Wasserfilterung kann dabei durchaus hilfreich bei der Einschätzung von Vor- und Nachteilen dieser Einsatzorte von Filtern sein.

Betrieb in Endgeräten: Die Filterung im Endgerät erfolgt durch die Installation einer Filtersoftware, zum Beispiel eines Jugendschutzprogramms auf jedem Endgerät, das Kindern oder Jugendlichen zugänglich ist und über eine Internetverbindung verfügt. Dazu zählen Computer, Laptops, Tablets, Smartphones, Fernseher, Spielkonsolen und zunehmend weitere Geräte.

Solche Software kann über gängige Mechanismen (wie etwa ein zugehöriges Installationsprogramm oder ein Internet-Verkaufsportal wie der Google Play Store[7]) installiert und konfiguriert werden. Dabei ist es leicht, auf einem Gerät verschiedene Nutzerprofile zu unterstützen, das heißt für verschiedene Mitglieder eines Haushalts unterschiedliche beziehungsweise keine Filterung vorzunehmen. Darüber hinaus ist es in diesem Fall leicht möglich, die Filter zu personalisieren, das heißt, gezielt vom Filter gesperrte Inhalte freizugeben oder zusätzliche Inhalte zu sperren. Über solche Konfigurationsmöglichkeiten hat der Nutzer die volle Kontrolle über die Filterung. Für Software in Endgerä-

[7] siehe https://play.google.com/store, Stand 1.12.2012

ten sind zudem eine Reihe sinnvoller Erweiterungen denkbar, die zum Beispiel die Ausführung bestimmter Programme ganz oder zeitlich beschränken. Für Geräte mit einem Windows-Betriebssystem sind zudem bereits zwei von der KJM anerkannte Jugendschutzprogramme verfügbar (siehe Abschnitt 1.4.1).

Problematisch bei dieser Lösung ist allerdings die Vielfalt der Betriebssysteme. Um eine große Abdeckung über die potentiellen Endgeräte zu erreichen, müssten Lösungen für eine große Zahl von Betriebssystemen entwickelt, getestet und auf ihre Sicherheit hin überprüft werden. Andererseits werden zu den einzelnen Betriebssystemen meist eine Reihe verschiedener Versionen betrieben, von denen zumindest ein relevanter Anteil stets abgedeckt werden muss. Bei der zunehmenden Anzahl und Vielfalt an internetfähigen Endgeräten gilt es zudem, immer komplexere Spezialsysteme abzudecken. Als besonders relevante Beispiele seien hierzu Spielkonsolen und Fernseher genannt, die inzwischen standardmäßig mit Internetzugang ausgestattet sind. Schließlich ist je nach Filtertechnik auch die Kooperation der Betriebssystem-Hersteller notwendig, die über die Freigabe von Software und Apps gegebenenfalls mit entscheiden (zum Beispiel Apple).

Betrachtet man über die bisher gängige Praxis zur Filterung hinaus technische Verfahren zum Jugendmedienschutz, wie sie in dieser Studie behandelt werden, so ist zudem mit Effizienz- und Bandbreitenproblemen beim Einsatz auf Endgeräten zu rechnen (siehe auch Abschnitt 6.1).

Betrieb in Netzwerkgeräten: Eine Filterung in den lokalen Netzwerkgeräten eines Haushalts durchzuführen, zum Beispiel im Router, ist als Lösung ebenfalls bereits verfügbar (siehe Abschnitt 1.4). Prinzipiell ist es damit möglich, an einer zentralen Stelle alle Endgeräte abzudecken, die über das entsprechende Netzwerkgerät versorgt werden. So kann mit einer gemeinsamen Lösung zum Beispiel ein ganzer Haushalt oder eine ganze Schule versorgt werden. Dabei kann eine Unabhängigkeit vom Betriebssystem und den Besonderheiten der Endgeräte erreicht werden. Die Kontrolle über die Filterung wird dabei allerdings ebenfalls zentralisiert. Einerseits wird es so leichter, unerwünschte Zugriffe auf die Konfiguration des Filters zu vermeiden, andererseits wird es aber auch schwieriger, unterschiedliche Nutzergruppen zu verwalten, zum Beispiel einzelne Klassen einer Schule.

Der Einsatz auf Netzwerkgeräten wird dadurch erschwert, dass es üblicherweise nicht möglich ist, auf diesen Geräten zusätzliche Software zu installieren. Dies bedeutet, dass die entsprechenden Lösungen von den Herstellern der Netzwerkgeräte vorinstalliert und mit ausgeliefert werden müssten. Dabei ist insbesondere die Wartung deutlich schwieriger als in anderen Szenarien. Das Einrichten individueller Konfigurationen für unterschiedliche Nutzer ist ebenfalls schwerer umzusetzen. Mobile Endgeräte werden in dieser Lösung zudem nur dann mit einbezogen, wenn sie zum Beispiel über WLAN auch tatsächlich ein Netzwerkgerät für den Internetzugang nutzen.

Darüber hinaus ist in einem Netzwerkgerät – und ebenso in Systemen, die sich noch weiter vom Endgerät des Nutzers entfernt befinden (siehe die nächsten beiden Abschnitte) – sicherzustellen, dass die Nutzerautonomie gewahrt bleibt insofern, dass der Nutzer selbst entscheiden kann, ob ein bestimmter Inhalt freigegeben oder zusätzlich gesperrt werden soll, und dass auch für unterschiedliche Nutzer beziehungsweise Altersstufen – Eltern gegenüber Kinder und Jugendlichen – unterschiedliche Nutzungsszenarien realisiert werden können.

Zu beachten ist auch, dass der Einsatz von Software, welche die Web-Inhalte unabhängig von der Filterliste direkt, also anhand ihres Inhalts bewertet, in diesem Szenario nicht ohne weiterführende Kooperation mit den Herstellern der Netzwerkgeräte umgesetzt werden kann.

Betrieb über einen Proxyserver: Das Umleiten aller Netzwerkverbindungen eines Haushalts über einen Zwischenserver (Proxyserver), auf dem dann die Filterung erfolgt, wird ebenfalls bereits als Lösung angeboten[8].

Bei diesem Vorgehen kann nicht nur eine Umleitung von Verbindungen, die über einen lokalen Netzwerkzugang erfolgen, sondern auch von Mobilfunkverbindungen realisiert werden. Über unterschiedliche Zwischenserver für verschiedene Endgeräte und Nutzer lässt sich zudem die nutzerspezifische Konfiguration ermöglichen. Prinzipiell kann diese Lösung auch mit Einsatzmöglichkeiten von Netzwerkgeräten kombiniert werden, die dann die Umleitung auf den jeweils zuständigen Zwischenserver vornehmen. Auf einem Zwischenserver lassen sich sehr komplexe Mechanismen zur Verwaltung der Filterlisten realisieren. Sie können etwa durch den Einsatz von menschlichen Experten, durch Crowdsourcing oder durch technische Bewertungsverfahren erstellt und aktualisiert werden. Auch eine direkte Analyse von Webseiten, die nicht in den Filterlisten vorkommen, ist hier bei entsprechender Infrastruktur möglich.

Allerdings ist in diesem Szenario auch eine entsprechend aufwändige Infrastruktur beim Anbieter des Proxy-Services notwendig. Diese muss nicht nur die zuvor erwähnten Funktionalitäten erfüllen, sondern auch rund um die Uhr verfügbar sein und damit über einen entsprechenden Support verfügen. Da der Betrieb in diesem Fall nicht mehr in der Hand des Endnutzers liegt, sind die Kontrollmöglichkeiten auch weniger stark ausgeprägt. Für manche Endgeräte (zum Beispiel Spielkonsolen) ist darüber hinaus zu beachten, dass Umleitungen auf Zwischenserver dort leicht zu deaktivieren sind. In diesem Fall wäre also Unterstützung durch die Gerätehersteller notwendig.

Betrieb durch den Netzanbieter: Es besteht auch die Möglichkeit zur Filterung von Web-Inhalten direkt bei den Netzanbietern. Da jede Auslieferung von Daten über den Anbieter erfolgt, kann an dieser Stelle auch entschieden werden, bestimmte Inhalte nicht auszuliefern. Diese Filterung ist im Prinzip nur durch Wechsel des Anbieters beziehungsweise der Konfiguration beim Anbieter zu beeinflussen und deshalb für Kinder und Jugendliche kaum zu umgehen.

Auch in diesem Fall muss die Filterung nicht auf die Besonderheiten bestimmter Endgeräte oder Betriebssysteme eingehen und die Aktualität der Listen kann zentral sichergestellt werden. Netzanbieter haben zudem Erfahrung im Umgang mit komplexer IT-Infrastruktur. Allerdings müssten die dazu passenden Software-Strukturen, die die Erstellung, Aktualisierung und Konfiguration von Filtern ermöglichen, erst geschaffen werden. Für eine entsprechende Investition sind für die Netzanbieter zunächst entsprechende Anreize nötig.

Neben der rein technischen Betrachtung in diesem Abschnitt ist zu beachten, dass die Einrichtung einer solchen Filter-Infrastruktur bei den Netzanbietern durchaus kontrovers diskutiert wird [Halfacree, 2012, Arthur, 2012].

Zusammenfassend lässt sich sagen, dass die Anwendungsumgebung von Filterlisten entscheidenden Einfluss auf die Bedeutung und die tatsächlichen praktischen Auswirkungen des technischen Jugendschutzes hat.

6.3 Crowdsourcing im technischen Jugendmedienschutz

Im Oktober 2012 gab es im World Wide Web circa 188 Millionen aktive Webangebote [Netcraft Ltd., 2012]. Allein seit Oktober 2011 wuchs das World Wide Web im Saldo um etwa 19 Millionen aktive Angebote, also etwa 50 000 pro Tag. Darüber hinaus werden, insbesondere im

[8] Zum Beispiel durch OpenDNS, siehe http://www.opendns.com/home-solutions/parental-controls, Stand 30.11.2012.

Bereich User-Generated Content, auch auf einzelnen Angeboten sehr viele Inhalte produziert. Zum Beispiel wird auf YouTube pro Sekunde eine Stunde neues Videomaterial hochgeladen.[9] Es lässt sich somit zweifelsfrei feststellen, dass eine umfassende manuelle Bewertung, wenn überhaupt, nur mit sehr hohem Personalaufwand möglich wäre. Crowdsourcing, speziell die Umsetzung einer auf freiwilliger Basis nutzbaren Microtasking-Umgebung, könnte hier potentiell Abhilfe schaffen.

Crowdsourcing bietet grundsätzlich die Möglichkeit, die unerschöpfliche und extrem heterogene Menge an Webauftritten im Netz in Hinblick auf eine etwaige Entwicklungsbeeinträchtigung zu bewerten – insbesondere, wenn die Zahl von partizipierenden Eltern und Experten sehr groß ist. Zum Beispiel kann es Ziel einer solchen Maßnahme sein, zu ermitteln, wie viele Nutzer ein bestimmtes Webangebot als entwicklungsbeeinträchtigend oder als unbedenklich bewerten und welche Altersseinstufung sie ansetzen würden. Unter der Voraussetzung einer geeignet großen und repräsentativen Menge an Bewertungen könnte dann aus den ermittelten Daten eine Gesamtbewertung für den Webauftritt oder für eine einzelne Webseite abgeleitet und als Indizienlieferant in die weitere Bewertung einbezogen werden.

Der folgende Abschnitt befasst sich mit dem Potenzial, den technischen und konzeptionellen Rahmenbedingungen und den Risiken derartiger Beteiligungsmechanismen und mit der Möglichkeit, diese über das Mittel des Crowdsourcings geeignet umzusetzen.

6.3.1 Definition von Crowdsourcing

Der Begriff Crowdsourcing geht auf den amerikanischen Autor Jeff Howe und seinen 2006 im Wired Magazin erschienenen Artikel »The Rise of Crowdsourcing« zurück [Howe, 2006]. Crowdsourcing ist ein Neologismus aus dem Wort *Crowd* (zu deutsch »Menschenmenge«) und dem Ökonomie-Begriff *Outsourcing* (zu deutsch »Auslagern«). Ziel des Crowdsourcing ist es, eine zeit- oder kostenintensive Aufgabe in kleine Teilaufgaben, sogenannte *Microtask*s, zu zerlegen und diese an eine Menge von Internetnutzern zu verteilen. Die Nutzer werden dabei häufig als Crowdworker bezeichnet [Pelzer et al., 2012]. Somit folgt das Crowdsourcing dem Prinzip der Arbeitsteilung. Dabei gibt es aus Sicht des Nutzers sowohl Crowdsourcing-Beispiele auf freiwilliger Basis als auch entgeltbasierte Ansätze. Ein Beispiel für freiwilliges Crowdsourcing ist das *Web of Trust* Projekt[10], bei dem Nutzer Seiten im World Wide Web freiwillig hinsichtlich Vertrauenswürdigkeit, Händlerzuverlässigkeit, Daten- und Jugendschutz auf einer fünfstufigen Skala bewerten können. Der Nutzer erhält im Gegenzug eine Browser-Erweiterung, die ihn vor dem Aufruf von nicht vertrauenswürdigen Webangeboten schützen soll. Ein Beispiel für entgeltbasiertes Crowdsourcing ist »Mechanical Turk« von Amazon.[11] Hierbei handelt es sich um einen Marktplatz für »Human Intelligence Tasks« (zu deutsch »Aufgaben der menschlichen Intelligenz«), in dem Nutzer Aufgaben wie zum Beispiel die Kategorisierung eines Amazon-Produktes gegen einen Kleinstbetrag abarbeiten können. Amazon stellt hierbei die technische Infrastruktur bereit, die prinzipiell von jedem beliebigen Anbieter genutzt werden kann, um eigene Aufgaben zu veröffentlichen. Außerdem bedienen sich auch Unternehmen, unter anderem bei der Produktentwicklung, der Vorteile des Crowdsourcing. Beispielsweise hat die Fiat-Group zur Produktentwicklung des aktuellen Fiat 500 Modells die Internet-Community aufgerufen, vor der Produktion des Fahrzeugs Produktideen einzureichen. Das Projekt verzeichnete etwa 10 Millionen Seitenaufrufe und über 100 000 Designs wurden eingereicht [Fiat-Group, 2006]. Weitere Spielarten des Crowdsourcings sind *Crowdfunding*, also eine Finanzierungsform, bei der über die Crowd Geld eingesammelt und die Finanziers auf unterschiedliche Arten in das Projekt eingebunden werden [Pelzer et al., 2012], sowie *Creative Content Marktplätze* wie zum Beispiel freelan-

[9] siehe http://www.onehourpersecond.com/, Stand: 14.11.2012
[10] siehe http://www.mywot.com/, Stand 15.11.2012
[11] siehe https://www.mturk.com, Stand 15.11.2012

cer.com,[12] die insbesondere in kreativen Berufen Freiberufler und Auftraggeber zusammenbringen sollen [Pelzer et al., 2012]. Neben »Mechanical Turk« gibt es eine ganze Reihe weiterer spezialisierter Crowdsourcing-Plattformen zur Konfiguration eigener Crowd-basierter Projekte, darunter Mob4Hire für die Optimierung von Mobile-Angeboten, uTest für Crowd-basierte Software-Tests oder oDesk als Online-Arbeitsplatzanbieter [Doan et al., 2011].

Die vorhandenen Kennzahlen deuten darauf hin, dass Crowdsourcing eine Wachstumsbranche ist. Einer aktuellen Studie ist dazu zu entnehmen: »Kumulierte Zahlen zu Crowdsourcing Umsätzen gibt es bislang nicht. Auf oDesk, einer der größten Plattformen, wurden nach Angaben der Plattformbetreiber im November 2011 $23 423 972 umgesetzt (in 2 269 026 Arbeitsstunden). Der Crowd Census Report 2011 (siehe [Bratvold, 2011]) weist einige Zahlen der sechs großen globalen Microtask Anbieter (exklusive MTurk) aus, wonach in 2010 über 64 Millionen Tasks ausgeführt wurden, die vergleichbare Zahl für 2011 wird auf 292 Millionen geschätzt (was eine Steigerung um 355 % ausmachen würde). Im gleichen Report geht man (im dritten Quartal 2011) von einem durchschnittlichen Verdienst von 0,0179 Dollar pro Task aus. Dies würde bedeuten, dass auf den sechs Plattformen (darunter u.a. der deutsche Anbieter Clickworker) in 2011 etwa 5,2 Millionen an die Crowdworker ausgeschüttet wurden.« [Pelzer et al., 2012]

Crowdsourcing ist eng mit dem Begriff des Web 2.0 verbunden und teilweise werden auch Projekte wie die Wikipedia als Crowdsourcing-Beispiele aufgeführt – genauer als Collective Knowledge Crowdsourcing [Pelzer et al., 2012]. Die Trennung von Web 2.0, also der Idee des kollaborativen World Wide Webs, wo Nutzer gleichzeitig Konsument und Produzent von Inhalten sein können, und Crowdsourcing ist daher etwas unscharf. In den folgenden Abschnitten wird allerdings nur ein begrenzter Teil dieser Begriffswelt betrachtet, nämlich Crowdsourcing im Sinne des freiwilligen Abarbeitens von Microtasks.

6.3.2 Anwendungsszenarien

Im Folgenden werden verschiedene Anwendungsszenarien für den Einsatz von Crowdsourcing im Kontext des Jugendmedienschutzes beschrieben. Sie unterscheiden sich insbesondere in der Frage, welche Inhalte von den Crowdworkern manuell klassifiziert werden. Dabei würden Eltern und Experten, im Folgenden unter Crowdworker subsummiert, ein webbasiertes System nutzen, um qualitative Bewertungen für Webseiten abzugeben. Verschiedene Konfigurationen sind denkbar.

Variante 1: Bewertung von selbstgewählten Webangeboten oder Webseiten: In dieser Konfiguration, skizziert in Abbildung 6.3, würden die Crowdworker frei gewählte Webseiten oder ganze Webangebote bewerten. Die Art des Bewertungswerkzeugs ist dabei nicht festgelegt. In der Umsetzung könnte dies zum Beispiel über ein Webbrowser-Plug-In ermöglicht werden, das erkennt, auf welcher Webseite sich der Crowdworker gegenwärtig befindet. Das Browser-Plug-In könnte dann die neue Bewertung für die Webseite oder das Webangebot an eine serverseitige Crowdsourcing-Plattform übermitteln und diese fügt die neue Bewertung in eine Datenbank ein. Aus der Datenbank heraus können diese Datensätze dann für die weitere Bewertung verwendet werden. Dieses Vorgehen entspricht demjenigen im *Web of Trust* Projekt.[13] Dabei wäre auch eine bidirektionale Variante denkbar, in der das Webbrowser-Plug-In nach dem Aufrufen der Seite die aktuelle Gesamtbewertung des Angebots bei der Crowdsourcing-Plattform abfragt und im Webbrowser darstellt – zum Beispiel, um dem Crowdworker eine Orientierung zu geben, wie andere Nutzer die Seite bisher bewertet haben oder wie viele Bewertungen existieren. Der Vorteil bei dieser Konfiguration ist, dass die Bewertung mit wenigen Klicks erfolgen kann – im Gegensatz zu einem Webangebot zur Bewer-

[12] siehe http://www.freelancer.com/, Stand 15.11.2012
[13] siehe http://www.mywot.com/, Stand 15.11.2012

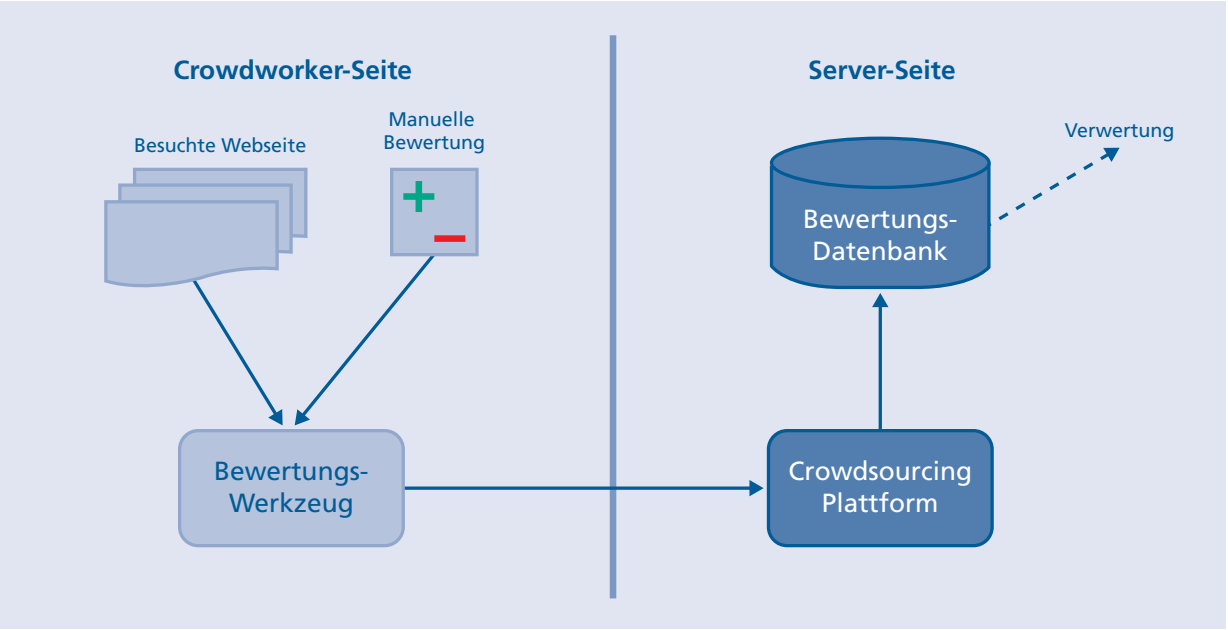


Abbildung 6.3
Crowdsourcing-Konfiguration: Bewertung von selbstgewählten Webangeboten oder Webseiten.

tung entfällt die Auswahl der zu bewertenden Seite – und der Crowdworker diese weitestgehend neben dem Surfen durchführen kann. In Kombination mit einem Jugendschutzprogramm könnten Eltern zum Beispiel auch zuerst diejenigen Seiten bewerten, die ihre Kinder aufgerufen haben. Die Protokollierung der angesteuerten Seiten kann durch das Jugendschutzprogramm ohne Interaktion mit einer Serverseite stattfinden. Eine mögliche Auswirkung bei diesem Vorgehen ist, dass sich die Bewertung eher auf beliebte, oft besuchte Seiten konzentrieren könnte und unbekanntere Angebote nur wenige oder keine Bewertungen erhalten. Demzufolge bietet sich diese Variante am ehesten dafür an, um sichere Surfräume zum Aufbau oder zur Wartung von Whitelists zu bestimmen. Das Bewertungswerkzeug könnte statt einem Webbrowser-Plug-In auch als Webseite bereitgestellt werden, auf der die Crowdworker eine URL zu der zu bewertenden Seite sowie die von ihnen getroffene Altersklassifizierung übermitteln – ähnlich dem Angebot *altersklassifizierung.de* der FSM.

Variante 2: Bewertung von nicht klassifizierten Webangeboten oder Webseiten: Eine andere mögliche Konfiguration ist in Abbildung 6.4 dargestellt. Hier treten die beiden Seiten stärker in Interaktion. Der Crowdworker würde in diesem Fall nicht mehr selbst die Webangebote auswählen die er bewerten möchte, sondern bekommt über das Bewertungs-Werkzeug eine Auswahl bisher nicht bewerteter Seiten angeboten. Die Auswahl wird Server-seitig auf Basis der existierenden Bewertungen und der Menge der bisher nicht bewerteten Webseiten getroffen. Die Seitenauswahl kann dabei zum Beispiel auf Basis des Alexa-Rankings[14] oder vergleichbarer Erhebungen zur Relevanz von Webangeboten getroffen werden. In diesem Fall hieße das konkret, die am häufigsten besuchten aber bisher nicht bewerteten Webangebote durch Crowdworker bewerten zu lassen. Ein prinzipieller Vorteil bei diesem Vorgehen ist, dass der Bewertungsprozess gesteuert und somit eine bessere Abdeckung über die Gesamtauswahl erzielt werden kann.

Variante 3: Bewertung von unsicher automatisch klassifizierten Webangeboten oder Webseiten: Eine dritte Variante ist in Abbildung 6.5 skizziert. Hier würde zusätzlich noch ein Serverseitiges System vor die Crowdsourcing-Plattform geschaltet werden, das verschiedene technische Verfahren für die automatische Bewertung von Webangeboten integriert – beispielsweise ein Verfahren, das Webseiten auf Basis von Textanalyse vorklassifiziert (vgl. Kapitel 4). Die Bewertungsdaten-

[14] siehe http://www.alexa.com/, Stand 15.11.2012

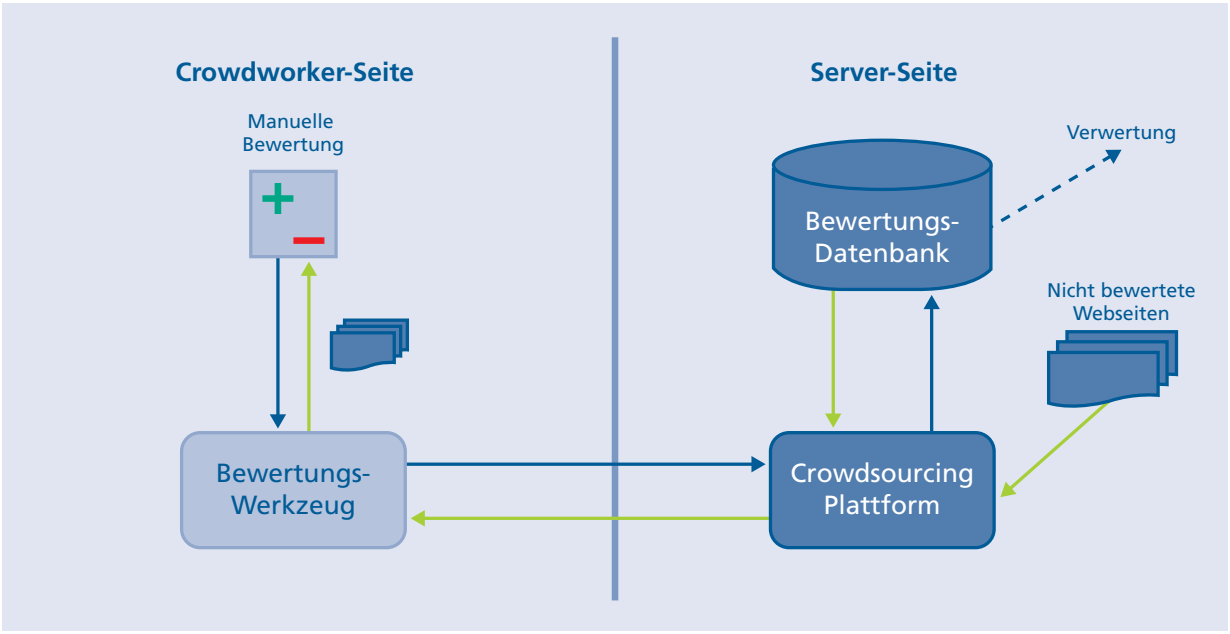


Abbildung 6.4
Crowdsourcing-Konfiguration: Bewertung von nicht klassifizierten Webangeboten oder Webseiten.

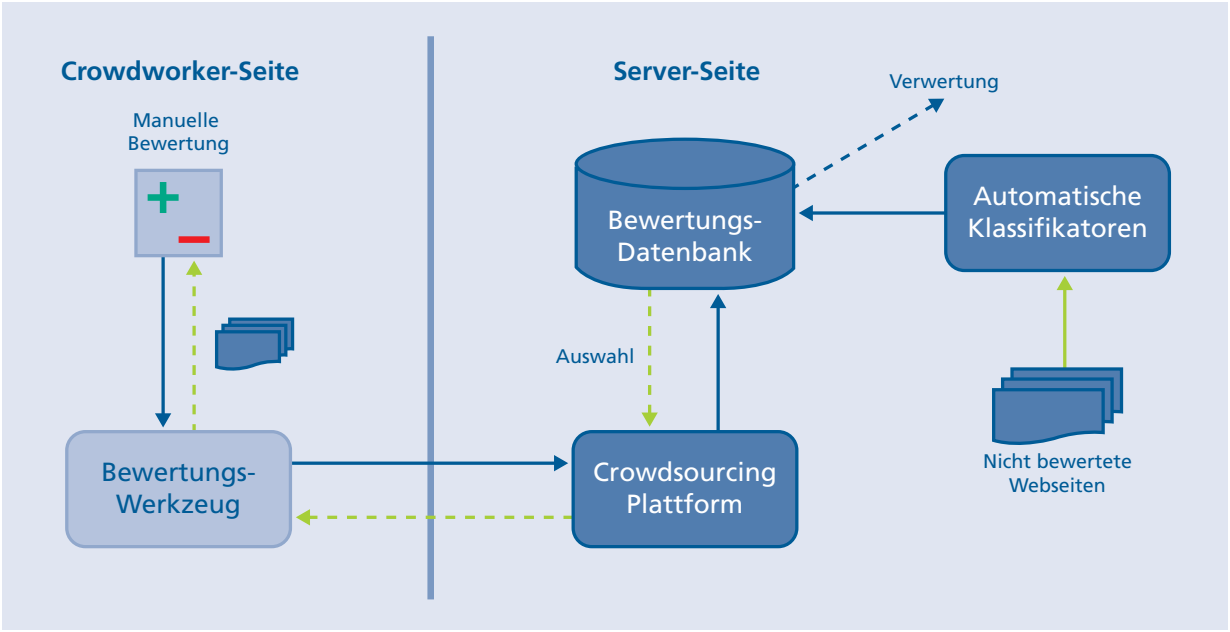


Abbildung 6.5
Crowdsourcing-Konfiguration: Bewertung von unsicher automatisch klassifizierten Webangeboten oder Webseiten.

bank würde somit nicht nur die manuellen Bewertungen, sondern auch die Klassifikationsergebnisse der verschiedenen technischen Verfahren sowie die zugehörigen Konfidenzen beinhalten. In einer solchen Datenbank können Webangebote in mehrere Gruppen unterteilt werden:

1. Bisher nicht bewertete Webangebote
2. Automatisch klassifizierte Webangebote, ggf. mit zusätzlicher Unterscheidung:
   (a) oberhalb einer Entscheidungsgrenze für eine »sichere Klassifikation«
   (b) unterhalb einer Entscheidungsgrenze für eine »sichere Klassifikation«
3. Manuell klassifizierte Webangebote, ggf. mit zusätzlicher Unterscheidung:
   (a) Manuell durch Experten klassifizierte Webangebote
   (b) Manuell durch Crowdworker klassifizierte Webangebote

Die Gruppen 2.a) und 2.b) unterscheiden Klassifikationen daran, ob sie unterhalb oder oberhalb einer sicheren Entscheidungsgrenze liegen. Beispielsweise könnte dies für ein Textklassifikationsverfahren bedeuten, dass dessen Entscheidung oberhalb einer Konfidenz von 80 Prozent als sichere Entscheidung angenommen wird – das heißt die Wahrscheinlichkeit einer Fehlklassifikation vernachlässigbar klein ist. Die Zahlenwerte sind hier frei gewählt und hängen stark vom eingesetzten Verfahren und der experimentell ermittelten Entscheidungsgrenze ab (vgl. Kapitel 5). Die Unterscheidung der beiden Gruppen 3.a) und 3.b) ist notwendig, um die Klassifikation eines Crowdworker, der möglicherweise ohne Vorkenntnisse Bewertungen vornimmt, von einem Experten zu unterscheiden und zu gewichten.

Die Grundlage für die Vorlage eines Webangebots an einen Crowdworker könnte somit aus einer gewichteten Kombination der oben genannten Gruppen erfolgen. Verschiedene Strategien sind denkbar. Sollen möglichst viele Bewertungen für Webangebote erhoben werden, unabhängig davon, ob sie automatisch oder manuell klassifiziert worden sind, so würde das vorgelegte Webangebot ausschließlich aus Gruppe 1. ausgewählt werden. Es könnten auch nur solche Angebote vorgelegt werden, die von einem Klassifikator mit geringer Konfidenz klassifiziert werden konnten (Gruppe 2.a)). Dies würde bedeuten, es werden vorzugsweise unsichere automatische Entscheidungen verifiziert und somit bei gleichem manuellen Aufwand eine größere Zahl sicherer Klassifikationen erreicht. Auch mehrere Iterationsstufen lassen sich abbilden. Beispielsweise könnte ein Webangebot von mehreren Crowdworkern als entwicklungsbeeinträchtigend markiert werden, ohne dass bisher eine Experten-Klassifikation vorliegt. In einer weiteren Iterationen würde ein Experte die Webangebote mit der höchsten Anzahl an Negativbewertungen zuerst klassifizieren.

6.3.3 Risiken

Crowdsourcing birgt neben den genannten Möglichkeiten auch eine Reihe von Risiken und Problemen, die beim Einsatz zu berücksichtigen sind.

Fehlende Supervision für Crowdworker: Ein erstes Problem besteht bei der Bewertung von entwicklungsbeeinträchtigenden Webseiten darin, dass Nicht-Experten, also etwa Eltern oder andere Freiwillige, zur Bewertung eines Angebots die zugehörigen Webseiten sichten müssen. In den bestehenden Expertenteams (z. B. bei jugendschutz.net) existieren für diese manuelle Klassifikation sehr strenge Rahmenbedingungen. Dazu zählen unter anderen die Begrenzung der Zeit, in der ein Experte pro Tag Web-Inhalte bewerten darf, oder die Möglichkeit der Supervision durch einen externen Berater. Hinzu kommt, dass Experten über fundierte Erfahrung in diesem Themenfeld und durch die Arbeit auch über eine Professionalität im Selbstschutz verfügen, die für die Crowdworker keinesfalls vorausgesetzt werden kann. Ferner erfüllen die Arbeitgeber in den entsprechenden Einrichtungen eine Fürsorgepflicht gegenüber den Angestellten und können bei etwaigen Problemen unmittelbar reagieren. Solche Rahmenbedingungen können über ein Crowd-basiertes System nicht

direkt geschaffen werden. Daher muss genau überprüft werden, welche Inhalte in einem Crowdsourcing-System für die Bewertung vorgelegt werden – insbesondere, wenn es sich um strafrechtlich relevante Inhalte handelt. Etwaige Möglichkeiten wären entweder nur auf Variante 1 zu setzen, das heißt, dass freiwillige Crowdworker die Webangebote für die Bewertung selbst auswählen, oder aber von einem offenen Crowdsourcing-Ansatz abzusehen und die Varianten 2 und 3 nur in einem kontrollierbaren Netzwerk von möglicherweise örtlich voneinander getrennten Experten einzusetzen. Gerade bei der letzten Variante besteht somit immer noch ein Potenzial, mehr Webangebote bei gleichem Personalaufwand zu bewerten – ein Teil der Bewertung wird somit von automatischen Verfahren übernommen.

Spannungsfeld zwischen klassischer Bewertungspraxis und Crowd-Bewertungen: Ein wesentlicher Faktor für die Akzeptanz eines Crowdsourcing-Systems ist eine möglichst niedrige Einstiegshürde für die potentiellen Crowdworker. Daher ist es notwendig, das Bewertungssystem so einfach wie möglich zu gestalten. Der Bewertungsprozess von Expertenteams baut dagegen auf einer umfangreichen Zahl von verschiedenen Kriterien auf (vgl. Abschnitt 3.1), deren direkte Abbildung in ein wie im vorherigen Abschnitt skizziertes Bewertungswerkzeug zu komplex sein und damit der Motivation und der Teilnahmebereitschaft von freiwilligen Crowdworkern entgegenstehen könnte. Damit ist also kritisch zu hinterfragen, ob die gängigen Kriterien für ein Crowdsourcing-Werkzeug geeignet sind und wenn nicht, welche Alternativen geeignet sind und wie diese mit der klassischen Bewertungspraxis in Beziehung zu setzen sind. In [Hopf und Braml, 2010] wird diese Problematik bereits angedeutet:

»Gegenwärtig sind im Rahmen der Bewertung von deutschen Telemedienangeboten – sei es durch den Anbieter oder durch die FSM – erste Versuche erkennbar, pauschale Bewertungsansätze in Form von standardisierten Kategorisierungsmethoden zu Grunde zu legen. So gibt es Tendenzen, bei erotischen textlastigen Angeboten, wie beispielsweise Erotikteletextangeboten, auf eine isolierte Bewertung von einzelnen Worten beziehungsweise Wortgruppen sowie Grafiken in Listenform abzustellen, um eine Einschätzung mittels einer Art Checkliste zu ermöglichen.« Weiterhin wird dazu ausgeführt: »Mögen solch kategoriengestützte Ansätze für den Anbieter bei der Einstufung seines Angebotes sicherlich zur Orientierung hilfreich sein, stellen sie jedoch eine Missachtung der allgemein anerkannten Bewertungsgrundsätze dar. Sowohl nach den Kriterien für die Aufsicht im Rundfunk und in den Telemedien als auch nach den Prüfgrundsätzen der FSM ist bei einer Bewertung eines Telemedienangebotes stets zu beachten, in welchem Gesamtkontext die Inhalte eingebunden sind und welche Intention und Funktion sie verfolgen.« [Hopf und Braml, 2010]

Konkret stellt sich das Spannungsfeld so dar, dass ein sehr einfaches Bewertungssystem[15] aufgrund der damit einhergehenden Abstraktion nur ergänzend eingesetzt werden kann, demgegenüber aber die direkte Umsetzung anerkannter Prüfgrundsätze zu komplex und damit abschreckend wirken kann.

Repräsentativität: Außerdem ist in einem System, das Crowdsourcing-Bewertungen unterstützt, klar zu definieren, ab wann und unter welchen Bedingungen eine Menge von Bewertungen repräsentativ und damit als geeignetes Mittel im Gesamtbewertungsprozess einsetzbar ist. Die möglichen Faktoren sind sehr vielfältig und gerade bei Erhebungen im Internet nicht zweifelsfrei zu überprüfen. Neben einer unzureichenden Menge an abgegebenen Bewertungen können zum Beispiel auch ungeeignete demografische Verteilungen oder Verzerrungen zu Negativ-Meinungen auftreten. Dies liegt in der Selbstselektivität der Teilnehmer begründet. Interessierte Nutzer können selbstständig und je nach Eigeninteresse als Crowdworker auftreten. Allerdings haben auch jüngere Studien gezeigt, dass webbasierte Befragungen bei entsprechender Konfiguration vergleichbare Ergebnisse wie klassische Befragungen liefern, aber nicht als Ersatz für diese dienen können [Gosling et al., 2004].

[15] zum Beispiel eine »Jugendschutz-Ampel« oder »Pro-Contra« Entscheidungen.

Dabei ist auch zu berücksichtigen, dass die in Deutschland etablierte und institutionalisierte Spruchpraxis hinsichtlich der Jugendschutzrelevanz von Inhalten Wertvorstellungen folgt, die als gemeinsame Wertmaßstäbe der Bevölkerung in Deutschland angesehen werden können. Dabei bleibt aber unberücksichtigt, dass verschiedene – zum Beispiel religiöse oder weltanschauliche – Gruppen in Deutschland eigene Wertmaßstäbe besitzen und entsprechende manuelle Klassifikationen entlang dieser Maßstäbe ausrichten.

Manipulation durch Dritte: Mit Crowdturfing wird der Einsatz von Crowdsourcing für die systematische Manipulation von sozialen Netzwerken bezeichnet [Wang et al., 2012]. Ein Beispiel dafür sind Anbieter, die Facebook-Likes oder Twitter-Follower für Geld anbieten und die eigentlichen Like- oder Follow-Vorgänge über Crowdsourcing realisieren, das heißt, Internetnutzern Beträge im Cent-Bereich dafür bieten, die gewünschten Seiten zu *liken* oder diesen *zu folgen*. Im Kontext des Jugendmedienschutzes wäre es durch Crowdturfing möglich, dass Anbieter eigene Crowdworker einsetzen, um falsch klassifizierte Bewertungen für ihre Angebote abzugeben. Aus technischer Sicht können derartige Falschbewertungen ohne ausreichende Schutzmaßnahmen zum Beispiel durch Bots sogar automatisiert werden. Zur Vermeidung von Botmanipulationen lassen sich allerdings gängige Schutzmechanismen wie Registrierungszwang kombiniert mit Captcha-Systemen einsetzen. Freiwillige Crowdworker müssten sich also zunächst einen Nutzer-Account anlegen, der gegen automatische Anmeldevorgänge geschützt ist und würden ihre Bewertungen dann nach Login beim Bewertungs-Werkzeug durchführen. Einen wirksamen Schutz gegen manuelle Manipulationen beziehungsweise Falschbewertungen bietet dieses Vorgehen allerdings nicht. Typischerweise wird versucht, solchen Manipulationen mit Reputationssystemen zu begegnen. Zur Verbesserung der Qualität der Beiträge hat die Wikipedia beispielsweise das WikiTrust-Projekt[16] ins Leben gerufen. Das zugehörige System basiert im Wesentlichen auf automatischen Verfahren zur Erkennung von möglicherweise manipulierten oder vandalisierten Wikipedia-Seiten. Die Basis der Analyse bildet die Autorenreputation, die sich aus dem manuellen Feedback anderer Autoren bestimmt. Im Rahmen der vorgestellten Crowdsourcing-Ansätze ließe sich eine derartige Funktion zum Beispiel über einfache Meldevorgänge für verdächtige Crowdworker realisieren – etwa bei auffällig hoher Zahl von offensichtlichen Falschbewertungen. Im Bereich der Dokumenten-Klassifikation mittels Crowdsourcing wurden in der Literatur ähnliche Systeme ebenfalls untersucht.

[Ipeirotis et al., 2010] diskutieren Qualitätssicherungsmaßnahmen für das Amazon Mechanical Turk System. Die Autoren stellen ein Reputationssystem vor, welches böswillige Annotatoren von solchen trennt, welche konsistent eine abweichende Meinung haben. Als Beispiel nennen sie Eltern von jungen Kindern, die im Durchschnitt wesentlich strengere Maßstäbe für die Freigabe von Filmen anlegen. Dazu erzeugen sie eine Qualitätsmaßzahl für jeden Annotator, der die Rate der böswilligen Annotationen und der persönlichen Tendenz des Annotators reflektiert. Das System wurde mit synthetischen und empirischen Dokumenten getestet. Es kann ohne Modifikationen auf den Fall übertragen werden, dass eine Gruppe von Freiwilligen neue Webangebote klassifiziert, um Indizien für die Gesamtbewertung zu liefern.

[Laws et al., 2011] verwenden das Amazon Mechanical Turk System direkt zum Aktiven Lernen von Dokumentklassifikatoren. Sie stellen in realen Experimenten fest, dass bei vier Kategorien die Antwort in 50–80 % der Fälle korrekt ist. Dabei waren sehr produktive Annotatoren wesentlich unzuverlässiger. Die Autoren schlagen daher vor, Annotationen mehrfach durchzuführen, um die Qualität sicherzustellen. Sie weisen zudem die Annotationen von schlecht bewerteten Annotatoren zurück.

[Pantola et al., 2011] präsentieren ein Reputationssystem, das die Annotationsqualität (Reputation) der einzelnen Annotatoren beurteilt. Ziel ist die Kategorisierung von URLs in einzelne Kategorien. Als Kategorien werden betrachtet: Pornographie, Gewalt etc. Jedes neue Dokument wird von drei Annotatoren klassifiziert. Das System wurde durch Simulationen geprüft, in denen die Anzahl der

[16] siehe http://www.wikitrust.net/, Stand: 15.11.2012

»guten«, »böswilligen« und »unzuverlässigen« Annotatoren variiert und die Systemperformanz gemessen wurde. In seiner ursprünglichen Konzeption wurde das System zur Annotation aller neuen Webseiten verwendet, um Blacklists zu aktualisieren. [Pantola et al., 2011] diskutieren ausführlich diese und andere Techniken, die bei der Filterung jugendschutzrelevanter Seiten in Deutschland genutzt werden könnten.

Einen weiterführenden Überblick über die Möglichkeiten des Schutzes vor Manipulation in Crowdsourcing-System bietet [Doan et al., 2011].

Betriebskosten: Abschließend ist zu erwähnen, dass die notwendige Infrastruktur für ein solches System Aufwände in Konzeption, Implementierung, Betrieb und Wartung verursacht und somit einen permanenten Kostenfaktor darstellt.

6.4 Teilautomatische Qualitätssicherung und Priorisierung im Bewertungsprozess

Über den Verlauf der bisherigen Abschnitte wurden verschiedene Aspekte der Integration technischer Verfahren diskutiert. Neben der Inbetriebnahme der Verfahren, Konzipierung und Implementierung von Filterlisten und möglichen Chancen durch Crowdsourcing gibt es weitere Bereiche im Bewertungsprozess, die für eine effektive Filterleistung ausschlaggebend sind. Dazu zählen automatische und semiautomatische Maßnahmen zur Qualitätssicherung, also die kontinuierliche Prüfung von bereits klassifizierten Webangeboten auf Veränderungen, die gegebenenfalls eine Neubewertung erfordern und dabei eine möglichst vorteilhaft priorisierte Auswahl an manuell zu bewertenden Angeboten.

6.4.1 Teilautomatische Qualitätssicherung

Die Qualitätssicherung von Filterlisten ist ein wichtiges Element für den Jugendmedienschutz. Bestehende Klassifikationen beziehen sich auf Webseiten oder ganze Webangebote, die sich über die Zeit inhaltlich verändern können. Dabei können sie gänzlich auf andere Eigentümer und demnach auf andere Inhalteanbieter übertragen werden (zum Beispiel bei einem Domainverkauf und Wiederankauf). Zudem können sie in unterschiedlichen Kontexten durch ihre Verlinkung auf anderen Seiten auftauchen oder gänzlich verschwinden. Um Filterlisten aktuell zu halten, gibt es unterschiedliche (semi-)automatische Prozesse, die unterstützend zur manuellen Qualitätssicherung eingesetzt werden können.

Testen auf Verfügbarkeit: Eine erste Möglichkeit, um für bestehende Filterlisten eine Qualitätssicherung durchzuführen ist es, einen regelmäßigen Verfügbarkeitstest durchzuführen. Mit einem solchen Test, der in der Regel auch Bestandteil typischer Webcrawler ist, können zweierlei Dinge überprüft werden:

- Generelle Überprüfung eines Webangebots auf Verfügbarkeit
- Überprüfung eines Webangebots mit bekannter age-de.xml Kennzeichnung auf Verfügbarkeit und Validierung der Alterskennzeichnung

Die Verfügbarkeit kann mit einer Reihe von kostenlosen und teilweise quelloffenen Werkzeugen durchgeführt werden. Beispielsweise kann auf Linux-Systemen das Open-Source Werkzeuge *wget* verwendet werden, um die wichtigsten Webcrawler-Aufgaben auszuführen – zum Beispiel Prüfung auf valide Serverantwort mit unterschiedlichen Timeouts, HTTPS-Verbindungen, rekursives *crawlen*

von Webseiten eines Angebots usw. Für *wget* gibt es auch Windows-Portierungen[17] oder die Möglichkeit, das Werkzeug über *Cygwin*[18] zu verwenden. Die Überprüfung von validen age-de.xml Kennzeichnungen kann ebenfalls mit einfachen Kommandozeilen-Werkzeugen durchgeführt werden. Mit dem Werkzeug *xmllint* kann zum Beispiel unter Linux eine age-de.xml Datei auf Korrektheit gegenüber der bestehenden Betaversion des XML-Schemas[19] untersucht werden.

Der Nutzen dieser Werkzeuge hängt davon ab, welche Crawler-Software aktuell bei den Herstellern eingesetzt wird. Wie bereits erwähnt, unterstützen gängige Systeme wie beispielsweise Apache Nutch,[20] Mechanismen, um während des Crawling-Vorgangs temporär oder gänzlich nicht verfügbare Webangebote oder XML-Inhalte wie die age-de.xml Auszeichnung gesondert zu behandeln.

Kollaborative Qualitätssicherung von Whitelists: Zur Qualitätssicherung von sicheren Surfräumen bietet sich neben den genannten Möglichkeiten auch das in Abschnitt 6.3 vorgestellte Crowdsourcing an. Wie dort bereits angedeutet, ist die Qualitätssicherung von »guten« Webangeboten einfacher zu realisieren, da für diesen Anwendungsfall eher eine Alarm- statt einer Klassifizierungsfunktion geschaffen werden müsste. Da sich innerhalb des sicheren Surfraums nur manuell klassifizierte Webangebote befinden, kann das Feedback der Nutzer darauf konzentriert werden, ob diese Angebote ihrer Ansicht nach immer noch korrekt klassifiziert sind. Dies ist eine einfache Ja-Nein-Entscheidung, die zum Beispiel über ein Webbrowser-Plugin mit wenigen Klicks ermöglicht werden könnte. Der klassischen Bewertungspraxis (vgl. Kapitel 3) stünde eine solche Alarmfunktion nicht entgegen, sondern ist eher ergänzend. Zur Qualitätssicherung könnten die Hersteller ab einer kritischen Anzahl von Alarmierungen eine Neubewertung der Seite anstoßen. Eine Aussage über die notwendige Anzahl kann hier nicht getroffen werden. In der Praxis könnte der Schwellwert von der Anzahl der teilnehmenden Nutzer abhängig gemacht oder aber, insbesondere bei wenigen Teilnehmern, jedem Hinweis nachgegangen werden.

Vergleich von Webseiten-Versionen: Für die Erkennung von Duplikaten von Webseiten gibt es in der Forschung und in kommerziellen Anwendungen eine Reihe von automatischen Ansätzen. Die verwendeten Ansätze haben eine ähnliche Basis wie die Fingerprint-Verfahren für Audio und Video (vgl. Abschnitt 4.5.1). Obwohl die Verfahren für die Suche nach duplizierten Web-Inhalten (zum Beispiel für die Suche nach Plagiaten, wie etwa die Ephorus Software[21]) konzipiert sind, können sie auch für die Qualitätssicherung eingesetzt werden. Dabei werden für eine Webseite nicht alle Duplikate aus einer Menge von Webseiten gesucht, sondern es wird bestimmt, ob eine bestimmte Webseite zu einem bestimmten Zeitpunkt identisch ist mit derselben Seite zu einem anderen Zeitpunkt.

[Kumar und Govindarajulu, 2009] geben einen Überblick über Verfahren zur Suche von Teilduplikaten. Eine wichtige Version dieser Verfahren nutzt eine Auswahl von n-Grammen von Wörtern (Shingles) und vergleicht diese für die verschiedenen Dokumente. Alternative Verfahren nutzen Fingerabdrücke (engl.: Fingerprints). Fingerabdrücke von Dokumenten sind eine Sammlung von Zahlen, welche den Inhalt eines Dokuments charakterisieren und durch Hashfunktionen aus einzelnen Wörtern (z. B. Wörtern mit hohem tf-idf-Wert[22]) oder Sätzen berechnet werden. Local Sensitivity Hashing ist auf ein Ähnlichkeitsmaß für Dokumente, die Jaccard-Ähnlichkeit, abgestimmt und skaliert für große Dokumentensammlungen.

[Theobald et al., 2008] beschreiben einen robusten Algorithmus, der auf der Verwendung von sehr häufigen Ankerworten beruht. Oft werden einige Stoppworte (z.B. ist, der, die, das …) als Anker-

[17] siehe http://gnuwin32.sourceforge.net/packages/wget.htm, Stand 1.12.2012
[18] siehe http://www.cygwin.com/, Stand 1.12.2012
[19] siehe http://www.online-management-kontor.de/downloads/age-de-xsd-beta_110815.xsd, Stand 1.12.2012
[20] siehe http://nutch.apache.org/, Stand 1.12.2012
[21] siehe https://www.ephorus.com/, Stand 1.12.2012
[22] Eine Kennzahl zur Beurteilung der Relevanz von Begriffen in Dokumenten einer Dokumentenkollektion – je höher desto besser.

worte ausgesucht. Das Wort-Bigramm (oder Trigramm) hinter diesem Ankerwort wird durch eine Signatur charakterisiert und in einem Index gespeichert. In einer empirischen Untersuchung hatte diese Technik einen F-Wert von 94 % bei der Entdeckung von Teilduplikaten und damit bei hoher Verarbeitungsgeschwindigkeit eine höhere Genauigkeit als andere Verfahren. Die Software ist frei verfügbar.[23] [Das et al., 2012] verwenden eine gewichtete Term-Dokument-Matrix zur Repräsentation der Webseiten. Dabei wird die Struktur der Webseite mit berücksichtigt. In Experimenten auf einem Benchmark-Problem ergibt sich eine Präzision von 94 % bei einer Trefferquote von 94 %. Insgesamt arbeiten die Verfahren mittlerweile zuverlässig und können auf große Dokumentensammlungen angewendet werden.

Diese qualitätssichernde Maßnahme kann sicherlich nicht so unmittelbar eingesetzt werden wie einfache Verfügbarkeitstests, allerdings hängen die eingesetzten Verfahren eng mit den in Abschnitt 4.4 (Textanalyse) und Abschnitt 4.6 (Umfeldanalyse) vorgestellten Verfahren zusammen. Auch diese können prinzipiell so eingesetzt werden, dass sie die extrahierten Merkmale nicht für eine Klassifikation, sondern für einen Vergleich zwischen zwei Versionen desselben Webangebots nutzen. Interessant könnten hier beispielsweise Veränderungen bei den Ergebnissen der Umfeldanalyse, insbesondere wenn Webangebote über die Zeit verstärkt in Kontexten mit bekannten Altersklassifizierungen auftreten, oder starke strukturelle oder inhaltliche Veränderungen sein. Bei interaktiven Angeboten, wie zum Beispiel Blogs mit Kommentarfunktionen, sind auch Differenzbildungen zwischen zwei Webseitenversionen denkbar – die Differenz wären in diesem Fall die neuen Kommentare seit der letzten Prüfung – um neue Kommentare auf Badwords zu untersuchen.

Auch ohne den Einsatz von Verfahren zur Textklassifikation gibt es Indizien für Änderungen auf Webseiten. HTTP-Header können zum Teil unmittelbar dazu verwendet werden, um die Änderung einer Seite und somit die Notwendigkeit einer Neubewertung festzustellen. Die Header enthalten gegebenenfalls eine oder mehrere der folgenden Informationen:

- Last-Modified (Response): Zeitpunkt der letzten Änderung
- Expires (Response): Zeitpunkt, bis zu dem die gelieferte Antwort gültig ist
- If-Modified-Since (Request): lädt Webseite nur, wenn sie sich seit einem bestimmten Zeitpunkt geändert hat
- ETag (Response), If-None-Match (Request), (If-Range (Request)): identifiziert eine Webseitenversion und lädt die Webseite nur, wenn sie sich seit einem bestimmten Zeitpunkt geändert hat
- Cache-Control (Response), Age (Response): Caching erlaubt/nicht erlaubt, Zeit in Proxy Cache

6.4.2 Priorisierung von Bewertungsgegenständen

Der Einsatz von technischen Verfahren und die Einführung teilautomatisierter Prozesse birgt das Potenzial, den gesamten Bewertungsprozess von Webangeboten effektiver zu gestalten. Auch in den Prozessen der Hersteller von Jugendschutzprogrammen werden schon teilautomatisierte oder ausgelagerte Prozesse eingesetzt, beispielsweise bei JusProg e.V.:

- Einsatz von Webcrawlern, um neue Webangebote im World Wide Web aufzuspüren,
- Einsatz von automatisierten Tests, um nicht mehr vorhandene Webangebote zu identifizieren,
- Automatische Textklassifikation zur Bewertung und Vorbereiten der redaktionellen Klassifikation,
- Anbieten von Meldestellen für verdächtige Webangebote (Vorbereiten der redaktionellen Klassifikation durch manuelle Priorisierung).

[23] siehe http://sourceforge.net/projects/spotsigs/, Stand 1.12.2012

Da die Anzahl von Experten für die manuelle Bewertung von Webangeboten begrenzt und nicht ausreichend für eine flächendeckende Bewertung von neuen Webangeboten ist – und weniger noch für den existierenden Bestand – ist eine gesteuerte Auswahl der Bewertungsgegenstände, also den zu bewertenden Webangeboten, entscheidend. Die Wichtigkeit einer geeigneten Auswahl lässt sich an einem Beispiel erläutern. Werden zwei Expertengruppen der selben Stärke je eine aus zwei disjunkten Listen von Webangeboten ähnlicher Größe vorgelegt, so werden die Gruppen die Altersklassifikation in etwa derselben Zeit durchführen können. Das resultierende Schutzniveau dieser Listen in einem Jugendschutzprogramm hängt allerdings von der Zusammensetzung der Angebote ab, konkret von der Wahrscheinlichkeit, dass Kinder und Jugendliche dieses Angebot letztlich überhaupt aufrufen. Beispielsweise könnte die erste Liste eintausend sehr selten besuchte, in gängigen Suchmaschinen eventuell gar nicht auffindbare Webangebote[24] und die zweite Liste eintausend sehr weit verbreitete und häufig besuchte, gut auffindbare Angebote enthalten. Da auch mit dem Einsatz von technischen Verfahren nicht alle Webangebote im World Wide Web bewertet werden können, ist es also notwendig, Strategien zu entwickeln, möglichst effizient genau diejenigen Webangebote zu identifizieren, die sehr stark entwicklungsbeeinträchtigende Inhalte enthalten, und die von den Nutzern mit der höchsten Wahrscheinlichkeit aufgerufen werden. Hierbei können verschiedene Anhaltspunkte für die Priorisierung nicht bewerteter Angebote ausgewertet werden.

Indizien über technische Verfahren: Wie in den vorherigen Kapiteln 4 und 5 ausführlich besprochen, können vollautomatische technische Verfahren zum Sammeln von Indizien herangezogen werden. Die Anhaltspunkte sind vielfältig – von der Wiedererkennung bereits bewerteter Inhalte bis zur Wahrscheinlichkeit, dass einzelne Konzepte oder Indizien vorliegen – und können auf unterschiedliche Weise zur Auswahl der Webangebote herangezogen werden.

Implizites und explizites Nutzerfeedback: Für Nutzerfeedback kann man generell zwei Arten unterscheiden: Zum einen implizites Feedback, also nicht bewusst durch den Nutzer initiiertes Feedback wie zum Beispiel Klickraten zur Bewertung der Relevanz von Webseiten. Zum anderen gibt es explizites Feedback. In diesem Fall muss der Nutzer interagieren, um eine Bewertung abzugeben. Ein klassisches Beispiel ist der Like-Button von Facebook. Wie bereits erwähnt, bietet zum Beispiel JusProg e. V. die Möglichkeit, verdächtige Webangebote für eine manuelle Bewertung vorzuschlagen. Wenn solche Angebote gemeldet werden, werden diese redaktionell klassifiziert. Hierbei handelt es sich um explizites Feedback.

Feedback via Crowdsourcing: Die kollaborative Bewertung von Webangeboten mittels Crowdsourcing bietet eine weitere Möglichkeit, eine Priorisierung für die Gesamtbewertung der möglichen Angebote zu sammeln. Diese Möglichkeit wurde in Abschnitt 6.3 im Detail besprochen.

Auswertung der Popularität: Für die Bewertung der Popularität eines Webangebots können verschiedene Suchmaschinen-Kennzahlen herangezogen werden. Beispiele dafür sind die Link- oder Click-Popularity oder der Alexa-Rang des Angebots.[25] Mit steigender Popularität eines Angebots steigt auch die Wahrscheinlichkeit, dass ein Kind oder Jugendlicher das Angebot aufsucht.

Auswertung von existierenden Filter- oder Bewertungslisten: Es gibt verschiedene Filter- und Bewertungslisten im Netz, die entweder für sich genommen oder über eine geeignete Kombination als Anhaltspunkt für eine Priorisierung verwendet werden können. Beispiele dafür sind die öffentliche Blacklist urlblacklist.com[26], die Bewertungen aus dem Web of Trust[27] oder webutation.net[28]. Denkbar sind auch die Nutzung von organisatorischen Informationen, wie zum Beispiel Informa-

[24] Zum Beispiel können Anbieter über eine robots.txt die Google Crawler anweisen, das eigene Angebot für die Suche zu ignorieren, siehe http://support.google.com/webmasters/bin/answer.py?hl=en&answer=35303, Stand 1.12.2012.

[25] siehe http://www.alexa.com/, Stand 1.12.2012

[26] siehe http://urlblacklist.com, Stand 1.12.2012

[27] siehe http://www.mywot.com/, Stand 1.12.2012, Web of Trust verfügt außerdem über ein öffentliches API.

[28] siehe http://www.webutation.net, Stand 1.12.2012

tionen über den Besitzerwechsel von Domains. Fraglich ist allerdings, ob die Registrierungsstellen wie etwa DENIC eG[29] diese Informationen im Rahmen des technischen Jugendmedienschutzes zur Verfügung stellen würden.

6.5 Zusammenfassung

Die Integration technischer Verfahren in entsprechende IT-Infrastrukturen beziehungsweise in die Jugendschutzprogramme selbst eröffnet eine Reihe von Fragestellungen für die Hersteller solcher Produkte. In den vier Abschnitten wurden jeweils unterschiedliche Konzepte vorgestellt und, je nach Notwendigkeit, deren Vor- und Nachteile gegenübergestellt.

Betriebsumgebungen für technische Verfahren: In Abschnitt 6.1 wurden drei Alternativen für die Einbindung technischer Verfahren in Jugendschutzprogramme vorgestellt: Direkt, über eine ausgelagerte Cloud-Architektur oder gekapselt durch ein Proxysystem. Das Spektrum der offenen Fragestellungen berührt dabei vor allem die erreichbare Echtzeitfähigkeit der Verfahren, die Flexibilität bei der Auswahl von Betriebssystemen und bei der Aktualisierung von Filterlisten sowie die notwendigen Investitionen in IT-Infrastruktur.

Zentrale Lösungen für die Klassifikation ermöglichen gegenüber dem Betrieb auf dem Endgerät eine flexible und skalierbare Betriebsumgebung für die Klassifikationsverfahren und verhindern die redundante, wiederholte Bewertung der immer wieder gleichen Web-Inhalte. Zentrale Filterlisten können unmittelbar aktualisiert werden und neue redaktionelle wie automatische Klassifikationen stehen allen Nutzern zum frühestmöglichen Zeitpunkt zur Verfügung. Für ausgewählte Verfahren der Textanalyse sind darüber hinaus Echtzeit-Klassifikationen möglich. Diese können aber ebenfalls von einem zentralen Dienst (zum Beispiel in einer Cloud-Infrastruktur) durchgeführt werden. Beim Betrieb von Proxy- und Cloudsystemen zur teilautomatisierten Bewertung von Webangeboten entstehen natürlich nicht zu vernachlässigende Kosten für die zu betreibende Infrastruktur. Auch wenn die Kosten für Cloud-Infrastrukturen in der letzten Zeit fallen – insbesondere die großen Anbieter Google und Amazon unterbieten sich zum Zeitpunkt der Erstellung der vorliegenden Studie mit immer günstigeren Cloud-Angeboten [Greif, 2012] – und im Unternehmensbereich auf Akzeptanz stoßen (2012 stuften 56 % von 114 deutschen IT-Anbietern Infrastrukturservices wie das Amazon EC-Angebot als sehr oder eher attraktiv ein [Spath et al., 2012]), so muss der Einsatz doch im Einzelfall geprüft werden. Auch muss die Akzeptanz der Nutzer in Bezug auf den Datenschutz sichergestellt werden, da zwangsläufig eine umfangreiche Kommunikation des eigenen Rechners mit einem Proxy- oder Clouddienst benötigt wird. Typischen Herausforderungen im Cloud Computing (Datenschutz- und Sicherheits-Aspekte, Ausfallsicherheit oder die Auswahl geeigneter Anbieter) steht die Möglichkeit kostengünstiger, flexibel einsetzbarer Hardware entgegen. Weiterführende Literatur zu diesen Themen bieten [Fujitsu Services, 2011] und [Spath et al., 2012].

Konzept und Implementierung von Filterlisten: In Abschnitt 6.2 wurden Filterlisten als ein zentrales Werkzeug im technischen Jugendschutz vorgestellt. Je nach gewünschter Granularität, den eingesetzten Verbindungsprotokollen und der Anwendungsumgebung ergeben sich eine Vielzahl an neuen Möglichkeiten und Herausforderungen für Jugendschutzprogramme.

Am einfachsten gestaltet sich hierbei eine Erweiterung der Filterliste um zusätzliche Informationen zur Verbesserung der Listenqualität. Diese lassen sich abwärtskompatibel gestalten und können so bestehende Filterlisten rasch ergänzen. Durch den Einsatz von automatischen Verfahren zur inhaltsbasierten Klassifikation ist eine feinere Granularität mittels einer Bewertung von einzelnen URLs sinnvoll. Diese soll dabei keineswegs die bestehende Domain-Klassifikation ersetzen, sondern vielmehr

[29] siehe http://www.denic.de/, Stand 1.12.2012

ergänzen. In großen heterogenen Videoportalen kann so die bestehende Altersklassifikation auf Domainbasis – Portal ist »ab 12 Jahre« – um Einzelklassifikationen – dieser spezielle Inhalt ist aber »ab 18 Jahre«, dieser spezielle Inhalt ist aber »ab 6 Jahre« – verfeinert werden.

Vor dem Hintergrund einer stetig steigenden Zahl internetfähiger Geräte und der damit verbundenen Vielzahl an Betriebssystemen und Anwendungsumgebungen für Jugendschutzprogramme ist die Orientierung weg vom Endgerät und hin zur Quelle ein mögliches Element zur Erhöhung der praktischen Relevanz von Jugendschutzfiltern. Dies kann zum Beispiel bedeuten, dass eine Filterliste nicht am Endgerät, sondern in einem Router für einen ganzen Haushalt, mittels eines zentralen Proxy-Servers oder direkt beim Netzanbieter abgeprüft wird. Allerdings benötigt man bei solchen Lösungen entsprechende Konzepte zur Gewährleistung der Nutzerautonomie, wie etwa individuelle Konfiguration von Filterlisten durch Eltern und individuelle Benutzerprofile für einzelne Familienmitglieder.

Crowdsourcing im technischen Jugendmedienschutz: Das in Abschnitt 6.3 vorgestellte Crowdsourcing bietet in Hinblick auf die enorm umfangreiche und praktisch nicht durch einzelne Expertengruppen zu bewältigende Bewertungsarbeit von Web-Inhalten eine Möglichkeit, Arbeiten aufzuteilen und an eine größere Community zu verteilen. Diese Community kann entweder aus freiwilligen Teilnehmern (z.B. Eltern), aber auch aus voneinander unabhängigen Expertengruppen bestehen, die durch dieses technische Mittel ihre Kräfte bündeln können. Crowdsourcing bietet sich hier insbesondere als Indizienlieferant bei der Erstellung von Filterlisten und der Priorisierung von durch Experten zu bewertenden Inhalten, birgt jedoch auch diverse Herausforderungen, wie in Abschnitt 6.3 dargestellt.

Teilautomatische Qualitätssicherung und Priorisierung im Bewertungsprozess: Darüber hinaus gibt es, wie in Abschnitt 6.4 dargelegt, eine Vielzahl teilautomatischer Verfahren, die den Klassifizierungsprozess für Web-Inhalte im Gesamten optimieren können. Am vielversprechendsten sind solche für die teilautomatisierte Qualitätssicherung und zur Priorisierung von manuell zu bewertenden Webangeboten. Exemplarische Untersuchungen konnten im Rahmen dieser Studie nicht durchgeführt werden, daher muss der Beitrag der vorgestellten Verfahren auf die Effektivität der Filterleistung noch überprüft werden.

KAPITEL 7

HANDLUNGSEMPFEHLUNGEN

Im Rahmen der vorliegenden Studie wurden technische Verfahren untersucht, die zur automatischen Erstellung und Pflege von Filterlisten im Jugendmedienschutz eingesetzt werden können. Es wurde gezeigt, dass Kinder und Jugendliche durch den Einsatz ausgewählter Verfahren vor bestimmten Inhalten besser geschützt werden können. Dies gilt insbesondere für die Bereiche Gewalt, Extremismus und Selbstgefährdung.

Aus der detaillierten Betrachtung der Verfahren ergeben sich eine Reihe von unmittelbaren Handlungsempfehlungen, die im Folgenden dargestellt werden.

7.1 Kurzfristig einsetzbare Verfahren

Auf Basis der vorliegenden Studie ergeben sich folgende Empfehlungen, die kurzfristig zu einer Verbesserung des technischen Jugendmedienschutzes führen können. Die zugehörigen Techniken und Software-Module sind im industriellen Einsatz erprobt und die verfügbaren Lösungen müssen nicht an die Besonderheiten im Jugendmedienschutz angepasst werden. Durch die Einführung der genannten Technologien kann die Leistungsfähigkeit der bestehenden Schutzsysteme erhöht werden.

Fingerprinting zur Wiedererkennung bereits klassifizierter Inhalte: Bekannte jugendschutzrelevante Inhalte können in allen Modalitäten (Text, Audio, Video, Bild) wiedererkannt werden. Dies ermöglicht eine vollautomatische Filterung solcher Inhalte ohne erneute redaktionelle Prüfung. Dies ist insbesondere für besonders drastische, reale Gewaltdarstellungen relevant: so gibt es zum Beispiel nur sehr wenige Videoaufzeichnungen von realen Morden, diese werden aber immer wieder im Netz verfügbar gemacht und könnten über ein Fingerprinting-System effektiv detektiert und gesperrt werden.

Zur Umsetzung ist zum Beispiel eine öffentliche Datenbank mit Fingerprints jugendschutzrelevanter Web-Inhalte denkbar, die von verschiedenen öffentlichen Stellen und Filter-Herstellern gepflegt und genutzt werden kann.

Automatische Selektion und Wartung von Schlagwort-Listen: In bestehenden Filterprogrammen werden Webseiten unter anderem aufgrund des gehäuften Vorkommens von bestimmten Schlagwörtern klassifiziert. Dies erfordert großen manuellen Aufwand zur Erstellung und Wartung der entsprechenden Schlagwortlisten. Neue relevante Themen und Schlagwörter kommen hinzu und müssen manuell bewertet werden. In Rahmen der vorliegenden Studie konnte gezeigt werden, dass eine *automatische Selektion passender Schlagwörter* nicht nur (i) den manuellen Aufwand drastisch reduziert sondern auch (ii) bessere Klassifikationsergebnisse liefert als die manuelle Auswahl.

Das entsprechende Verfahren zur automatischen Auswahl von Schlagwörtern könnte unmittelbar von den Herstellern von Jugendschutzprogrammen zur direkten Verbesserung der Filterqualität eingesetzt werden.

7.2 Mittelfristig einsetzbare Verfahren

Die folgenden Empfehlungen erfordern den Aufbau von neuen organisatorischen und technischen Strukturen und eine Anpassung der zugrunde liegenden Technologien. Von diesem Empfehlungen ist ebenfalls ein deutlicher positiver Einfluss auf die Qualität der Filterlisten zu erwarten.

Automatische Text-Klassifikation für Inhalte »ab 18«: Die automatische Text-Klassifikation erzielt in den hier betrachteten inhaltlichen Kategorien (Gewalt, Extremismus, Selbstgefährdung) bereits mit vergleichsweise kleinen Trainingsdatensätzen sehr gute Erkennungsergebnisse. Die entsprechenden Klassifikationsmechanismen können daher zum Aufbau und zur Pflege von Filterlisten eingesetzt werden. In einer späteren Ausbaustufe sollte ein Workflow zur manuellen Korrektur aufgebaut werden, um die Ergebnisse kontinuierlich zu verbessern *(Active Learning)*.

Detektion jugendschutzrelevanter Symbole: In vielen Fällen deuten eindeutige Symbole und Logos darauf hin, dass ein Video oder Bild möglicherweise jugendschutzrelevant sein könnte. Dazu zählen zum Beispiel eingeblendete Logos oder URLs einschlägiger gewaltverherrlichender Seiten, aber auch extremistische Symbolik wie etwa Logos von rechtsextremen Bands oder Vereinigungen. Die automatische und robuste Erkennung solcher Symbole ist nach dem aktuellen Stand der Technik möglich und könnte zur Detektion noch ungeprüfter jugendschutzrelevanter Multimedia-Inhalte eingesetzt werden.

7.3 Empfehlungen für den Einsatz von Filterlisten

Filterlisten spielen im technischen Jugendmedienschutz eine wesentliche Rolle. Die folgenden Empfehlungen beziehen sich auf die effiziente und effektive Erstellung, Wartung und Nutzung dieser wichtigen Wissensbasis.

Feinere Granularität von Filterlisten: Angebote sollten auf Domain-Ebene und ergänzend auch auf URL-Ebene gefiltert werden. Die Ergänzung durch die feingranulare Filterung mittels URL ist nötig, um insbesondere populäre Portale mit nutzergenerierten Inhalten für eine bestimmte Altersklasse nicht in Gänze ein- oder ausschließen zu müssen. Statt dessen können einzelne Inhalte – unter Berücksichtigung der Herausforderungen an die nötige Infrastruktur – ergänzend gesperrt oder freigegeben werden.

Aufbau eines Systems zur automatischen Priorisierung von zu bewertenden Webseiten: Technische Klassifikationsverfahren können auch eingesetzt werden, um die redaktionelle manuelle Bewertung zu unterstützen und zu beschleunigen. So kann ein Verfahren zum Beispiel erkennen, wenn in einem Video Symbole verbotener Organisationen enthalten sind, und die entsprechenden Webseiten dann für eine redaktionelle Prüfung vorschlagen.

Crowdsourcing: Angesichts der umfangreichen Menge von Web-Inhalten bietet Crowdsourcing eine Möglichkeit, aus den enormen Datenmengen gezielt Inhalte für die redaktionelle Bewertung vorzuschlagen. Dadurch können freiwillige Teilnehmer (z. B. Eltern) und unabhängige Expertengruppen ihre Kräfte bündeln. Wird eine Seite zum Beispiel im Jugendschutzprogramm häufig durch die

Eltern für eine bestimmte Altersklasse gesperrt, so könnte diese Seite zur manuellen Prüfung und Einstufung an den Jugendschutzprogrammhersteller weitergeleitet werden. Diese Einstufung könnte dann in die Filterliste des Jugendschutzprogramms aufgenommen oder das Angebot bei unzulässigen oder strafrechtlich relevanten Inhalten an die jeweils zuständige Stelle weitergeleitet werden. Die beschriebenen Herausforderungen beim Crowdsourcing – etwa die möglicherweise fehlende Repräsentativität – spielen in diesem Anwendungsszenario keine Rolle.

Erstellung von Filterlisten: Kräfte bündeln: Die Erstellung von parallelen, redundanten Listen durch verschiedene Akteure im Jugendmedienschutz erscheint nicht sinnvoll – schließlich verwenden die verschiedenen Akteure vergleichbare Kriterien zur Klassifizierung von Inhalten basierend auf den vom Gesetzgeber gesteckten Rahmenbedingungen – und ein einzelner Inhalt sollte somit eigentlich eine eindeutige Altersklassifizierung besitzen. Die gemeinsame Arbeit an einer gebündelten Filterliste könnte die Qualität und Aktualität der Filterliste deutlich erhöhen. Hier könnte ein System etabliert werden, das die manuelle Bewertung durch verschiedene institutionelle Clearing-Stellen, die aktive Einbeziehung der Eltern mittels Crowdsourcing sowie die automatische Detektion von jugendschutzrelevanten Inhalten effizient kombiniert.

Alternative Geräte zur Filterung: Eine gemeinsame, nach transparenten Kriterien erstellte Filterliste könnte in vielfältigen Betriebsumgebungen eingesetzt werden – etwa in Jugendschutzprogrammen oder in Netzwerkgeräten wie dem heimischen Router. Die Möglichkeit, die Filterung näher an der Quelle durchzuführen, ist insbesondere vor dem Hintergrund der großen Vielfalt internetfähiger Geräte sinnvoll. Allerdings erfordert eine solche Lösung entsprechende Konzepte zur Gewährleistung der Nutzerautonomie, wie etwa individuelle Konfiguration von Filterlisten durch Eltern und individuelle Benutzerprofile für einzelne Familienmitglieder.

Automatische Qualitätssicherung bestehender Filterlisten: Automatische Verfahren können die Qualitätssicherung bestehender Filterlisten unterstützen. Im Rahmen einer regelmäßigen, automatischen Neubewertung bestehender Angebote können verschiedene Schritte zur Erhöhung der Listenqualität durchgeführt werden, zum Beispiel eine Detektion von inhaltlichen Veränderungen (die eine neue Einstufung oder erneute redaktionelle Überprüfung erforderlich machen) oder eine Überprüfung auf die Einführung einer validen age-de.xml-Kennzeichnung.

Infrastruktur für automatische Klassifikation: Da die automatische Klassifikation eines Web-Inhalts je nach Verfahren nicht zu vernachlässigende Rechenressourcen benötigt, ist eine Klassifikation auf dem Endgerät des Nutzers sowie eine wiederholte Klassifikation des gleichen Inhalts möglichst zu vermeiden. Daher bietet sich eine Cloud-Infrastruktur zur Bewertung von Web-Inhalten an, die zentral gewartet werden kann.

7.4 Empfehlungen für die technische Begleitforschung

Die rasante Entwicklung in der Forschung zu Text- und Multimedia-Analyseverfahren hat eine Reihe von Anwendungen ermöglicht, die schon jetzt unterstützend im Jugendmedienschutz verwendet werden können. Die Forschung sollte für dieses gesellschaftlich und politisch wichtige Thema weiter sensibilisiert und motiviert werden. Dies kann zum Beispiel durch entsprechende Förderprogramme mit dem Ziel einer **Fokussierung der Forschung auf die konkrete Anwendung im Jugendmedienschutz** erreicht werden. Auf Basis der vorliegenden Studie empfehlen wir exemplarisch die Weiterentwicklung auf folgenden Gebieten:

Bei der direkten Klassifikation von Webseiten sollte die **weitere Verbesserung von Klassifikationsmodellen im Kontext des Jugendmedienschutzes** angestrebt werden. Die Anwendbarkeit

solcher Techniken für den Bereich des Jugendmedienschutzes konnte in dieser Studie zwar exemplarisch belegt werden. Jedoch ist für eine Reihe von Verfahren noch eine spezielle Anpassung an dieses Einsatzszenario möglich, die einerseits zu besseren Ergebnissen führen kann und andererseits eine bessere Einschätzung der genauen Einsatzparameter dieser Verfahren ermöglicht.

Für den Einsatz **indirekter Merkmale**, wie sie zum Beispiel mit Hilfe von Symbolerkennung, Spracherkennung, optischer Texterkennung und Konzepterkennung gewonnen werden können, besteht noch Forschungsbedarf. Bei ausgereiften Verfahren wie Symbol-, Sprach- und Texterkennung muss inbesondere untersucht werden, wie aus den ermittelten Indizien eine tatsächliche Entscheidung über die Aufnahme in eine Filterliste getroffen werden kann. Dies gilt auch für die Konzepterkennung, jedoch müssen hier zunächst die eigentlichen Verfahren durch weitere Forschungsarbeit verbessert und an die Herausforderungen des Jugendmedienschutzes angepasst werden.

GLOSSAR

age-de.xml

age-de.xml ist ein technischen Label-Format für die Altersklassifizierung von Inhalten in Webangeboten und anderen Telemedien. Weiterführende Informationen und die technische Spezifikation sind auf den Webseiten von Online Management Kontor zur finden (siehe http://www.online-management-kontor.de/jugendschutz/altersklassifizierung.html, Stand 29.10.2012).

Altersklassifikation

Altersklassifikation bezeichnet im Sinne des Jugendmedienschutzes und der vorliegenden Studie die altersdifferenzierte Beurteilung eines Web-Inhaltes oder Webangebotes hinsichtlich der Eignung zur Entwicklungsbeeinträchtigung Minderjähriger einer bestimmten Altersstufe. Die Begriffe Altersklassifikation und Altersklassifizierung finden synonyme Anwendung.

Altersklassifizierung

siehe Altersklassifikation.

App

App (als Kurzform für Applikation bzw. englisch application) bezeichnet im Allgemeinen jede Form von Anwendungsprogrammen. »Als Teil von betriebssystembezogenen Softwareökosystemen sind sog. Appmärkte entstanden. Dabei wird unter dem Begriff App (als Kurzform für application) eine nicht integrierte Anwendung für Smartphones und Tabletcomputer verstanden, die über Onlineshops erhältlich ist und dann direkt auf dem Gerät installiert werden kann. Beispiele sind der Appstore von Apple und Androidmarket von Google.« [Krcmar, 2012].

ASCII

Genormter Zeichencode, der in Rechnern zur Darstellung bestimmter Informationen (Ziffern, Buchstaben und Sonderzeichen) verwendet wird.

Badwords

»Unerwünschte Begriffe«, können in zentralen Listen geführt werden.

Blacklist

Auf einer Blacklist (Negativliste) werden Angebote gesammelt, die jugendschutzrelevant sind und zu denen Kinder und Jugendliche – zumindest bis zu einer bestimmten Altersstufe – keinen Zugang erhalten sollen.

Bot

Aus dem Englischen abgekürzt für »Robot« (dt.: Roboter) wird damit ein Computerprogramm umschrieben, dass ohne menschliche Interaktion automatisch wiederholende Aufgaben durchführt. Suchmaschinen setzen Bots beispielsweise für die Erschließung des World Wide Web ein.

BPjM-Modul

»Die Bundesprüfstelle für jugendgefährdende Medien (BPjM) stellt in Kooperation mit der Freiwilligen Selbstkontrolle Multimedia-Diensteanbieter e. V. (FSM e. V.) Herstellern nutzerautonomer Filterprogramme das BPjM-Modul zur Verfügung. Das Modul ermöglicht die Filterung

der von der BPjM indizierten Online-Angebote. […] Bei den indizierten und mittels des BPjM-Moduls filterbaren Online-Angeboten handelt es sich zum einen um Angebote, die gemäß § 4 Absatz 1 Jugendmedienschutz-Staatsvertrag (JMStV) unzulässig sind. […] Das BPjM-Modul enthält zum anderen solche Angebote, deren Verbreitung außerhalb einer geschlossenen Benutzergruppe unzulässig ist (§ 4 Absatz 2 JMStV).« [BPjM, 2012b].

Crowdsourcing
Neologismus aus dem Wort *Crowd* (zu deutsch »Menschenmenge«) und den Ökonomie-Begriff *Outsourcing* (zu deutsch »Auslagern«), der das Konzept beschreibt eine zeit- oder kostenintensive Aufgabe in kleine Teilaufgaben, sogenannte Microtasks, zu zerlegen und diese an eine Menge von Internetnutzern zu verteilen.

Crowdturfing
Systematischer Einsatz von Crowdsourcing zur Manipulation von Marken-Reputation, Artikelbewertungen oder ähnliches in Social Media Angeboten [Wang et al., 2012].

Crowdworker
Beschreibt einen Internetnutzer der innerhalb eines Crowdsourcing Systems eine Microtask abarbeitet.

Cyberbullying
Cyberbullying ist eine Form des Cybermobbing, die insbesondere unter Kindern und Jugendlichen vorzufinden ist. »Die Täter – meistens Mitschüler – nennt man Bullys. Bullys beleidigen, bedrohen oder erpressen ihre Opfer entweder direkt oder aber sie versuchen, psychischen Druck auszuüben: Sie verbreiten Gerüchte und sorgen dafür, dass nimand mehr mit dem Betroffenen etwas zu tun haben will.« [ServiceBureau Jugendinformation, oJ].

Cybermobbing
Mit den aus dem Englischen kommenden Begriffen Cybermobbing (oder Cyber-Mobbing), auch Internet-Mobbing, Cyber-Bullying sowie Cyber-Stalking werden verschiedene Formen der Diffamierung, Belästigung, Bedrängung und Nötigung anderer Menschen oder Firmen mit Hilfe elektronischer Kommunikationsmittel über das Internet, in Chatrooms, beim Instant Messaging und/oder auch mittels Mobiltelefonen bezeichnet. Dazu gehört auch der Diebstahl von (virtuellen) Identitäten, um in fremden Namen Beleidigungen auszustoßen oder Geschäfte zu tätigen (Quelle: http://de.wikipedia.org/wiki/Cybermobbing, Stand 15.10.2012).

DNS
Domain Name System, Dienst zur Beantwortung von Anfragen zur Namensauflösung in einem IT-Netzwerk. Siehe http://de.wikipedia.org/wiki/Domain_Name_System, Stand 20.11.2012.

Effektivität
Effektivität ist ein Beurteilungskriterium, mit dem sich beschreiben lässt, ob eine Maßnahme geeignet ist, ein vorgegebenes Ziel zu erreichen. [Thommen et al., 2012] Effektivität ist ein Maß für die Zielerreichung (Wirksamkeit, Qualität der Zielerreichung).

Effizienz
Effizienz ist ein Beurteilungskriterium, mit dem sich beschreiben lässt, ob eine Maßnahme geeignet ist, ein vorgegebenes Ziel in einer bestimmten Art und Weise zu erreichen. Sie ist ein Merkmal der Softwarequalität. [Schaich et al., 2012].

Extremismus
Im Sinne des Jugendmedienschutzes und der vorliegenden Studie umfasst *Extremismus* Angebote, die weltanschauliche oder politische Totalitarismen oder religiöse Fundamentalismen andeuten, darstellen oder propagieren.

fragFINN
fragFINN ist eine Whitelist-basierte Kindersuchmaschine für das WWW, siehe http://www.fragfinn.de/, Stand 20.11.2012.

fraggen
Der im Vietnamkrieg eingesetzte Begriff *fraggen*, der ursprünglich das absichtliche Töten oder Verwunden eines Angehörigen der eigenen Mannschaft durch eine Handgranate beschrieb, wurde später in die Spielesprache übernommen und bezeichnet spätestens seit dem bekannten Spiel *Doom* das virtuelle Töten einer Spielerfigur.

FSM
Die Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e. V. ist ein gemeinnütziger eingetragener Verein, der sich in erster Linie dem Jugendmedienschutz widmet.

Gewalt
Im Sinne des Jugendmedienschutzes und der vorliegenden Studie beschreibt der Begriff *Gewalt* die »Gewalt ist die physische, psychische oder materielle Schädigung von personalen und dinglichen Objekten durch Subjekte, Naturgewalten oder Unfälle.« [KJM, 2010]. »Gewaltdarstellungen sind mediale Angebote, bei denen eine physische, psychische oder materielle Schädigung von Personen und Tieren sowie dinglichen Objekten dargestellt ist.« [FSM, 2011d].

Happy Slapping
»Umschreibung für das Phänomen, fremde Passanten, Mitschüler oder Lehrer zu schlagen und das auf Videokamera aufzunehmen, um es eventuell später ins Netz zu stellen.« [Duden, 2012].

HTML
Mithilfe der als ISO-Norm 8879 festgeschriebenen SGML (Standard Generalized Markup Language) definierte Auszeichnungssprache, die die logischen Bestandteile eines Dokuments wie Überschriften und Aufzählungen beschreibt. HTML wird dazu benutzt, Dokumente für das World Wide Web zu erstellen, die daraufhin unter Zuhilfenahme eines Browsers angezeigt werden [Lackes und Siepermann, 2012a].

HTTP
Protokoll zur Übertragung von Webseiten.

HTTPS
Protokoll zur sicheren Übertragung von Webseiten, via SSL/TLS.

ICRA
Die Internet Content Rating Association ist eine internationale non-profit Organisation, die sich selbst-regulierend für den Jugendschutz einsetzt. ICRA Deutschland betrachtet dabei die speziellen Anforderungen in Deutschland.

ISP
Internet Service Provider (Internetdienstanbieter oder Internetdienstleister), im deutschsprachigen Raum auch oft nur Provider, weniger häufig auch nur Internetanbieter oder Internetprovider genannt, sind Anbieter von Diensten, Inhalten oder technischen Leistungen, die für die Nutzung oder den Betrieb von Inhalten und Diensten im Internet erforderlich sind. (Quelle: http://de.wikipedia.org/wiki/Internetdienstanbieter, Stand 10.10.2012).

JMStV
Gesetz zum Staatsvertrag über den Schutz der Menschenwürde und den Jugendschutz in Rundfunk und Telemedien (Jugendmedienschutz-Staatsvertrag) [JMStV, 2003].

jugendschutzrelevant
Der Begriff *jugendschutzrelevant* bzw. *Jugendschutzrelevanz* bezieht sich in der vorliegenden Studie auf Inhalte oder Angebote, im im Sinne des geltenden Rechts als jugendgefährdend oder potentiell entwicklungsbeeinträchtigend anzusehen sind.

JuSchRiL
Gemeinsame Richtlinien der Landesmedienanstalten zur Gewährleistung des Schutzes der Menschenwürde und des Jugendschutzes [JuSchRiL, 2005].

KJM
Die Kommission für Jugendmedienschutz ist zuständig ist für die inhaltliche Kontrolle im Bereich des privaten Rundfunks und im Internet, siehe http://www.kjm-online.de, Stand 27.9.2012.

Klassifikator
Ein Klassifikator ist in der Informatik ein Algorithmus (typischerweise realisiert durch ein Computerprogramm), der eine mathematische Klassifikationsfunktion implementiert. Das bedeutet, dass er für jede Eingabe entscheidet, in welche einer vorgegebenen Menge von Klassen die Eingabe einzuordnen ist. Zum Beispiel kann ein Klassifikator als Eingabe eine Webseite erhalten und für jede Seite entscheiden, ob sie in die Klasse »jugendschutzrelevant« oder in die Klasse »nicht jugendschutzrelevant« gehört. Manche Klassifikatoren erlauben nur für einen Teil der möglichen Eingaben eine Klassifikation oder liefern zusätzlich eine Konfidenz für die vorgenomme Einordnung mit.

Konfidenz
»Ein Konfidenzintervall (auch ›Vertrauensbereich‹ oder ›Vertrauensintervall‹ genannt) ist ein Begriff aus der Statistik. Er sagt etwas über die Präzision der Lageschätzung eines Parameters (zum Beispiel eines Mittelwertes) aus. Konfidenzintervalle schließen einen Bereich um den geschätzten Wert des Parameters ein, der mindestens mit einer zuvor festgelegten Wahrscheinlichkeit (dem Konfidenzniveau) die wahre Lage des Parameters trifft« (Quelle: http://de.wikipedia.org/wiki/Konfidenzintervall, Stand 18.12.2012).

Korrelation
Eine Korrelation beschreibt den Zusammenhang von Merkmalen. Bei einer hohen positiven Korrelation treten Merkmale häufig gemeinsam und selten getrennt voneinander auf. Bei einer negativen Korrelation verhält es sich analog, die Merkmale treten also selten gemeinsam auf und häufig getrennt voneinander.

Kriterium
Im vorliegenden Zusammenhang sind mit Kriterien überprüfbare Eigenschaften von Internetinhalten gemeint. Insbesondere bei der Einordnung von Medien für die Altersfreigabe führt die Prüfung bzw. Bewertung entlang eines Kriteriums zu einer Klassifikation eines Inhaltes. Typischerweise liegen dazu eine Reihe von Kriterien in einem Kriterienkatalog vor.

Microtask
Beschreibt eine abgeschlossene, in kurzem zeitlichen Rahmen abzuarbeitende Teilaufgabe im Rahmen eines Crowdsourcing Projektes.

Negativliste
siehe Blacklist.

Overblocking
Im Sinne des Jugendmedienschutzes für das WWW und der vorliegenden Studie bezeichnet *Overblocking* ein Fehlverhalten von Filterprogrammen, bei dem nicht-jugendschutzrelevante

Inhalte als jugendschutzrelevant identifiziert werden und dadurch ein Zugriff auf die Inhalte verhindert wird.

PC
Personal Computer (Einzelplatzrechner), der im Gegensatz zu Großrechnern oder Workstations preisgünstig genug für Privathaushalte ist und die Anforderungen an einen Heimcomputer erfüllt (Quelle: http://de.wikipedia.org/wiki/Personal_Computer, Stand 10.10.2012).

PEGI-System
Das Pan-European Game Information System ist ein Kennzeichnungssystem zur Alterseinstufung der Spiele-Industrie.

Perplexität
Perplexität ist eine Maßgröße, die zur Qualitätsbeurteilung von Wahrscheinlichkeitsmodellen herangezogen werden kann.

Positivliste
siehe Whitelist.

Pro-Ana
Pro-Ana bezeichnet eine Bewegung, die die Anorexia nervosa (Magersucht) als Lebensideal darstellt.

Pro-Mia
Pro-Mia bezeichnet eine Bewegung, die die Bulimia nervosa (Ess-Brechsucht) als Lebensideal darstellt.

Proxy
»Proxy heißt Bevollmächtigter oder Stellvertreter. Ein Proxy-Server ermöglicht Systemen, die keinen direkten Zugang zum Internet haben, den indirekten Zugang zum Netz. Das können solche Systeme sein, die durch eine Firewall aus Sicherheitsgründen vom unmittelbaren Zugang ausgeschlossen sind. Damit die Unternehmensnetzstruktur und die Teilnehmeradressen dem Beobachter aus Internet gegenüber verborgen bleiben, tauscht der Proxy-Server die Client-Adressen gegen Proxy-Adressen aus. Ein Proxy kann außerdem einzelne Datenpakete aus dem Datenstrom zwischen dem Internet und einem lokalen Netz über Paketfilter ausfiltern und so zur Erhöhung der Sicherheit beitragen.« [ITwissen.info, 2012c].

Router
Router sind Koppelelemente in der Netzwerktechnik, die zwei oder mehr Subnetze auf der Vermittlungsschicht miteinander verbinden und die Funktionen bis zu dieser Schicht realisieren [ITwissen.info, 2012e]. Router stellen im typischerweise den Übergabepunkt zwischen dem Netzwerk des ISP (z. B. DSL-Zugang) und dem Heimnetzwerk dar.

RTA
»Restricted To Adults«, Auszeichnung der amerikanischen Non-Profit Organisation *Association of Sites Advocating Child Protection (ASACP)*.

Schockerseite
Eine Schockerseite ist eine Webseite, die absichtlich und meist schon auf der Startseite besonders anstößiges, verstörendes oder obszönes Material zeigt. Oft werden im Zuge eines üblen Scherzes nichts ahnende Internetbenutzer dorthin gelockt, etwa durch einen harmlos wirkenden Link mit dem Verweis »niedliche Katze« o. ä. [Kirkpatrick, 2004].

Selbstgefährdung
In der vorliegenden Studie werden aufgrund der unterschiedlichen Definitionen (vgl. Abschnitt 3.4.1) und der ausgewählten Beispieldaten unter dem Begriff Selbstgefährdung (auch: selbstgefährdendes Verhalten) zusammengefasst.

Smart Device
Als Smart Devices werden informationstechnisch aufgerüstete Alltagsgegenstände bezeichnet, die einen Mehrwert durch sensorgestützte Informationsverarbeitung und Kommunikation erhalten und im Sinne der vorliegenden Studie über einen Internetzugang verfügen (Quelle: http://wirtschaftslexikon.gabler.de/Archiv/76910/smart-devices-v5.html, Stand 12.10.2012).

Smart TV
Smart TV, manchmal auch Hybrid-TV, ist die Bezeichnung für Fernsehgeräte mit Computer-Zusatzfunktionen, insbesondere Internet-Fähigkeiten. (Quelle: http://de.wikipedia.org/wiki/Smart_TV, Stand 10.10.2012.

Social Tagging
Social Tagging ist eine Form der freien Verschlagwortung (Indexierung), bei der Nutzer von Inhalten die Deskriptoren (Schlagwörter) mit Hilfe verschiedener Arten von sozialer Software ohne Regeln zuordnen. Die bei diesem Prozess erstellten Sammlungen von Schlagwörtern werden Folksonomien genannt, siehe http://de.wikipedia.org/wiki/Social_Tagging, Stand 20.11.2012.

Streisand-Effekt
Wenn bei dem Versuch, eine bestimmte Information oder einen Web-Inhalt zu unterdrücken, das genaue Gegenteil erreicht wird, dass nämlich besonders viele Sekundärquellen die Information beschreiben oder spiegeln, spricht man vom Streisand-Effekt. Der Name geht darauf zurück, dass die Sängerin und Schauspielerin Barbara Streisand gegen die Veröffentlichung eines Luftfotos ihres Strandhauses geklagt hatte und erst daraufhin die Aufmerksamkeit der Medien und des Internets auf ihr Grundstück lockte [Rogers, 2003].

TRECVid
Text Retrieval Conference Video Retrieval Evaluierung ist eine jährlich stattfinde Aufgabe, aus Video-Material eine Inhaltsbasierte Analyse durchzuführen, die dann von einem Expertenteam bewertet wird.

Underblocking
Im Sinne des Jugendmedienschutzes für das WWW und der vorliegenden Studie bezeichnet *Underblocking* ein Fehlverhalten von Filterprogrammen, bei dem jugendschutzrelevante Inhalte nicht als solche identifiziert werden.

ungeeignete Inhalte
Inhalte, die gemäß gesetzlichen Vorgaben oder dem erzieherischen Konzept der Eltern zu Folge nicht für den kindlichen bzw. jugendlichen Internetnutzer geeignet sind. *Inhalte* kann sich ebenfalls auf Kommunikation beziehen.

URL
Ein Uniform Resource Locator identifiziert und lokalisiert eine Ressource wie beispielsweise eine Webseite über die zu verwendende Zugriffsmethode und den Ablageort.

vertikale Suchmaschine
»Vertikale Suchmaschinen konzentrieren sich auf bestimmte Themengebiete […] oder spezielle Zielgruppen […]. Vertikale Suchmaschinen betrachten im Unterschied zu allgemeinen Suchmaschinen nur einen Teilbereich des World Wide Web und erstellen einen Index von Webdokumenten, die mit dem betrachteten Themengebiet korrespondieren. Technisch funktioniert eine vertikale Suchmaschine wie eine Universal-Suchmaschine« (Quelle: http://de.wikipedia.org/wiki/Vertikale_Suchmaschine, Stand 18.12.2012).

VO-S
Verfahrensordnung Freiwillige Selbstkontrolle Multimedia e. V. für Verfahren bei Beschwerden gegen Verstöße gegen den Verhaltenssubkodex der Suchmaschinenanbieter.

Webangebot
siehe Webauftritt.

Webauftritt
Ein *Webauftritt* (engl.: Website) ist eine Sammlung von zusammengehörigen Webseiten die unter einer gemeinsamen Domain in einem Netzwerk erreichbar sind. Alle im Internet öffentlich verfügbaren *Webauftritte* zusammen bilden das World Wide Web.

Webseite
Eine *Webseite* (engl.: Webpage) ist ein einzelnes Dokument das unter Angabe eines Uniform Resource Locators (URL) über einen Webbrowser abgerufen werden kann. In der Regel meint man dabei ein einzelnes HTML-Dokument. Zusammengehörige Webseiten bilden einen Webauftritt.

Web 2.0
»Web 2.0 ist ein Schlagwort, das für eine Reihe interaktiver und kollaborativer Elemente des Internets, speziell des World Wide Webs, verwendet wird. Hierbei konsumiert der Nutzer nicht nur den Inhalt, er stellt als Prosument selbst Inhalt zur Verfügung« (Quelle: http://de.wikipedia.org/wiki/Web_2.0, Stand 29.10.2012).

Whitelist
Webseiten speziell für Kinder und Angebote ohne Jugendschutzrelevanz können auf einer Whitelist (Positivliste) zusammengefasst werden, die als Grundlage für einen sicheren Surfraum dienen kann.

WLAN
Wireless Local Area Network – drahtloses lokales Netzwerk, bei dem die Rechnerkommunikation über Funk nach dem IEEE 802.11 Standard abgewickelt wird.

ZDF
Zweites Deutsches Fernsehen, Anstalt des öffentlichen Rechts.

ABKÜRZUNGEN

aj	Aktion Jugendschutz, Landesarbeitsstelle Bayern.
AJS	Arbeitsgemeinschaft Kinder- und Jugendschutz.
AKJS	Aktion Kinder- und Jugendschutz, Landesarbeitsstelle Brandenburg e. V..
ASACP	Association of Sites Advocating Child Protection, Webadresse: http://www.asacp.org/, Stand 20.11.2012.
ASCII	American Standard Code for Information Interchange (Glossar: ASCII).
BAJ	Bundesarbeitsgemeinschaft Kinder- und Jugendschutz e. V..
BBFC	British Board of Film Classification, Webadresse: http://www.bbfc.co.uk/, Stand 9.10.2012.
BfV	Bundesamt für Verfassungsschutz.
BKM	Der Beauftragte der Bundesregierung für Kultur und Medien, Webadresse: http://www.kulturstaatsminister.de, Stand 26.9.2012.
BLM	Bayerische Landeszentrale für neue Medien, Webadresse: http://www.blm.de, Stand 11.10.2012.
BMFSFJ	Bundesministerium für Familie, Senioren, Frauen und Jugend, Webadresse: http://www.bmfsfj.de, Stand 27.9.2012.
BMI	Bundesministerium des Innern.
bpb	Bundeszentrale für politische Bildung.
BPjM	Bundesprüfstelle für jugendgefährdende Medien, Webadresse: http://www.bundespruefstelle.de, Stand 27.9.2012.
BSI	Bundesamt für Sicherheit in der Informationstechnik.
BZgA	Bundeszentrale für gesundheitliche Aufklärung, Webadresse: http://www.bzga.de/, Stand 25.10.2012.
CERO	Computer Entertainment Rating Organization, Webadresse: http://www.cero.gr.jp/e/index.html, Stand 9.10.2012.
CPU	Central Processing Unit, dt. Hauptprozessor.
CRF	Conditional Random Field.
DMK	Dynamische Markov-Ketten.
DNS	Domain Name System (Glossar: DNS).
DsiN	Deutschland sicher im Netz e. V..
DVO JuSchG	Verordnung zur Durchführung des Jugendschutzgesetzes [DVO JuSchG, 2003].
EKD	Evangelische Kirche Deutschland, Webadresse: http://www.ekd.de, Stand 11.10.2012.
EKM	Evangelische Kirche in Mitteldeutschland, Webadresse: http://www.ekmd.de, Stand 11.10.2012.
ESRB	Entertainment Software Rating Board, Webadresse: http://www.esrb.org/, Stand 9.10.2012.

FOSI	Family Online Safety Institute, Webadresse: http://www.fosi.org/, Stand 20.11.2012.
FQDN	Fully Qualified Domain Name (dt.: vollständig qualifizierter Domainname).
FSF	Freiwillige Selbstkontrolle Fernsehen e.V., Webadresse: http://www.fsf.de, Stand 27.9.2012.
FSK	Freiwillige Selbstkontrolle der Filmwirtschaft GmbH, Webadresse: http://www.fsk.de, Stand 27.9.2012.
FSK.online	Freiwillige Selbstkontrolle der Filmwirtschaft Abteilung FSK.online.
FSM	Freiwillige Selbstkontrolle Multimedia-Diensteanbieter(Glossar: FSM).
GG	Grundgesetz für die Bundesrepublik Deutschland [Grundgesetz, 2012].
HBI	Hans-Bredow-Institut für Medienforschung an der Universität Hamburg, Webadresse: http://www.hans-bredow-institut.de, Stand 5.10.2012.
HMM	Hidden Markov Modell.
HTML	Hypertext Markup Language (Glossar: HTML).
HTTP	Hypertext Transfer Protocol (Glossar: HTTP).
HTTPS	Hypertext Transfer Protocol Secure (Glossar: HTTPS).
ICRA	Internet Content Rating Association (Glossar: ICRA).
ISP	Internet Service Provider (Glossar: ISP).
JK	Juristen-Kommission.
JMStV	Jugendmedienschutz-Staatsvertrag (Glossar: JMStV).
JuSchG	Jugendschutzgesetz des Bundes [JuSchG, 2008].
JuSchRiL	Jugendschutzrichtlinien (Glossar: JuSchRiL).
KJF	Kinder- und Jugendfilmzentrum in Deutschland, Webadresse: http://www.kjf.de/, Stand 25.10.2012.
KJM	Kommission für Jugendmedienschutz (Glossar: KJM).
LDA	Latent Dirichlet Allocation.
LfM	Landesanstalt für Medien Nordrhein-Westfalen, Webadresse: http://www.lfm-nrw.de, Stand 27.9.2012.
LGPL	Lesser General Public License, Webadresse: http://www.gnu.org/licenses/lgpl.html, Stand 20.11.2012.
LMK	Landeszentrale für Medien und Kommunikation Rheinland-Pfalz, Webadresse: http://www.lmk-online.de, Stand 11.10.2012.
LPR	Hessische Landesanstalt für privaten Rundfunk und neue Medien, Webadresse: http://www.lpr-hessen.de, Stand 11.10.2012.
MPAA	Motion Picture Association of America, Webadresse: http://www.mpaa.org/, Stand 9.10.2012.
MPFS	Medienpädagogischer Forschungsverbund Südwest, Webadresse: http://www.mpfs.de, Stand 5.10.2012.
NER	Named Entity Recognition.
NIST	National Institute of Standards and Technology, Webadresse: http://www.nist.gov, Stand 23.10.2012.
OFLC	Office of Film and Literature Classification, Webadresse: http://www.classification.gov.au, Stand 9.10.2012.

PC	Personal Computer (Glossar: PC).
PEGI	Pan European Game Information.
PEGI-System	Pan-European Game Information System (Glossar: PEGI-System).
Pro-Ana	Pro Anorexie nervosa (Glossar: Pro-Ana).
Pro-Mia	Pro Bulimia nervosa (Glossar: Pro-Mia).
RTA	Restricted To Adults (Glossar: RTA).
SPIO	Spitzenorganisation der Filmwirtschaft e. V..
SPIO/JK	Juristen-Kommission der Spitzenorganisation der Filmwirtschaft e. V..
StGB	Strafgesetzbuch [StGB, 2012].
SVM	Support Vektor Maschine.
TLM	Thüringer Landesmedienanstalt, Webadresse: http://www.tlm.de, Stand 11.10.2012.
TRECVid	Text Retrieval Conference Video Retrieval (Glossar: TRECVid).
UGC	User Generated Content (dt.: von Nutzern erzeugt Inhalte), siehe http://de.wikipedia.org/wiki/User-Generated-Content, Stand 20.11.2012.
URL	Uniform Resource Locator (Glossar: URL).
USK	Unterhaltungssoftware Selbstkontrolle, Webadresse: http://www.usk.de, Stand 27.9.2012.
USK.online	Unterhaltungssoftware Selbstkontrolle für den Bereich der Telemedien.
VK-C	Verhaltenssubkodex für Chatanbieter der FSM.
VK-S	Verhaltenssubkodex für Suchmaschinenanbieter der FSM.
VO-S	Verfahrensordnung Freiwillige Selbstkontrolle Multimedia (Glossar: VO-S).
WLAN	Wireless Local Area Network (Glossar: WLAN).
WWW	World Wide Web.
ZDF	Zweites Deutsches Fernsehen (Glossar: ZDF).

LITERATURVERZEICHNIS

[Antoniskek et al., 2011] Antoniskek, B., Fiscus, J., Over, P., Michel, M., Strassel, S., und Morris, A. (2011). 2011 TRECVID Workshop – Multimedia Event Detection. In *TRECVID 2011 Folien*. 181

[Ars Technica, 2011] Ars Technica (2011). Another fraudulent certificate raises the same old questions about certificate authorities. http://arstechnica.com/security/2011/08/earlier-this-year-an-iranian/, Stand 30.11.2012. 271

[Arthur, 2012] Arthur, C. (2012). David Cameron's 'porn filter' will make no difference, say ISPs. http://www.guardian.co.uk/technology/2011/oct/11/david-cameron-porn-filter-isps, Stand 1.12.2012. 274

[Assion, 2012] Assion, S. (2012). JMStV: Nachrichten-Privileg für Blogger? *Telemedicus*. 51

[Asthana, 2012] Asthana, S. (2012). Rolling out HTTPS for all Users. https://developers.facebook.com/blog/post/2012/11/14/platform-updates--operation-developer-love/, Stand 1.12.2012. 270

[AVG Technologies, 2012] AVG Technologies (2012). Introducing AVG Family Safety for Windows Phone and Major iOS Update. http://blogs.avg.com/view-from-the-top/introducing-avg-family-safety-windows-phone-major-ios-update/, Stand 1.11.2012. 40

[Backes, 2001] Backes, U. (2001). Gestalt und Bedeutung des intellektuellen Rechtsextremismus in Deutschland. *Aus Politik und Zeitgeschichte (B 46/2001)*. 102

[BAJ, 2012] BAJ (2012). Themenschwerpunkte Jugendschutz. http://www.jugendschutzlandesstellen.de/jugendschutz_sucht_medien_gewalt.html, Stand 30.10.2012. 23

[Bardeli et al., 2012] Bardeli, R., Schwenninger, J., und Stein, D. (2012). Audio fingerprinting for media synchronisation and duplicate detection. In *Media Synchronization Workshop (MEDIASYNC 2012)*. 159, 203

[Bay et al., 2008] Bay, H., Ess, A., Tuytelaars, T., und Van Gool, L. (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359. 232

[BBFC, 2009] BBFC (2009). BBFC Classification Guidelines. http://www.bbfc.co.uk/download/guidelines/BBFCClassificationGuidelines2009.pdf, Stand 10.10.2012. 56

[Bergholz, 2009] Bergholz, A. (2009). Antiphish – lessons learnt. Technischer Bericht, Invited talk on the workshop on CyberSecurity and Intelligence Informatics (CSI-KDD), 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 154

[Bergholz et al., 2008] Bergholz, A., Chang, J.-H., Paaß, G., Reichartz, F., und Strobel, S. (2008). Improved phishing detection using model-based features. In *Proc. Conference on Email and AntiSpam CEAS 2008*, Mountain View, CA. 147

[Bergholz et al., 2009] Bergholz, A., De Beer, J., Glahn, S., Moens, M.-F., Paaß, G., und Strobel, S. (2009). New filtering approaches for phishing email. *Journal of Computer Security*. accepted for publication. 146

[Bergholz et al., 2010] Bergholz, A., De Beer, J., Glahn, S., Moens, M.-F., Paaß, G., und Strobel, S. (2010). New filtering approaches for phishing email. *Journal of Computer Security*, 18:5–35. accepted for publication. 146, 151

[Bird et al., 2009] Bird, S., Klein, E., und Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. 149

[BITKOM, 2012] BITKOM (2012). Smartphone-Funktionen: Internet wichtiger als Telefonieren. http://www.bitkom.org/files/documents/BITKOM_Presseinfo_Smartphone-Funktionen_05_07_2012.pdf, Stand 16.10.2012. 73

[BKM, 2012a] BKM (2012a). Medienkompetenz - Projekte und Initiativen des Beauftragten der Bundesregierung für Kultur und Medien. http://www.bundesregierung.de/Content/DE/_Anlagen/BKM/2012-09-25-flyer-medienkompetenz.pdf?__blob=publicationFile&v=1, Stand 26.9.2012. 12

[BKM, 2012b] BKM (2012b). Staatsminister für Kultur und Medien Bernd Neumann – Deutscher Computerspielpreis. http://www.bundesregierung.de/Webs/Breg/DE/Bundesregierung/BeauftragterfuerKulturundMedien/medien/medienkompetenz/dtComputerspielpreis/_node.html, Stand 30.10.2012. 12

[BKM, 2012c] BKM (2012c). Staatsminister für Kultur und Medien Bernd Neumann – Medienkompetenz. http://www.bundesregierung.de/Webs/Breg/DE/Bundesregierung/BeauftragterfuerKulturundMedien/medien/medienkompetenz/_node.html, Stand 30.10.2012. 11

[Blei et al., 2003] Blei, D. M., Ng, A. Y., und Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022. 146

[Blue Coat Systems, 2012] Blue Coat Systems (2012). Blue Coat präsentiert sicheren, kostenlosen und familienfreundlichen Webbrowser für Android. http://www.bluecoat.de/news/deutschenews.php?id=231, Stand 1.11.2012. 40

[BMFSFJ, 2011] BMFSFJ (2011). Bundesfamilienministerin stellt Initiativen für eine Kinder- und Jugendnetzpolitik vor. http://www.bmfsfj.de/BMFSFJ/kinder-und-jugend,did=175940.html, Stand 30.10.2012. 12, 13

[BMFSFJ, 2012a] BMFSFJ (2012a). I-KiZ – Zentrum für Kinderschutz im Internet – Einladung zur Auftaktveranstaltung. http://www.bmfsfj.de/RedaktionBMFSFJ/Pressestelle/Pdf-Anlagen/0910-Programm-I-Kiz,property=pdf,bereich=bmfsfj,sprache=de,rwb=true.pdf, Stand 30.10.2012. 13

[BMFSFJ, 2012b] BMFSFJ (2012b). Kristina Schröder: »Offline und Online brauchen wir ein soweit wie möglich vergleichbares Niveau des Jugendschutzes«. http://www.bmfsfj.de/BMFSFJ/kinder-und-jugend,did=185766.html, Stand 30.10.2012. 12

[Bordes et al., 2011] Bordes, A., Weston, J., Collobert, R., und Bengio, Y. (2011). Learning structured embeddings of knowledge bases. In *AAAI 2011*. 140

[BPjM, 2012a] BPjM (2012a). Aufgaben und Angebote der Bundesprüfstelle. http://www.bundespruefstelle.de/bpjm/die-bundespruefstelle.html, Stand 30.10.2012. 13, 14

[BPjM, 2012b] BPjM (2012b). Filterung indizierter Online-Angebote und Hinweise zum BPjM-Modul. http://www.bundespruefstelle.de/bpjm/redaktion/PDF-Anlagen/bpjm-modul,property=pdf,bereich=bpjm,sprache=de,rwb=true.pdf, Stand 2.10.2012. 294

[Bratko et al., 2006] Bratko, A., Cormack, G. V., Filipic, B., Lynam, T. R., und Zupan, B. (2006). Spam filtering using statistical data compression models. *Journal of Machine Learning Research*, 6:2673–2698. 146

[Bratvold, 2011] Bratvold, D. (2011). Crowd Census I: Microtasking Market Research Report. Technischer Bericht, dailycrowdsource.com. 276

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140. 143

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32. 140

[Brin und Page, 1998] Brin, S. und Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 107–117, Brisbane, Australia. 182

[BSI, 2008] BSI (2008). Internet safety - Access control systems for the protection of children online - Specification (PAS 74:2008). Publicly available specification. 77

[BSI, 2012] BSI (2012). Das BSI: Chancen nutzen – Risiken vermeiden. https://www.bsi.bund.de/DE/DasBSI/dasbsi_node.html, Stand 20.11.2012. 15

[Cesarini et al., 1997] Cesarini, F., Francesconi, E., Gori, M., Marinai, S., Sheng, J., und Soda, G. (1997). A neural-based architecture for spot-noisy logo recognition. In *ICDAR*, pages 175–179. 176

[Cesarini et al., 1995] Cesarini, F., Gori, M., Marinai, S., und Soda, G. (1995). A hybrid system for locating and recognizing low level graphic items. In *GREC*, pages 135–147. 176

[Chen et al., 2011] Chen, J., Wang, L., und Chen, D. (2011). *Logo Recognition: Theory and Practice*. Taylor & Francis. 176

[Ciocca und Schettini, 2001] Ciocca, G. und Schettini, R. (2001). Content-based similarity retrieval of trademarks using relevance feedback. *Pattern Recognition*, 34(8):1639–1655. 176

[ContentWatch Inc., 2011] ContentWatch Inc. (2011). New Net Nanny to expand internet safety to mobile devices. http://www.netnanny.com/company/press/expanding_safety_to_mobile, Stand 1.11.2012. 40

[Cosmos Media, 2012] Cosmos Media (2012). Über »Helles-Koepfchen.de« – Das Wissensportal für Kinder und Jugendliche. , Stand 30.10.2012. 24

[Cybion et al., 2012] Cybion, Chancen, S. D., und Europe, I. (2012). Benchmarking of parental control tools for the online protection of children SIP-Bench II - Assessment results and methodology 4th Cycle. Technischer Bericht, INNOVA Europe, Brüssel, Belgien. 37, 38, 39, 40, 44, 68, 81

[Das et al., 2012] Das, S. N., Mathew, M., und K.Vijayaraghavan, P. (2012). An efficient approach for finding near duplicate web pages using minimum weight overlapping method. In *Proc. 2012 Ninth International Conference on Information Technology - New Generations*. 284

[Dialog Internet, 2011] Dialog Internet (2011). Thesenpapier – Auf dem Weg zu einer zeitgemäßen Kinder- und Jugendnetzpolitik. , Stand 30.10.2012. 13

[Dialog Internet, 2012a] Dialog Internet (2012a). I-KiZ – Zentrum für Kinderschutz im Internet. http://dialog-internet.de/web//initiativen_internet-kinderschutz-zentrum, Stand 30.10.2012. 13

[Dialog Internet, 2012b] Dialog Internet (2012b). Initiative »sicher online gehen – Kinderschutz im Internet«. http://dialog-internet.de/web//initiativen_jugendschutz-software, Stand 30.10.2012. 13

[Dika, 1990] Dika, V. (1990). *Games of terror: Halloween, Friday the 13th, and the films of the stalker cycle*. Fairleigh Dickinson University Press. 99

[DJI, 2011] DJI (2011). Ausgewählte Links zum Kinder- und Jugendschutz im Internet. http://www.dji.de/cgi-bin/projekte/output.php?projekt=898&Jump1=RECHTS&Jump2=32, Stand 30.10.2012. 15, 22

[Doan et al., 2011] Doan, A., Ramakrishnan, R., und Halevy, A. (2011). Crowdsourcing systems on the world-wide web. In *Communications of the ACM Vol. 54, No. 4*, Communications of the ACM. 276, 282

[Doermann et al., 1996] Doermann, D., Rivlin, E., und Weiss, I. (1996). Applying algebraic and differential invariants for logo recognition. *Mach. Vis. Appl.*, 9(2):73–86. 176

[Douze et al., 2010] Douze, M., Jégou, H., Schmid, C., und Pérez, P. (2010). Compact Video Description for Copy Detection with Precise Temporal Alignment. In *European Conference on Computer Vision (ECCV '10)*, Heraklion, Greece. 205

[Dreyer und Hajok, 2012] Dreyer, S. und Hajok, D. (2012). Lieber mit oder ohne? Die Sicht der Eltern auf Jugendschutzprogramme. *tv diskurs. 16. Jahrgang, Heft 02/12. Berlin 2012, S. 54 - 59*. 44

[DsiN, 2012a] DsiN (2012a). Mitglieder – Starke Partnerschaft für Deutschland sicher im Netz. https://www.sicher-im-netz.de/wir_ueber_uns/Mitglieder.aspx, Stand 31.10.2012. 22

[DsiN, 2012b] DsiN (2012b). Wir über uns – Deutschland sicher im Netz e.V. https://www.sicher-im-netz.de/wir_ueber_uns/wir_ueber_uns.aspx, Stand 30.10.2012. 22

[Duden, 2012] Duden (2012). Duden Neues Wörterbuch der Szenesprachen, Eintrag Happy Slapping. http://szenesprachenwiki.de/definition/happy-slapping/, Stand 30.11.2012. 295

[DVO JuSchG, 2003] DVO JuSchG (2003). Verordnung zur Durchführung des Jugendschutzgesetzes vom 9. September 2003 (BGBl. I S. 1791), die durch Artikel 4 Absatz 56 des Gesetzes vom 5. Mai 2004 (BGBl. I S. 718) geändert worden ist. http://www.gesetze-im-internet.de/bundesrecht/juschgdv/gesamt.pdf, Stand 10.10.2012. 5, 301

[EC, 2012a] EC (2012a). Bürgerinfo – EU-Strategie für ein besseres Internet für Kinder. http://ec.europa.eu/information_society/activities/sip/docs/bik/de.pdf, Stand 13.11.2012. 23

[EC, 2012b] EC (2012b). COM(2012) 196 final – Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen – Europäische Strategie für ein besseres Internet für Kinder. http://eur-lex.europa.eu/LexUriServ/site/de/com/2012/com2012_0196de01.pdf, Stand 26.11.2012. 23

[EC, 2012c] EC (2012c). Digitale Agenda: Neue Strategie für ein sicheres Internet und bessere Online-Inhalte für Kinder und Jugendliche – Pressemitteilung. http://europa.eu/rapid/press-release_IP-12-445_de.htm, Stand 13.11.2012. 23

[eco, 2012a] eco (2012a). Kompetenzgruppe Jugendschutz – Zielsetzung. http://jugendschutz.eco.de/zielsetzung/, Stand 30.10.2012. 27

[eco, 2012b] eco (2012b). Über eco: Wir gestalten das Internet. http://www.eco.de/about.html, Stand 30.10.2012. 27

[Eickhoff et al., 2011] Eickhoff, C., Serdyukov, P., und de Vries, A. P. (2011). A combined topical/non-topical approach to identifying web sites for children. In *Proc. WSDM 2011*. 153, 184

[Eifler, 2011] Eifler, S. (2011). Das System des Jugendmedienschutzes in Jugendschutzgesetz und Jugendmedienschutz-Staatsvertrag (JurPC Web-Dok. 40/2011, Abs. 1 - 43). JurPC Web-Dok. 40/2011 - DOI 10.7328/jurpcb/201126340, http://www.jurpc.de/jurpc/show?id=20110040, Stand 21.09.2012. 3, 4

[Ellis et al., 2011] Ellis, D. P., Whitman, B., und Porter, A. (2011). Echoprint: An open music identification service. In *ISMIR 2011 Miami: 12th International Society for Music Information Retrieval Conference*. 227

[Erdmann et al., 2012] Erdmann, M., Nguyen, D. D., Takeyoshi, T., Hattori, G., Matsumoto, K., und Ono, C. (2012). Hierarchical training of multiple svms for personalized web filtering. In *Proc. PRICAI 2012*. 152

[Erfurter Netcode, 2012] Erfurter Netcode (2012). Folgende Institutionen unterstützen uns. http://www.erfurter-netcode.de/55.html, Stand 30.10.2012. 25

[ESRB, 2012] ESRB (2012). ESRB Ratings Guide. http://www.esrb.org/ratings/ratings_guide.jsp, Stand 20.11.2012. 56

[EU Kids Online, 2010a] EU Kids Online (2010a). Risiken im Internet – Erste Befunde einer empirischen Untersuchung zur Onlinenutzung von Kindern und Jugendlichen – Überblick über europäische Ergebnisse und ausgewählte Befunde für Deutschland. Technischer Bericht, EU Kids Online Network. 44, 45

[EU Kids Online, 2010b] EU Kids Online (2010b). Risks and safety on the internet - The perspective of European children - Initial findings from the EU Kids Online survey of 9-16 year olds and their parents. Technischer Bericht, EU Kids Online Network. 43, 44

[European Commission, 2012a] European Commission (2012a). PEGI Online. http://ec.europa.eu/information_society/activities/sip/projects/completed/filtering_content_labelling/content_labelling/pegi/index_en.htm, Stand 30.10.2012. 37

[European Commission, 2012b] European Commission (2012b). Safer Internet Programme: the main framework for European policy. http://ec.europa.eu/information_society/activities/sip/policy/programme/index_en.htm, Stand 30.10.2012. 37

[European Commission, 2012c] European Commission (2012c). SIP-BENCH. http://ec.europa.eu/information_society/activities/sip/projects/completed/filtering_content_labelling/filtering/sip_bench/index_en.htm, Stand 30.10.2012. 37, 78

[European Commission, 2012d] European Commission (2012d). SIP-BENCH 2: Benchmarking of parental control tools for the online protection of children. http://ec.europa.eu/information_society/activities/sip/projects/filter_label/sip_bench2/index_en.htm, Stand 30.10.2012. 37, 78

[F-Secure GmbH, 2012] F-Secure GmbH (2012). F-Secure Sicherheitslösungen für Smartphones und Tablets nun in Google Play und im Apple App Store erhältlich. http://www.f-secure.com/de/web/home_de/news-info/product-news-offers/view/story/656771/F-Secure%20Sicherheitsl%C3%B6sungen%20f%C3%BCr%20Smartphones%20und%20Tablets%20nun%20in%20Google%20Play%20und%20im%20Apple%20App%20Store%20erh%C3%A4ltlich, Stand 1.11.2012. 40

[Feierabend und Klingler, 2007] Feierabend, S. und Klingler, W. (2007). Kinder und Medien: Ergebnisse der KIM-Studie 2006. Der Medienumgang Sechs- bis 13-Jähriger nach Sinus-Milieus. *media perspektiven, Heft 10/2007, S. 492-505*. 43

[Feierabend und Klingler, 2009] Feierabend, S. und Klingler, W. (2009). Kinder und Medien: Ergebnisse der KIM-Studie 2006. Der Medienumgang Sechs- bis 13-Jähriger nach Sinus-Milieus. *media perspektiven, Heft 8/2009, S. 398-412*. 43

[Fiat-Group, 2006] Fiat-Group (2006). Fiat 500 feiert im Web Erfolge. http://www.fiatpress.de/press/article/15670, Stand 13.11.2012. 275

[Fiedler und Lindner, 2001] Fiedler, G. und Lindner, R. (2001). Suizidforen im Internet. http://www.suizidprophylaxe.de/Tagungen/suizidforen.pdf, Stand 25.10.2012. 109

[Fischler und Bolles, 1981] Fischler, M. A. und Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395. 232

[forsa, 2011] forsa (2011). Jugendschutzprogramme aus Sicht der Eltern. Unveröffentlichter Ergebnisbericht vom 2. Dezember 2011. 43

[FOSI, 2012] FOSI (2012). About ICRA. http://www.fosi.org/icra/, Stand 20.11.2012. 34

[fragFINN, 2012a] fragFINN (2012a). Pressemitteilung – fragfinn präsentiert app zum sicheren surfen für kinder mit smartphones und tablets. http://www.fragfinn.de/download/presse/121114_Jubilaeum_fragFINN_Pressemappe.pdf, Stand 26.11.2012. 24

[fragFINN, 2012b] fragFINN (2012b). Pressemitteilung: Zum Kindertag starten Videos der Kinderreporter auf fragFINN.de. http://www.fragfinn.de/download/presse/120920_PM_fragFINN_Kinderreporter.pdf, Stand 30.10.2012. 24

[fragFINN, 2012c] fragFINN (2012c). Pädagogen – Informationen für Lehrer und Multiplikatoren. http://www.fragfinn.de/kinderliste/eltern/paedagogen.html, Stand 31.10.2012. 79

[Fraunhofer-Gesellschaft, 2007] Fraunhofer-Gesellschaft (2007). Nichtfunktionale Anforderungen erheben. http://www.re-wissen.de/Wissen/Anforderungserhebung/Praktiken/Nichtfunktionale_Anforderungen_erheben.html, Stand 30.10.2012. 64

[Freund und Schapire, 1995] Freund, Y. und Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceeding EuroCOLT '95 Proceedings of the Second European Conference on Computational Learning Theory*. 144

[Freund und Schapire, 1997] Freund, Y. und Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139. 144

[FSF, 2012a] FSF (2012a). Geschichte der FSF – ein Überblick. http://fsf.de/die-fsf/geschichte/, Stand 16.11.2012. 22, 54

[FSF, 2012b] FSF (2012b). Prüfungsordnung der freiwilligen Selbstkontrolle Fernsehen e.V. (PrO-FSF). http://fsf.de/data/user/Dokumente/Downloads/FSF_PrO.pdf, Stand25.10.2012. 94

[FSK, 2010] FSK (2010). Grundsätze der freiwilligen Selbstkontrolle der Filmwirtschaft GmbH. http://www.fsk.de/media_content/422.pdf, Stand 28.9.2012. 8, 20, 93

[FSK, 2011a] FSK (2011a). Altersfreigaben im internationalen Vergleich. http://www.spio.de/media_content/1392.pdf, Stand 6.9.2012. ix, 61

[FSK, 2011b] FSK (2011b). Übersicht der FSK-Prüfverfahren für filmische Inhalte, die in Deutschland zur öffentlichen Vorführung und Veröffentlichung auf Trägermedien nach dem Jugendschutzgesetz vorgesehen sind. http://www.fsk.de/index.asp?SeitID=505&TID=72, Stand 28.9.2012. 20

[FSK, 2012a] FSK (2012a). Alterseinstufungen und FSK-Kennzeichen. http://www.fsk.de/index.asp?SeitID=508&TID=72, Stand 31.10.2012. 93

[FSK, 2012b] FSK (2012b). Aufgaben und Strukturen. http://www.fsk.de/index.asp?SeitID=504&TID=473, Stand 30.10.2012. 20

[FSK, 2012c] FSK (2012c). Beschwerdestelle. http://www.fsk.de/index.asp?SeitID=1676&TID=466, Stand 30.10.2012. 20

[FSK, 2012d] FSK (2012d). Online-Medien – Die FSK gewährleistet Jugendschutz für Webangebote. http://www.fsk-online.de/index.asp?SeitID=466&TID=466, Stand 30.10.2012. 53

[FSK, 2012e] FSK (2012e). Prüfverfahren und Ausschüsse. http://www.fsk.de/index.asp?SeitID=505&TID=72, Stand 30.10.2012. 20

[FSK, 2012f] FSK (2012f). Serviceleistungen von FSK.online. http://www.fsk.de/index.asp?SeitID=559&TID=466, Stand 30.10.2012. 20

[FSM, 2004a] FSM (2004a). Der Begriff der Eintwicklungsbeeinträchtigung in § 5 des Jugendmedienschutz-Staatsvertrags. Technischer Bericht, Freiwillige Selbstkontrolle Multimedia Diensteanbieter e.V. 7

[FSM, 2004b] FSM (2004b). Verhaltenssubkodex für Suchmaschinenanbieter der FSM (VK-S). http://fsm.de/de/Subkodex_Suchmaschinenanbieter, Stand 9.10.2012. 26

[FSM, 2005] FSM (2005). Verhaltenskodex der Mobilfunkanbieter in Deutschland zum Jugendschutz im Mobilfunk. http://www.fsm.de/de/subkodex_mobilfunk, Stand 9.10.2012. 25

[FSM, 2007a] FSM (2007a). Jugendmedienschutz im Mobilfunk - Selbstverpflichtung der Mobilfunkanbieter. http://fsm.de/inhalt.doc/Selbstverpflichtung_Mobilfunk.pdf, Stand 9.10.2012. 25

[FSM, 2007b] FSM (2007b). Verhaltenssubkodex für Chatanbieter der FSM (VK-C). http://www.fsm.de/inhalt.doc/Verhaltenssubkodex_Chat.pdf, Stand 9.10.2012. 27

[FSM, 2009a] FSM (2009a). Beschwerdeordnung Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V. http://fsm.de/de/Beschwerdeordnung, Stand 28.9.2012. 20

[FSM, 2009b] FSM (2009b). Verhaltenskodex für Betreiber von Social Communities bei der FSM. http://www.fsm.de/inhalt.doc/VK_Social_Networks.pdf, Stand 9.10.2012. 25

[FSM, 2010] FSM (2010). Verhaltenskodex Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V. http://fsm.de/de/Verhaltenskodex, Stand 28.9.2012. 19

[FSM, 2011a] FSM (2011a). FAQ Selbstkontrolle Suchmaschinen – Fragen zum BPjM-Modul. http://fsm.de/de/FAQs_Selbstkontrolle_Suchmaschinen#BPjM-Modul, Stand 30.10.2012. 14

[FSM, 2011b] FSM (2011b). Gesetzliche Bestimmungen. https://www.altersklassifizierung.de/de/faq/gesetzliche-bestimmungen, Stand 30.10.2012. 54

[FSM, 2011c] FSM (2011c). Herzlich Willkommen auf der Website der Freiwilligen Selbstkontrolle Multimedia. http://fsm.de/, Stand 30.10.2012. 19

[FSM, 2011d] FSM (2011d). Prüfgrundsätze der FSM. Technischer Bericht, Freiwillige Selbstkontrolle Multimedia Diensteanbieter e.V. 8, 90, 91, 96, 97, 98, 102, 103, 104, 108, 110, 295

[FSM, 2011e] FSM (2011e). Selbstkontrolle Suchmaschinen. http://www.fsm.de/de/Web_2_0, Stand 30.10.2012. 26

[FSM, 2011f] FSM (2011f). Verhaltenskodex für Betreiber von Social Communities bei der FSM. http://www.fsm.de/de/Web_2_0, Stand 30.10.2012. 26

[Fu et al., 2010] Fu, J., Wang, J., und Lu, H. (2010). Effective logo retrieval with adaptive local feature selection. In *Proceedings of the international conference on Multimedia*, MM '10, pages 971–974. 178

[Fujisawa, 2008] Fujisawa, H. (2008). Forty years of research in character and document recognition-an industrial perspective. *Pattern Recognition*, 41(8):2435–2446. 174

[Fujitsu Services, 2011] Fujitsu Services (2011). The White Book of Cloud Adoption - The definitive guide to a business technology revolution. Technischer Bericht, Fujitsu Services Ltd., Tokio. 286

[Gamestar, 2011] Gamestar (2011). The Death of Osama bin Laden - Videospiel von Kuma Games. http://www.gamestar.de/news/pc/2322824/the_death_of_osama_bin_laden.html, Stand 30.10.2012. 51

[Garfield, 1972] Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178:471–479. 182

[Garfinkel und Rosenberg, 2009] Garfinkel, S. und Rosenberg, B. (2009). Gesichtserkennung: Clever oder unheimlich? *Technology Review*. 80

[Gärtner et al., 2003] Gärtner, T., Flach, P. A., und Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. In *COLT*, pages 129–143. 142

[Giese und Poggio, 2003] Giese, M. A. und Poggio, T. (2003). Neural mechanisms for the recognition of biological movements and action. *Nature Reviews Neuroscience*, 4:179–192. 179

[Golub und Ardö, 2005] Golub, K. und Ardö, A. (2005). Importance of html structural elements and metadata in automated subject classification. In *Proceedings of the 9th European conference on Research and Advanced Technology for Digital Libraries ECDL'05*, pages 368–378, Berlin, Heidelberg. Springer-Verlag. 135

[Google, 2012a] Google (2012a). Sucheinstellungen – SafeSearch-Filter. http://support.google.com/websearch/bin/answer.py?hl=de&answer=35892#safe, Stand 31.10.2012. 85

[Google, 2012b] Google (2012b). Sucheinstellungen – SafeSearch-Filter. https://www.google.de/preferences?hl=de, Stand 31.10.2012. 85

[Gori et al., 2003] Gori, M., Maggini, M., Marinai, S., Sheng, J., und Soda, G. (2003). Edge-backpropagation for noisy logo recognition. *Pattern Recognition*, 36(1):103–110. 176

[Gosling et al., 2004] Gosling, S. D., Vazire, S., und Srivastava, S. (2004). Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. In *American Psychologist*, Volume 59, No. 2, pages 93–103. 280

[Gottron und Lipka, 2010] Gottron, T. und Lipka, N. (2010). A comparison of language identification approaches on short, query-style texts. In *Proc. ECIR*, pages 611–614. 135

[Grega et al., 2011] Grega, M., Glowacz, A., Anzel, W., Lach, S., und Musial, J. (2011). INCR – INDECT Multimedia Crawler. In Dziech, A. und Czyzewski, A., editors, *Multimedia Communications, Services and Security*, volume 149 of *Communications in Computer and Information Science*, pages 46–54. Springer Berlin Heidelberg. 184

[Greif, 2012] Greif, B. (2012). Amazon und Google liefern sich Preiskampf bei Cloud Storage. http://www.zdnet.de/88134069/amazon-und-google-liefern-sich-preiskampf-bei-cloud-storage/, Stand 1.12.2012. 286

[Grundgesetz, 2012] Grundgesetz (2012). Grundgesetz für die Bundesrepublik Deutschland (Ausfertigungsdatum: 23.05.1949, Zuletzt geändert durch Art. 1 G v. 11.7.2012). 4, 302

[Haffner et al., 2006] Haffner, J., Roos, J., Steen, R., Parzer, P., Klett, M., und Resch, F. (2006). Lebenssituation und Verhalten von Jugendlichen — Ergebnisse einer Befragung 14- bis 16-jähriger Jugendlicher und deren Eltern im Jahr 2005. *Gesundheitsbericht Rhein-Neckar-Kreis/Heidelberg*, 3. 109

[Halfacree, 2012] Halfacree, G. (2012). Cameron pushes on porn-filtering programme. http://www.bit-tech.net/news/bits/2012/11/19/cameron-porn/1, Stand 1.12.2012. 274

[Hasebrink et al., 2011] Hasebrink, U., Lampert, C., Schröder, H.-D., und Drosselmeier, M. (2011). Jugendmedienschutz aus Sicht der Eltern – Kurzbericht über eine Studie des Zweiten Deutschen Fernsehens. Technischer Bericht, ZDF, Hans-Bredow-Institut. 42

[Hasebrink et al., 2012] Hasebrink, U., Schröder, H.-D., und Schumacher, G. (2012). Kinder- und Jugendmedienschutz aus der Sicht der Eltern – Ergebnisse einer repräsentativen Elternbefragung. *Media Perspektiven 1/2012*. 42, 43, 45

[Hastie et al., 2011] Hastie, T., Tibshirani, R., und Friedman, J. (2011). *The Elements of Statistical Learning*. Springer. 137, 138, 139, 140, 141, 143, 144

[HBI, 2011] HBI (2011). Jugendschutzsoftware im Elternhaus: Kenntnisse, Erwartungen und Nutzung. Technischer Bericht, Hans-Bredow-Institut (HBI). 9, 31, 43, 80

[Heise Online, 2007] Heise Online (2007). Chat-Anbieter verpflichten sich zu mehr Jugendschutz. http://heise.de/-192369, Stand 30.10.2012. 27

[Hoffmann, 1845] Hoffmann, H. (1845). *Lustige Geschichten und drollige Bilder mit 15 schön kolorierten Tafeln für Kinder von 3–6 Jahren*. Literarische Anstalt, Frankfurt am Main, Deutschland. 199, 200

[Hopf und Braml, 2010] Hopf, K. und Braml, B. (2010). Bewertungsvorgänge im Jugendmedienschutz: Kategorienbildung anstelle differenzierter Jugendschutzprüfung : Paradigmenwechsel oder Feigenblatt? *Zeitschrift für Urheber- und Medienrecht (ZUM)*, 3:211–220. 52, 280

[Howe, 2006] Howe, J. (2006). The Rise of Crowdsourcing. http://www.wired.com/wired/archive/14.06/crowds.html, Stand 13.11.2012. 275

[IGN, 2012] IGN (2012). Current CERO Ratings . http://www.ign.com/wikis/content-ratings/CERO, Stand 20.11.2012. 58

[INNOVA Europe, 2012] INNOVA Europe (2012). SIP-Benchmark II, Ergebnisse des vierten Testzyklus, Sommer 2012. http://www.sipbench.eu/index.cfm/lang.1, Stand 30.10.2012. 37

[internet abc, 2012] internet abc (2012). Über uns. http://www.internet-abc.de/eltern/ueberuns.php, Stand 30.10.2012. 25

[Ipeirotis et al., 2010] Ipeirotis, P. G., Provost, F., und Wang, J. (2010). Quality management on amazon mechanical turk. In *KDD-HCOMP?10, July 25, 2010,*. 281

[itemis, 2012] itemis (2012). Scrum Kompakt – Grundlagen des Projektmanagements – Begriffe. http://www.scrum-kompakt.de/grundlagen-des-projektmanagements/begriffe/, Stand 30.10.2012. 49

[ITwissen.info, 2012a] ITwissen.info (2012a). Informationssicherheit. http://www.itwissen.info/definition/lexikon/Informationssicherheit-information-security.html, Stand 30.11.2012. 75

[ITwissen.info, 2012b] ITwissen.info (2012b). Peer-to-Peer-Netz. http://www.itwissen.info/definition/lexikon/Peer-to-Peer-Netz-P2P-peer-to-peer-network.html, Stand 20.11.2012. 63

[ITwissen.info, 2012c] ITwissen.info (2012c). Proxy-Server. http://www.itwissen.info/definition/lexikon/Proxy-Server-proxy-server.html, Stand 30.11.2012. 297

[ITwissen.info, 2012d] ITwissen.info (2012d). Rechenleistung. http://www.itwissen.info/definition/lexikon/Rechenleistung-Computing-power.html, Stand 21.11.2012. 67

[ITwissen.info, 2012e] ITwissen.info (2012e). Router. http://www.itwissen.info/definition/lexikon/Router-RO-router.html, Stand 30.11.2012. 297

[Jain und Vailaya, 1998] Jain, A. und Vailaya, A. (1998). Shape-based retrieval: A case study with trademark image databases. *Pattern Recognition*, 31:1369–1390. 176

[Jiang et al., 2011] Jiang, Y., Meng, J., und Yuan, J. (2011). Grid-based local feature bundling for efficient object search and localization. In *ICIP*, pages 113–116. 177

[Jiang et al., 2012] Jiang, Y., Meng, J., und Yuan, J. (2012). Randomized visual phrases for object search. In *CVPR*, pages 3100–3107. 176

[JMStV, 2003] JMStV (2003). Gesetz zum Staatsvertrag über den Schutz der Menschenwürde und den Jugendschutz in Rundfunk und Telemedien (Jugendmedienschutz-Staatsvertrag - JMStV), zuletzt geändert durch Artikel 2 des Dreizehnten Rundfunkänderungsstaatsvertrages vom 30. Oktober 2009. 7, 8, 18, 51, 53, 295

[Joachims, 2002] Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*. Kluwer/Springer. 142

[Joerges, 2012] Joerges, R. (2012). BITKOM: Marktanteile der Smartphone-Betriebssysteme in Deutschland. http://enterprisit.wordpress.com/2012/06/06/bitkom-marktanteile-der-smartphone-betriebssysteme-in-deutschland/, Stand 30.10.2012. 40

[Joly und Buisson, 2009] Joly, A. und Buisson, O. (2009). Logo retrieval with a contrario visual query expansion. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 581–584. 176, 178

[jugendschutz.net, 2007] jugendschutz.net (2007). 10 Jahre jugendschutz.net - 10 Jahre Jugendschutz im Internet. http://www.jugendschutz.net/pdf/festschrift.pdf, Stand 27.9.2012. 18

[jugendschutz.net, 2011] jugendschutz.net (2011). Jugendschutz im Internet, Ergebnisse der Recherchen und Kontrollen, Bericht 2011. Technischer Bericht, Jugendschutz.net. i, 42, 50

[jugendschutz.net, 2012] jugendschutz.net (2012). Auftrag von jugendschutz.net. http://www.jugendschutz.net/jugendschutz_net/Auftrag/index.html, Stand 30.10.2012. 17, 18

[JuSchG, 2008] JuSchG (2008). Jugendschutzgesetz vom 23. Juli 2002 (BGBl. I S. 2730), das zuletzt durch Artikel 3 Absatz 1 des Gesetzes vom 31. Oktober 2008 (BGBl. I S. 2149) geändert worden ist. http://www.gesetze-im-internet.de/bundesrecht/juschg/gesamt.pdf, Stand 21.9.2012. 5, 14, 53, 54, 102, 302

[JuSchRiL, 2005] JuSchRiL (2005). Gemeinsame Richtlinien der Landesmedienanstalten zur Gewährleistung des Schutzes der Menschenwürde und des Jugendschutzes (Jugendschutzrichtlinien - JuSchRiL) vom 8./9. März 2005. http://www.kjm-online.de/files/pdf1/JuSchRiL.pdf, Stand 26.9.2012. 9, 10, 52, 296

[Jüttner, 2009] Jüttner, J. (2009). Volksverhetzung vor Gericht – Fräulein Stolz und der Hitlergruß. *Spiegel Online*, 9. Mai 2009. 235

[Kan und Thi, 2005] Kan, M.-Y. und Thi, H. O. N. (2005). Fast webpage classification using url features. In *Proceedings of the 14th ACM international conference on Information and knowledge management CIKM '05*, pages 325–326, New York, NY, USA. ACM. 135

[Kim und Kim, 1998] Kim, Y.-S. und Kim, W.-Y. (1998). Content-based trademark retrieval system using a visually salient feature. *Image Vision Comput.*, 16(12–13):931–939. 176

[Kirkpatrick, 2004] Kirkpatrick, S. (2004). Lazy Guide to Net Culture: NSFW. *The Scotsman*. 297

[KJM, 2004] KJM (2004). Netz-Regeln: 10 Vorschläge für die Beachtung des Jugendschutzes im Internet. http://www.kjm-online.de/files/pdf1/Netz-Regeln1.pdf, Stand 26.9.2012. 10

[KJM, 2005] KJM (2005). Kriterien für die Aufsicht im Rundfunk und in den Telemedien. 98

[KJM, 2010] KJM (2010). Kriterien für die Aufsicht im Rundfunk und in den Telemedien. 53, 89, 96, 97, 100, 102, 103, 104, 108, 109, 110, 295

[KJM, 2011a] KJM (2011a). Informationen für Betreiber und Anbieter von Jugendschutzprogrammen: Kriterien der KJM für die Anerkennung von Jugendschutzprogrammen im Bereich des World Wide Web. http://www.kjm-online.de/files/pdf1/Informationen-fr-JSP-Anbieter_Stand_2011-05-11.pdf, Stand 16.11.2012. 16, 30, 32, 65, 66, 77, 81

[KJM, 2011b] KJM (2011b). Vierter Bericht der KJM über die Durchführung der Bestimmungen des Staatsvertrags über den Schutz der Menschenwürde und den Jugendschutz in Rundfunk und Telemedien (Jugendmedienschutz-Staatsvertrag - JMStV) gem. § 17 Abs. 3 JMStV, Berichtszeitraum: März 2009 bis Februar 2011. 16, 17, 18

[KJM, 2012a] KJM (2012a). Jugendmedienschutz-Staatsvertrag (JMStV). http://kjm-online.de/de/pub/recht/gesetze_und_staatsvertraege/jugendmedienschutz-staatsvertr.cfm, Stand 20.11.2012. 5

[KJM, 2012b] KJM (2012b). Jugendschutzprogramme. http://kjm-online.de/de/pub/jugendschutz_in_telemedien/jugendschutzprogramme.cfm, Stand 30.10.2012. 31

[KJM, 2012c] KJM (2012c). KJM: Auftrag. http://www.kjm-online.de/de/pub/die_kjm/auftrag.cfm, Stand 30.10.2012. 15, 16

[KJM, 2012d] KJM (2012d). KJM erkennt erstmals zwei Jugendschutzprogramme unter Auflagen an. http://www.kjm-online.de/de/pub/aktuelles/pressemitteilungen/pressemitteilungen_2012/pm_032012.cfm, Stand 30.10.2012. 33

[KJM, 2012e] KJM (2012e). KJM: Jugendschutz in den Telemedien – Selbstkontrolle. http://www.kjm-online.de/de/pub/jugendschutz_in_telemedien/selbstkontrolle.cfm, Stand 30.10.2012. 19

[KJM, 2012f] KJM (2012f). Kriterien zur Bewertung von Konzepten für Altersverifikationssysteme als Elemente zur Sicherstellung geschlossener Benutzergruppen in Telemedien nach § 4 Abs. 2 S. 2 JMStV (»AVS-RASTER«). http://www.kjm-online.de/files/pdf1/KJM-Kriterienraster_Stand_11.09.20121.pdf, Stand 9.10.2012. 29, 50

[KJM, 2012g] KJM (2012g). Technische Mittel. http://kjm-online.de/de/pub/jugendschutz_in_telemedien/technische_mittel.cfm, Stand 31.10.2012. 29

[Kleinberg, 1999] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632. 182

[klick-tipps.net, 2012] klick-tipps.net (2012). Frequently Asked Questions klick-tipps.net. http://www.klick-tipps.net/fileadmin/downloads/pdf/2012-02-01-FAQ.pdf, Stand 25.10.2012. 24

[klicksafe, 2012] klicksafe (2012). Wer ist klicksafe? http://www.klicksafe.de/ueber-klicksafe/die-initiative/projektinfo/wer-ist-klicksafe/, Stand 30.10.2012. 23

[Kraaij und Awad, 2011] Kraaij, W. und Awad, G. (2011). 2011 TRECVID Workshop – Content Based Copy Detection. In *TRECVID 2011 Folien*. 165

[Krcmar, 2012] Krcmar, H. (2012). Enzyklopädie der Wirtschaftsinformatik, Eintrag Softwaremarkt. http://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/uebergreifendes/Kontext-und-Grundlagen/Markt/Softwaremarkt/index.html/?searchterm=app, Stand 30.11.2012. 293

[Krüger und von Gottberg, 2012] Krüger, T. und von Gottberg, P. J. (2012). Die eigentliche Problematik liegt im Internet - Ordnungspolitischer Jugendschutz und Prävention müssen besser abgestimmt werden - Interview. *tv diskurs. Verantwortung in audiovisuellen Medien 16. Jg., 3/2012 (Ausgabe 61), S. 96-101*. 28

[Kuma Games, 2011] Kuma Games (2011). ›WAR‹ IS OVER! 106 MISSIONS LATER, GAMERS TAKE DOWN BIN LADEN IN FINAL EPISODE OF KUMA WAR II. http://www.kumagames.com/osama_2011.html, Stand 30.10.2012. 51

[Kumar und Govindarajulu, 2009] Kumar, J. P. und Govindarajulu, P. (2009). Duplicate and near duplicate documents detection: A review. *European Journal of Scientific Research*, pages 514–527. 283

[Lackes und Siepermann, 2012a] Lackes, R. und Siepermann, M. (2012a). Gabler Wirtschaftslexikon, Stichwort: HTML. http://wirtschaftslexikon.gabler.de/Archiv/75639/html-v6.html, Stand 10.12.2012. 295

[Lackes und Siepermann, 2012b] Lackes, R. und Siepermann, M. (2012b). Gabler Wirtschaftslexikon, Stichwort: Smart Devices. http://wirtschaftslexikon.gabler.de/Archiv/76910/smart-devices-v5.html, Stand 10.12.2012. 68

[Lafferty et al., 2001] Lafferty, J., McCallum, A., und Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, pages 282–289. 147

[Laws et al., 2011] Laws, F., Scheible, C., und Schütze, H. (2011). Active learning with amazon mechanical turk. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 281

[Lehner, 2012] Lehner, F. (2012). Enzyklopädie der Wirtschaftsinformatik, Eintrag Software-Wartung. http://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/is-management/Systementwicklung/Hauptaktivitaten-der-Systementwicklung/Software-Wartung, Stand 23.11.2012. 71

[LfM, 2012] LfM (2012). Das Projekt »Medienscouts NRW«. http://www.medienscouts-nrw.de/index.php?q=node/187, Stand 30.10.2012. 45

[LfM, 2012] LfM (2012). Zwischen Selbstoffenbarung und Privatheit: Wie schützen junge Menschen ihre Daten? – Landesanstalt für Medien NRW präsentiert neue Studie zum Datenschutzverhalten junger Menschen in sozialen Netzwerken. http://www.lfm-nrw.de/aktuell/pressemitteilungen/pressemitteilungen-detail/article/zwischen-selbstoffenbarung-und-privatheit-wie-schuetzen-junge-menschen-ihre-daten.html, Stand 30.10.2012. 80

[Liesching, 2011] Liesching, M. (2011). *Schutzgrade im Jugendmedienschutz; Begriffsbestimmungen, Auslegungen, Rechtsfolgen*. Nomos Verlagsgesellschaft, Baden-Baden. 28, 89, 92, 98, 104, 108, 109, 110

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110. 163

[Manning und Schütze, 2001] Manning, C. D. und Schütze, H. (2001). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA. 145

[Mehtre et al., 1997] Mehtre, B., Kankanhalli, M., und Lee, W. (1997). Shape measures for content based image retrieval: a comparison. *Information Processing and Management*, 33(3):319–337. 176

[Meiritz, 2009] Meiritz, A. (2009). Urteil wegen Volksverhetzung – Horst Mahler pöbelt im Gerichtssaal. *Spiegel Online*, 11. März 2009:1. 235

[Microsoft, 2011] Microsoft (2011). Nokia and Microsoft Announce Plans for a Broad Strategic Partnership to Build a New Global Mobile Ecosystem. http://www.microsoft.com/en-us/news/press/2011/feb11/02-11partnership.aspx, Stand 1.11.2012. 40

[Microsoft, 2012a] Microsoft (2012a). Allgemeine Einstellungen – SafeSearch. http://www.bing.com/preferences.aspx, Stand 31.10.2012. 85

[Microsoft, 2012b] Microsoft (2012b). Bereitstellen von Suchergebnissen in Bing, Abschnitt »So begegnen wir Problemen bezüglich des Zugriffs auf nicht jugendfreie Inhalte«. http://onlinehelp.microsoft.com/de-de/bing/ff808447.aspx, Stand 31.10.2012. 85

[Mitchell, 1997] Mitchell, T. (1997). *Machine Learning*. McGraw-Hill. 140

[MPFS, 2007] MPFS (2007). KIM-Studie 2006. Kinder + Medien, Computer + Internet. Basisuntersuchung zum Medienumgang 6- bis 13-Jähriger. Technischer Bericht, Medienpädagogischer Forschungsverbund Südwest (MPFS). 43

[MPFS, 2009] MPFS (2009). KIM-Studie 2008. Kinder + Medien, Computer + Internet. Basisuntersuchung zum Medienumgang 6- bis 13-Jähriger. Technischer Bericht, Medienpädagogischer Forschungsverbund Südwest (MPFS). 43

[MPFS, 2011a] MPFS (2011a). JIM-Studie 2011. Jugend, Information, (Multi-)Media. Basisuntersuchung zum Medienumgang 12- bis 19-Jähriger. Technischer Bericht, Medienpädagogischer Forschungsverbund Südwest (MPFS). 43, 73

[MPFS, 2011b] MPFS (2011b). KIM-Studie 2010. Kinder + Medien, Computer + Internet. Basisuntersuchung zum Medienumgang 6- bis 13-Jähriger. Technischer Bericht, Medienpädagogischer Forschungsverbund Südwest (MPFS). 43

[Müller und Clausen, 2007] Müller, M. und Clausen, M. (2007). Transposition-invariant self-similarity matrices. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 47–50, Vienna, Austria. 229

[Müller et al., 2005] Müller, M., Kurth, F., und Clausen, M. (2005). Audio matching via chroma-based statistical features. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 288–295. 229

[Netcraft Ltd., 2012] Netcraft Ltd. (2012). October 2012 Web Server Survey. http://news.netcraft.com/archives/2012/10/02/october-2012-web-server-survey.html, Stand 13.11.2012. 274

[Netzwelt, 2012] Netzwelt (2012). Samsung ST80 im Test: WLAN-Kompaktkamera. http://www.netzwelt.de/news/84813-samsung-st80-test-wlan-kompaktkamera.html, Stand 30.10.2012. 68

[Nikles, 2011] Nikles, B. W. (2011). Kinder- und Jugendschutz; Online-Handbuch; Oberste Landesjugendbehörde. http://www.handbuch-jugendschutz.de/index.php?postid=153&search=&abc=O, Stand 30.10.2012. 14

[NIST, 2009] NIST (2009). The History of Automatic Speech Recognition Evaluations at NIST – NIST SST Benchmark Test History – May '09. http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html, Stand 19.12.2012. 169

[Nocedal und Wright, 2006] Nocedal, J. und Wright, S. J. (2006). *Numerical Optimization*. Springer, Berlin, New York, 2nd edition. 138

[OFLC, 2012] OFLC (2012). Classification categories explained. http://www.classification.gov.au/Pages/Guidelines/Guidelines.aspx, Stand 20.11.2012. 59

[Olsson, 2009] Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing. Technischer Bericht T2009:06, Swedish Institute of Computer Science, Box 1263, SE-164 29 Kista, Sweden. 153

[OMK, 2012] OMK (2012). Definition des technischen Labels zur Altersklassifizierung in Telemedien. http://online-management-kontor.de/jugendschutz/altersklassifizierung.html, Stand 30.10.2012. 19, 31

[Over et al., 2011] Over, P., Awad, G., Fiscus, J., F. Smeaton, A., Kraaij, W., und Quénot, G. (2011). TRECVID 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2011*. NIST, USA. 163

[Pantola et al., 2011] Pantola, A. V., Pancho-Festin, S., und Salvador, F. (2011). TULUNGAN: A consensus-independent reputation system for collaboratve web filtering systems. *Science Diliman, A Journal of Pure and Applied Science*, 23. 281, 282

[PEGI, 2011] PEGI (2011). PEGI Annual Report 2011. Technischer Bericht, PEGI S.A., Belgien. 87, 93

[PEGI, 2012] PEGI (2012). Über PEGI – Was bedeuten die Kennzeichnungen? http://www.pegi.info/de/index/id/54/, Stand 20.11.2012. 55

[Pelzer et al., 2012] Pelzer, C., Wenzlaff, K., und Eisfeld-Reschke, J. (2012). Crowdsourcing Report 2012 - Neue Digitale Arbeitswelten. Technischer Bericht, Deutscher Crowdsourcing Verband e.V. 275, 276

[Pohl, 2010] Pohl, K. (2010). *Requirements Engineering: Fundamentals, Principles and Techniques*. Springer. 47

[Porter, 1980] Porter, M. (1980). An algorithm for suffix stripping. *Program*, pages 130–137. 136

[Projecter, 2012] Projecter (2012). Zahlen und Statistiken zur Internetnutzung 2012. http://www.projecter.de/blog/allgemein/zahlen-und-statistiken-zur-internetnutzung-2012.html, Stand 30.10.2012. 68

[Ptaszynski et al., 2010] Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., und Araki, K. (2010). Machine learning and affect analysis against cyber-bullying. In *Proc. of the Linguistic And Cognitive Approaches To Dialog Agents Symposium at the AISB 2010 convention,*. 152

[Puresight Technologies, 2012] Puresight Technologies (2012). Advanced Technology. http://www.puresight.com/Why-PureSight-product/advanced-technology.html, Stand 30.10.2012. 38

[Qi und Davison, 2009] Qi, X. und Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM Computing Surveys (CSUR)*, 41:12:1–12:31. 135, 182, 184

[Ramage et al., 2009] Ramage, D., Hall, D., Nallapati, R., und Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. EMNLP*, pages 248–256. 147

[Rauchfuß, 2008] Rauchfuß, K. (2008). Abschlussbericht der Recherche zu Pro-Anorexie-Angeboten 2006/2007. Technischer Bericht, jugendschutz.net, Mainz, Deutschland. 89, 109

[Reichartz et al., 2010] Reichartz, F., Korte, H., und Paass, G. (2010). Semantic relation extraction with kernels over typed dependency trees. In *Proc. ACM SIGKDD*, pages 773–782. 142

[Research, 2011] Research, G. (2011). Gartner Says Android to Command Nearly Half of Worldwide Smartphone Operating System Market by Year-End 2012. http://www.gartner.com/it/page.jsp?id=1622614, Stand 31.10.2012. 40

[Rogers, 2003] Rogers, P. (2003). Streisand's home becomes hit on Web – star had sued valley mogul to keep photograph off site. *Mercury News*. 298

[Schaich et al., 2012] Schaich, E., Schmidt, K., Feess, E., Lackes, R., Siepermann, M., Steven, M., und Thommen, J.-P. (2012). Gabler Wirtschaftslexikon, Stichwort: Effizienz. http://wirtschaftslexikon.gabler.de/Archiv/7640/effizienz-v10.html, Stand 10.12.2012. 294

[SCHAU HIN!, 2012] SCHAU HIN! (2012). SCHAU HIN! bietet Eltern und Erziehenden Orientierung. http://schau-hin.info/schau-hin/selbstverstaendnis.html, Stand 30.10.2012. 22

[Schellenberg, 2011] Schellenberg (2011). Label-Formate für die Altersklassifizierung in Telemedien age-de.xml. http://online-management-kontor.de/downloads/age-de-xml-Label_Definition_v3.0g.pdf, Stand 27.9.2012. 19, 20

[Schenk et al., 2012] Schenk, M., Niemann, J., Reinmann, G., Roßnagel, A., Jandt, S., und Schnurr, J.-M. (2012). Digitale Privatsphäre: Heranwachsende und Datenschutz auf Sozialen Netzwerkplattformen. *Schriftenreihe Medienforschung der LfM, Band 71*. 80

[Schleier und Schreiber, 2012] Schleier, S. und Schreiber, T. (2012). Web Application Security Untersuchung - Aktuelle Verbreitung von HTTP Strict Transport Security (HSTS). Technischer Bericht, SecureNet GmbH. 270

[Schneider et al., 2008] Schneider, D., Schon, J., und Eickeler, S. (2008). Towards Large Scale Vocabulary Independent Spoken Term Detection: Advances in the Fraunhofer IAIS Audiomining System. In *Proc. SIGIR*, Singapore. 235

[Schulz, 2004] Schulz, D. W. (2004). Basiswissen: Jugendschutz. http://www.hans-bredow-institut.de/ws-lehr/lehre/sose2004/wuv/bw_jugendschutz.pdf, Stand 21.09.2012. 4

[Sculley und Wachman, 2007] Sculley, D. und Wachman, G. M. (2007). Relaxed online svms in the trec spam filtering track. In Voorhees, E. M. und Buckland, L. P., editors, *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007*, Gaithersburg, Maryland, USA. 154

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47. 128, 143, 151

[secure.me, 2012] secure.me (2012). Schutz für Ihre Identität im Internet. Rund um die Uhr. https://www.secure.me/de/about/, Stand 31.10.2012. 85

[Seitenstark, 2012] Seitenstark (2012). Seitenstark – das kompetente Netzwerk für Kinder im Internet. http://seitenstark.de/erwachsene/das-netzwerk, Stand 30.10.2012. 24

[Serrà et al., 2008] Serrà, J., Gómez, E., Herrera, P., und Serra, X. (2008). Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech and Language Processing*, 16:1138–1151. 229

[Serrà et al., 2009] Serrà, J., Serra, X., und Andrzejak, R. G. (2009). Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017. 229

[ServiceBureau Jugendinformation, oJ] ServiceBureau Jugendinformation (o.J.). Cyber-bullying! http://jugendinfo.de/toleranz/admin/attachviewer.php?typ=Thema&dateiorig=CyberBullying.pdf&dateiverzeichnis=39001&dateiname=b0420e414787e9141611bc3f08d64a13, Stand 29.10.2012. 294

[Siepermann, 2012] Siepermann, M. (2012). Gabler Wirtschaftslexikon, Stichwort: Portabilität. http://wirtschaftslexikon.gabler.de/Archiv/55014/portabilitaet-v5.html, Stand 10.12.2012. 72

[Sivic und Zisserman, 2009] Sivic, J. und Zisserman, A. (2009). Efficient visual search of videos cast as text retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):591–606. 163

[Soffer und Samet, 1998] Soffer, A. und Samet, H. (1998). Using negative shape features for logo similarity matching. In *In Proceedings of the 14th International Conference on Pattern Recognition*, pages 571–573. 176

[Spath et al., 2012] Spath, D., Weiner, N., Renner, T., und Weisbecker, A. (2012). Neue Geschäftsmodelle für die Cloud entwickeln - Methoden, Modelle und Erfahrungen für Software-As-A-Service im Unternehmen. Technischer Bericht, Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO, Stuttgart. 286

[SPIO, 2007] SPIO (2007). Statut der Juristenkommission (JK). http://www.spio.de/media_content/281.pdf, Stand 20.11.2012. 20

[Statista, 2012a] Statista (2012a). Marktanteile der führenden Betriebssystemversionen weltweit von Januar 2009 bis September 2012. http://de.statista.com/statistik/daten/studie/157902/umfrage/marktanteil-der-genutzten-betriebssysteme-weltweit-seit-2009/, Stand 30.10.2012. 68

[Statista, 2012b] Statista (2012b). Prognose der globalen Marktanteile der Smartphone-Betriebssysteme 2012 und 2016. http://de.statista.com/statistik/daten/studie/232378/umfrage/prognose-marktanteile-smartphone-betriebssysteme/, Stand 30.10.2012. 40

[StGB, 2012] StGB (2012). Strafgesetzbuch in der Fassung der Bekanntmachung vom 13. November 1998 (BGBl. I S. 3322), das zuletzt durch Artikel 1 des Gesetzes vom 25. Juni 2012 (BGBl. I S. 1374) geändert worden ist. http://www.gesetze-im-internet.de/bundesrecht/stgb/gesamt.pdf, Stand 21.9.2012. 4, 303

[Theobald et al., 2008] Theobald, M., Siddharth, J., und Paepcke, A. (2008). Spotsigs: Robust and efficient near duplicate detection in large web collections. In *Proc. SIGIR 2008*. 283

[Thommen et al., 2012] Thommen, J.-P., Amelung, V. E., Mühlbacher, A., und Krauth, C. (2012). Gabler Wirtschaftslexikon, Stichwort: Effektivität. http://wirtschaftslexikon.gabler.de/Archiv/7639/effektivitaet-v9.html, Stand 10.12.2012. 294

[Thongsuk et al., 2011] Thongsuk, C., Haruechaiyasak, C., und Saelee, S. (2011). Multi-classification of business types on twitter based on topic model. In *Proc. 8th Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand - Conference 2011*. 147

[Ting, 1998] Ting, A. (1998). *Linear Document Processing*. PhD thesis, Nanyang Technological University, School of Applied Science. 176

[Troncy et al., 2011] Troncy, R., Huet, B., und Schenk, S. (2011). *Multimedia Semantics: Metadata, Analysis and Interaction*. John Wiley & Sons. 128

[USK, 2011a] USK (2011a). Grundsätze der Unterhaltungssoftware Selbstkontrolle. Technischer Bericht, Freiwillige Selbstkontrolle Unterhaltungssoftware GmbH, Berlin. 21

[USK, 2011b] USK (2011b). Kinder und Jugendliche schützen. Alterskennzeichen für Computer- und Videospiele in Deutschland. http://www.usk.de/fileadmin/documents/USK_Broschuere_Dt.pdf, Stand 28.9.2012. 21

[USK, 2011c] USK (2011c). Statut der Unterhaltungssoftware Selbstkontrolle für den Bereich der Telemedien (Statut USK.online). http://www.usk.de/fileadmin/documents/Publisher_Bereich/Statut_USK.online.pdf, Stand 27.9.2012. 21

[USK, 2012a] USK (2012a). Herzlich willkommen bei der USK. http://www.usk.de/, Stand 30.10.2012. 20

[USK, 2012b] USK (2012b). Online-Spiele (JMStV). http://www.usk.de/extramenue/login/publisher/angebot/online-spiele/, Stand 30.10.2012. 21

[Vapnik, 1995] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York. 142

[Vereinte Nationen, 1989] Vereinte Nationen (1989). UN-Kinderrechtskonvention – Übereinkommen über die Rechte des Kindes, vom 20. November 1989, am 26. Januar 1990 von der Bundesrepublik Deutschland unterzeichnet (Zustimmung von Bundestag und Bundesrat durch Gesetz vom 17. Februar 1992 – BGBl. II S. 121), am 6. März 1992 Hinterlegung der Ratifikationsurkunde beim Generalsekretär der Vereinten Nationen am 5. April 1992, für Deutschland in Kraft getreten (Bekanntmachung vom 10. Juli 1992 – BGBl. II S. 990). http://www.unicef.de/fileadmin/content_media/Aktionen/Kinderrechte18/UN-Kinderrechtskonvention.pdf, Stand 7.11.2012. 11, 80

[Viola und Jones, 2001] Viola, P. und Jones, M. (2001). Robust real-time object detection. *International Journal of Computer Vision*. 232

[W3C, 2010] W3C (2010). Web Security Context: User Interface Guidelines. http://www.w3.org/TR/wsc-ui/, Stand 30.11.2012. 271

[Wang et al., 2012] Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H., und Zhao, B. Y. (2012). Serf and turf: Crowdturfing for fun and profit. In *Proceedings of the International World Wide Web Conference(WWW 2012)*. 281, 294

[WDR, 2012] WDR (2012). Rheinkultur-Festival Bonn 2011 – Das größte deutsche Festival-Panorama. http://rheinkulturpanorama.de/, Stand 31.10.2012. 81

[WEB.DE, 2012] WEB.DE (2012). secure.me: Allgemeine Fragen. http://hilfe.freemail.web.de/faq/secureme/allgemein/, Stand 31.10.2012. 85

[Wei et al., 2009] Wei, C.-H., Li, Y., Chau, W.-Y., und Li, C.-T. (2009). Trademark image retrieval using synthetic features for describing global shape and interior structure. *Pattern Recognition*, 42(3):386–394. 176

[Weigand und Braml, 2012] Weigand, V. und Braml, B. (2012). Anerkannte Jugendschutzprogramme aus Sicht der Praxis. *BPJM-Aktuell 3/2012, S.19-22*. 9, 29

[Wendland, 2012] Wendland, C. W. (2012). Die Indizierung von Internetangeboten und das BPjM-Modul. *BPJM-Aktuell 3/2012, S. 12-19*. 14

[Yahoo!, 2012] Yahoo! (2012). Using Yahoo! SafeSearch. http://help.yahoo.com/kb/index?locale=en_US&page=content&y=PROD_SRCH&id=SLN2247, Stand 31.10.2012. 85

[Zauner, 2010] Zauner, C. (2010). Implementation and benchmarking of perceptual image hash functions. Master's thesis, Upper Austria University of Applied Sciences, Hagenberg Campus, Austria. 228, 238, 247

[ZBFS, 2012] ZBFS (2012). Jugendschutz – Jugendmedienschutz. http://www.blja.bayern.de/themen/jugendschutz/jugendmedienschutz/index.html, Stand 30.10.2012. 14

[Zhu und Doermann, 2009] Zhu, G. und Doermann, D. (2009). Logo matching for document image retrieval. In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ICDAR '09, pages 606–610, Washington, DC, USA. IEEE Computer Society. 176