Fraunhofer SIT

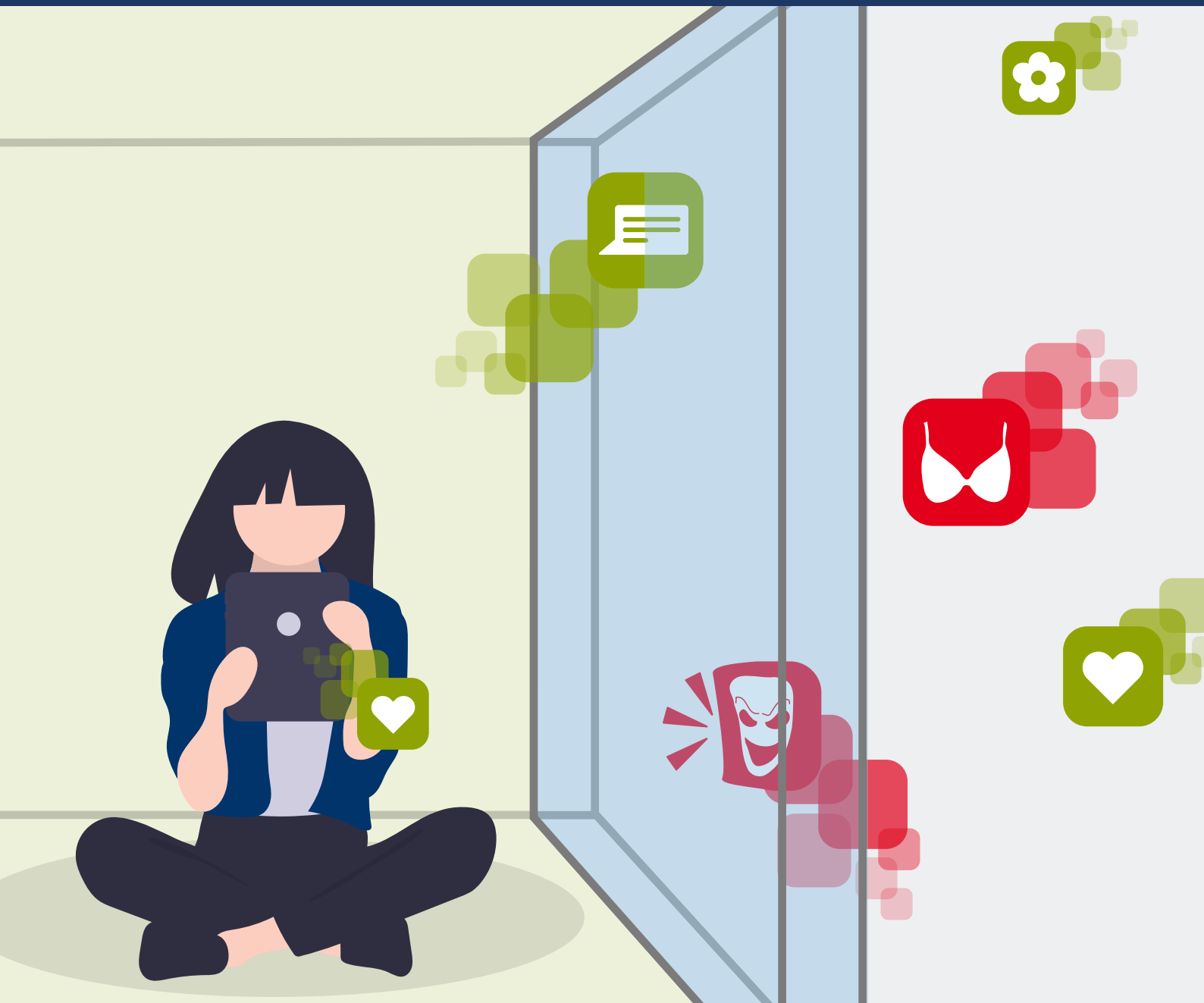
ATHENE
Nationales Forschungszentrum für angewandte Cybersicherheit

FORSCHUNGSFÖRDERUNG CYBERSICHERHEIT HESSEN

Hessisches Ministerium des Innern und für Sport

STUDIE DES VERBUNDPROJEKTS »CYBERSICHERHEITSFORSCHUNG FÜR DIE DIGITALISIERUNG IN VERWALTUNG UND GESELLSCHAFT« DES HESSISCHEN MINISTERIUMS DES INNERN UND FÜR SPORT

TECHNIK FÜR DEN DIGITALEN JUGENDSCHUTZ: AUTOMATISCHE ERKENNUNG VON SEXTING UND CYBERGROOMING



Hinweise

Diese Studie ist bereits 2018 fertiggestellt worden, eine Veröffentlichung war damals nicht vorgesehen. Allerdings sind die Ergebnisse vor dem Hintergrund des Digitalisierungsschubs während der Coronakrise gesellschaftlich so relevant, dass sich das Institut für eine Veröffentlichung im Nachhinein entschieden hat. Es wurden keine Aktualisierungen vorgenommen.

Alle Bezeichnungen für Personen, die in dieser Studie genannt werden, gelten sowohl für das männliche als auch das weibliche Geschlecht. Der Einfachheit halber wird durchgehend das generische Maskulinum verwendet.

Impressum

Layout und Satz
Paula Behnke

Kontakt
Fraunhofer-Institut für Sichere Informationstechnologie SIT
Rheinstraße 75
64295, Darmstadt

TECHNIK FÜR DEN DIGITALEN JUGENDSCHUTZ: AUTOMATISCHE ERKENNUNG VON SEXTING UND CYBERGROOMING

STUDIE

Verbundprojekt »Cybersicherheitsforschung für die Digitalisierung in Verwaltung und Gesellschaft« des Hessischen Ministeriums des Innern und für Sport

Inna Vogel
Martin Steinebach

26. September 2021

INHALTSVERZEICHNIS

VORWORT

Peter Beuth

Digitaler Jugendschutz ist essenziell für eine sichere und verantwortungsbewusste Teilhabe von Minderjährigen an der digitalen Welt. Kinder und Jugendliche haben keine Berührungsängste mit diesen Medien, sie spielen vielmehr eine wichtige Rolle in ihrer Entwicklung – besonders die Corona-Pandemie hat diesen Trend noch einmal beschleunigt. Die Kommunikation über digitale Kanäle hat viele analoge Treffen ersetzt, die teils nicht oder nur eingeschränkt möglich waren. Doch viele Minderjährige sind sich oft nicht bewusst, welche Fallstricke und Gefahren mit digitalen Medien einhergehen können: im Rahmen von „Sexting", also der privaten Kommunikation über sexuelle Themen, werden bspw. eigene intime Bilder versandt, was zur unkontrollierten Verbreitung dieses Bildmaterials führen kann. Häufige Folgen sind Mobbing, soziale Ausgrenzung und damit verbundener psychischer Stress, der vielfach mehrere Jahre anhält und auch gravierende Spätfolgen auslösen kann. Eine weitere Gefahr ist „Cybergrooming" – die digitale Kontaktaufnahme Erwachsener mit Minderjährigen, mit dem Ziel, diese zu missbrauchen.

Um Kindern und Jugendlichen eine möglichst gefahrenfreie Nutzung von digitalen Medien zu ermöglichen, ist Jugendschutz auch im Digitalen mit innovativer Cybertechnik unabdingbar. Das Hessische Ministerium des Innern und für Sport fördert seit 2016 anwendungsorientierte Forschung im Bereich der Cybersicherheit. Aus einem solchen Fördercluster für Cybersicherheitsforschung legt die vorliegende Machbarkeitsstudie mit dem Titel „Technik für den digitalen Jugendschutz: Automatische Erkennung von Sexting und Cybergrooming" ein nachhaltiges Ergebnis vor.

Die Studie stellt unterschiedliche technische Ansätze vor, mit denen sich der Jugendschutz in der digitalen Welt verbessern lässt. Sie skizziert die Herausforderungen, die mit einer Erkennung von „Sexting" und „Cybergrooming" einhergehen, und beschreibt dazu passende prototypische Lösungen. So kann beispielsweise mithilfe automatischer Bilderkennung verhindert werden, dass Apps auf intime Bilder zugreifen. Verdachtsfälle für „Cybergrooming" in Online-Foren lassen sich über technische Verfahren aus der Textforensik automatisiert erkennen. Ein solches System kann dabei unterstützen, potenzielle Täterinnen und Täter schnell zu erkennen, bevor es zum Missbrauch kommt.

Das Land Hessen ist ein Pionier in der anwendungsorientierten Förderung von Cybersicherheitsforschung – und dies nicht nur zum digitalen Jugendschutz. Das Hessische Ministerium des Innern und für Sport arbeitet an der Umsetzung innovativer Technologien und Verfahren, die auch in anderen Bereichen in die Anwendung kommen.

Peter Beuth
Hessischer Minister des Innern und für Sport

1 Einführung

Online-Foren, Chat-Portale und Messenger-Apps eröffnen neue Kommunikationsmöglichkeiten, bieten neue Geschäftsmodelle und erlauben viele Freiheitsgrade zur Selbstdarstellung. Problematisch werden die neuen Möglichkeiten dann, wenn damit Persönlichkeitsrechte verletzt werden, Inhalte (wie z. B. Fotos) ohne Einverständnis weitergegeben werden und wenn die neuen Kommunikationsmöglichkeiten für die Vorbereitung oder Anbahnung von Verbrechen genutzt werden – insbesondere bei Kindern und Jugendlichen als Opfer. Mit dem Internet kommt eine neue Qualität hinsichtlich Reichweite und Dauerhaftigkeit hinzu. Texte und Bilder sind, sobald im Umlauf, kaum noch zu stoppen.

Die vorliegende Studie betrachtet und bewertet technische Möglichkeiten, mit denen vor dem unachtsamen Verschicken von Nacktfotos durch Minderjährige (sogenanntes »Sexting«) gewarnt werden kann und mit denen erwachsene Personen erkannt werden können, die sich in Online-Foren oder Chat-Portalen als Minderjährige ausgeben.

Es hat sich gezeigt, dass hinsichtlich der Erkennung von Nacktfotos maschinelle Lernverfahren eine deutliche Verbesserung der technischen Machbarkeit herbeigeführt haben. Dementsprechend liegt der Fokus der Studie auch auf den Möglichkeiten des maschinellen Lernens, es werden aber auch die Defizite früherer Ansätze aufgezeigt. Im Rahmen der Studie zeigten sich auch weitere Nutzungsmöglichkeiten der Erkenntnisse für die Arbeit von Ermittlungsbehörden zur Kinderpornografie. Diese wurden bereits öffentlich auf Fachtagungen vorgestellt; ein entsprechendes Kapitel fasst das Potenzial zusammen. Eine Besonderheit ist, dass sich die Studie auf eine Erkennung von relevantem, vorher aber unbekanntem Material konzentriert. Aktuelle Methoden basieren auf White- und besonders Blacklisting bekannter Bilder mittels kryptographischer und robuster Hash-Verfahren.

Weiterhin wurden Methoden des Autorenprofilings und der Autorschaftsverifikation untersucht. Autorenprofiling kann dazu dienen, erwachsene Personen, die sich in Online-Foren oder Chat-Portalen als Minderjährige ausgeben (um an entsprechendes Bildmaterial zu kommen oder Treffen zu vereinbaren), hinsichtlich ihres Alters auf Basis von Sprachmustern zu identifizieren. Mithilfe der Autorschaftsverifikation können wiederum pädophile Wiederholungstäter, die sich hinter Pseudonymen tarnen, demaskiert werden, vorausgesetzt, dass von ihnen Referenztexte wie z. B. Chat-Verläufe zur Verfügung stehen.

1.1 Aufbau

In den beiden folgenden Kapiteln werden die Phänomene »Sexting« und »Cybergrooming« diskutiert. Dabei werden nicht nur technische, sondern auch rechtliche und soziologische Aspekte betrachtet.

Im Kapitel »Technische Grundlagen« beschreiben wir technologische Aspekte und Methoden, die als Grundbausteine in den darauf folgenden Kapiteln verwendet werden. So erfolgt eine kompakte Einführung in das maschinelle Lernen, aber auch Grundlagen der Bildverarbeitung werden erörtert.

Im »Stand der Technik« werden verschiedene Verfahren kritisch diskutiert, die heute im Umfeld der Erkennung von Sexting und Cybergrooming bekannt sind. Dabei liegt der Schwerpunkt auf Verfahren, die aus dem maschinellen Lernen stammen. Aber auch ältere Ansätze wie Hautpixelerkennung werden betrachtet.

Die Kapitel »Eignungsprüfung« und »Umsetzbarkeit« eruieren für die Themen Sexting und Cybergrooming jeweils, welche Methoden zur Erkennung geeignet sind und inwiefern eine konkrete Nutzung dieser Methoden auch tatsächlich realisierbar ist. Technologischer Kern sind dabei Verfahren des maschinellen Lernens.

Die »Handlungsempfehlungen« fassen die gewonnenen Erkenntnisse zusammen und liefern Impulse, wie Jugendschutz mit neuen technischen Möglichkeiten verbessert werden kann.

Am Ende steht ein Exkurs zum Thema »Kinderpornografie«. Im Rahmen der Studie wurde deutlich, dass die Erkennung vom Sexting Minderjähriger eng mit der Erkennung von Kinderpornografie verbunden ist. Wir fassen hier Impulse zusammen, die die Sichtung von Bildmaterial bei der Suche nach Kinderpornografie unterstützen können.

2 Sexting

Der Begriff Sexting stammt ursprünglich aus dem Englischen und ist eine Wortneuschöpfung aus den Begriffen »sex« und »texting«. Im englischen Sprachgebrauch bezeichnet er jegliche Form der erotischen Kommunikation über digitale Medien, z. B. den Versand erotischer oder pornografischer Bilder, aber auch den Austausch sexuell anregender Textnachrichten. Im deutschen Sprachgebrauch hingegen bezeichnet Sexting laut Döring lediglich den »private[n] Austausch selbst produzierter erotischer Fotos per Handy oder Internet« [1]. Von dieser Definition wird in der vorliegenden Studie ebenfalls ausgegangen. Ein weiteres Merkmal des Sextings ist nach Döring [2], dass es typischerweise einvernehmlich, privat und bidirektional zwischen zwei Personen stattfindet. Inhaltlich reicht der Begriff Sexting von aufreizenden Bildern in Unterwäsche bis hin zur Darstellung sexueller Praktiken. Ein im Rahmen des Sextings erstelltes Bild bezeichnet man als »Sext«, während Sexting betreibende Personen »Sexter« genannt werden.

2.1 Gesellschaftliche Funktion

Sexting hat nach Döring verschiedene Funktionen für die Beteiligten. Neben der Beziehungspflege innerhalb einer Partnerschaft sowie der Anbahnung einer Beziehung können dies auch unverbindliche Flirts mit Online-Bekanntschaften oder der Austausch im Freundeskreis sein [1]. Zum größten Teil wird Sexting innerhalb einer Paarbeziehung betrieben [1, 2, 3]. Dabei bitten männliche Sexter ihre weiblichen Partnerinnen deutlich häufiger um das Zusenden eines Sexts als umgekehrt. Jedoch verläuft das Sexting in den meisten Fällen bidirektional, d. h. das Versenden der Bilder wird von beiden Kommunikationspartnern vollzogen [1].

2.2 Sexting unter Minderjährigen

In einigen Studien [1, 4, 5, 2] untersuchte Döring das Sexting-Verhalten von Jugendlichen. Der Anteil der aktiven Sexter unter Jugendlichen wird laut Döring maßgeblich durch das Alter bestimmt. Während Sexting im präpubertären Entwicklungsstadium noch kaum eine Rolle spielt, steigt die Teilnahme mit zunehmendem Alter, insbesondere ab der Pubertät sowie dem Beginn von sexuellen Beziehungen [5].

Abbildung 2.1:
Während 10% der befragten Jugendlichen Sexts erstellt haben, gaben 17% an, Sexts von Jugendlichen erhalten zu haben.

Quellen: [1,4,5,2]

Während zum Sexting-Verhalten von Jugendlichen in Deutschland keine aussagekräftigen Zahlen vorliegen, wurden in den USA[1] bereits zahlreiche Erhebungen durchgeführt, welche jedoch deutlich voneinander abweichen. Eine Metaanalyse von insgesamt 12 unterschiedlichen US-Studien ergab eine Sexting-Beteiligung von 10,2% (bei einem Konfidenzniveau von 95% lag die Beteiligung zwischen 1,7% und 18,63%) [6]. Ebenfalls fällt auf, dass der Empfang von Sexts deutlich häufiger als das Versenden stattfindet, was darauf zurückzuführen ist, dass ein einmal produziertes Bild beliebig oft weitergeleitet werden kann. So gaben in den von Döring betrachteten Studien durchschnittlich 17% der befragten Jugendlichen an, bereits Sexts erhalten zu haben (im Gegensatz zu etwa 10%, die angaben, Sexts bereits selbst produziert und versendet zu haben) [1]. Wiederum 17% der Sext-Empfänger gaben an, erhaltene Sexts anderen Personen weitergeleitet oder gezeigt zu haben. Besonders die Weiterleitung von erhaltenen Sexts ist ein kritischer Aspekt, da dies zu einer durch den ursprünglichen Versender kaum kontrollierbaren Verbreitung des Bildes führen kann.

2.3 Gesetzeslage

Das unbefugte Weiterleiten eines Sexts stellt in Deutschland eine Straftat im Sinne des Allgemeinen Persönlichkeitsrechts[2] (APR) dar. Das tatsächliche Bewusstsein, gerade unter Jugendlichen, hierbei eine Straftat zu begehen, ist jedoch vermutlich eher gering ausgeprägt. Zahlreiche deutschsprachige Jugendschutzkampagnen wie Saferinternet.at, jugendundmedien.ch oder schau-hin.info weisen auf die Strafbarkeit eines solchen Weiterleitens hin und fördern somit die kritische Auseinandersetzung der Jugendlichen mit dem eigenen Handeln hinsichtlich des Umgangs mit erhaltenen Sexts.

Handelt es sich bei einem Sext um Kinder- oder Jugendpornografie, stellt seine Verbreitung eine Straftat gemäß § 184c StGB dar und kann mit einer »Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft« werden.

Anders als beispielsweise in den USA steht in Deutschland der Austausch von Sexts unter Jugendlichen ab 14 Jahren nicht unter Strafe. Konkret bedeutet dies, dass sowohl für eine jugendliche Person, die von sich selbst ein Sext anfertigt und dieses versendet, als auch für den jugendlichen Empfänger eines Sexts Straffreiheit gilt. Jedoch wird im Zuge der Umsetzung der EU-Richtlinie KOM (2010)94 eine Kriminalisierung von jugendlichen Sextern nach US-amerikanischem Vorbild befürchtet [4].

Sofern nur einer der beiden Kommunikationspartner minderjährig, der andere jedoch bereits erwachsen ist, liegt unter Umständen eine Straftat (durch die erwachsene Person) nach §176 StGB vor. Dies wird in Abschnitt 3.1 zum Thema Cybergrooming genauer beleuchtet.

1 In den USA wird das Thema u. a. deshalb deutlich stärker beforscht, da jugendliche Sexter sich nach der dortigen Gesetzeslage im Gegensatz zu Deutschland strafbar machen, wenn sie Nacktbilder von sich selbst versenden.

2 Das Allgemeine Persönlichkeitsrecht leitet sich aus Artt. 1, 2 GG ab und beinhaltet u. a. das Recht am eigenen Bild.

2.4 Risiken des Sexting

Mag die relative Zahl auch gering sein, kommt es dennoch immer wieder zu Sexting-Fällen mit verheerenden Folgen für die Opfer. Ausgangspunkt ist dabei meist das unerlaubte Weiterleiten eines oder mehrerer Sexts an Dritte durch den ursprünglichen Empfänger, welche zu Mobbing und Ausgrenzung des Sexters führen. So kann es zu einem Schneeballeffekt kommen, bei dem die Empfänger das ihnen weitergeleitete Sext wiederum an weitere Personen versenden und innerhalb kurzer Zeit ein weiter Personenkreis im Besitz des entsprechenden Bildes ist. Das Motiv für eine solche illegale Verbreitung kann beispielsweise Rache (z. B. aufgrund einer gescheiterten Beziehung) oder auch Geltungssucht (z. B. um im Freundeskreis zu prahlen) des Weiterleitenden sein [2].

FALLBEISPIEL

Die 13-jährige Hope Witsell aus Florida verschickte im Juni 2009 ein Sext an einen Jungen. Nachdem das Foto von dessen Handy aus weitergeleitet wurde, verbreitete es sich schnell weiter. An ihrer Schule kam es zu Mobbing und tätlichen Angriffen auf Witsell, ihre Eltern verhängten Hausarrest. Schließlich erhängte sie sich im September 2009 [1].

FALLBEISPIEL

Die 24-jährige Emma Jones aus Großbritannien nahm sich am 1. November 2013 das Leben. Zuvor hatte ihr ehemaliger Lebensgefährte Sexts, die sie ihm anvertraut hatte, auf Facebook veröffentlicht [7].

Neben Mobbing können erhaltene Sexts allerdings auch zur Erpressung des Sext-Produzenten genutzt werden (z. B. um weitere Bilder zu erzwingen). An diesem Punkt überschneidet sich das Sexting häufig mit Cybergrooming (Anbahnung eines sexuellen Kontakts zu einer minderjährigen Person über das Internet), welches in Kapitel 3 beschrieben ist.

FALLBEISPIEL

Die damals 12-jährige kanadische Schülerin Amanda Michelle Todd entblößte während eines Video-Chats mit einem unbekannten Mann ihren Oberkörper. Dieser erpresste sie anschließend damit, Aufnahmen des Chat-Verlaufs zu verbreiten, was er schließlich auch tat. Dies löste eine Mobbing-Welle gegen Todd aus, welche sogar nach dem Umzug in eine andere Stadt nicht nachließ. Mit 15 Jahren nahm Todd sich schließlich als Folge der andauernden psychischen Belastung durch das Mobbing am 10. Oktober 2012 das Leben, nachdem sie auf YouTube ein Video veröffentlicht hatte, in dem sie mit handgeschriebenen Karten auf ihre Situation aufmerksam machte [8].

2.5 Verwendete Kommunikationskanäle

Der mobile Internetzugriff ist heutzutage ein fester Bestandteil des Alltags von Jugendlichen in Deutschland. Das Internet- und Telekommunikationsverhalten deutscher Jugendlicher im Alter von 12 bis 19 wird durch die jährlich erscheinende Studie »Jugend, Information, (Multi-) Media« (JIM) vom »Medienpädagogischen Forschungsverbund Südwest« (mpfs) untersucht. Laut dieser besaßen

im Jahr 2017 97% aller Jugendlichen in Deutschland ein Smartphone, darunter in der Gruppe der 12- bis 13-Jährigen bereits 92%. Während 99% der 12- bis 19-Jährigen zumindest selten das Internet nutzen, sind laut JIM-Studie 89% Prozent von ihnen täglich online. Darüber hinaus erlaubt die Nutzung von öffentlichen oder Heim-WLAN-Netzen den permanenten Internetzugang nach dem »always on«-Prinzip.

Laut JIM-Studie 2015 versenden 94% der jugendlichen Smartphone-Nutzer regelmäßig Nachrichten via SMS oder Messenger-Dienste wie »WhatsApp«. Über 50% fertigen regelmäßig eigene Fotos oder Videos an [3]. Hierbei fällt auf, dass Mädchen (63%) deutlich häufiger Fotos oder Videos aufnehmen als Jungen (44%).

HINWEIS

In dieser Studie werden Sexts als Fotos betrachtet. Natürlich sind ebenso Videos dazu geeignet, erotische Inhalte von sich zu erstellen und zu verbreiten. Technisch gesehen macht es für diese Studie keinen Unterschied, ob ein Foto oder mehrere aufeinander folgende Fotos, die dann ein Video bilden, betrachtet werden.

Aufgrund der ständigen Verfügbarkeit per Smartphone und des unkomplizierten Austauschs von Textnachrichten sowie Mediendateien werden zum Sexting überwiegend Messenger-Dienste genutzt. Besonders häufig wird im Kontext von Sexting auf WhatsApp zurückgegriffen [9], was mit der generell hohen Beliebtheit, die der Messenger-Dienst bei Jugendlichen einnimmt, korreliert: Laut JIM-Studie 2015 wird WhatsApp von 95% der 12- bis 19-jährigen Smartphone-Nutzer regelmäßig genutzt und ist damit die am weitesten verbreitete App unter Jugendlichen [3]. Aber auch weniger oft genutzte Messenger-Dienste wie »Snapchat« oder das noch relativ neue »Kik« gehören zu den beliebten Kommunikationskanälen bei Sextern [10].

Die Beliebtheit von Snapchat lässt sich unter anderem dadurch erklären, dass die App damit beworben wird, geteilte Bilder auf Wunsch nach wenigen Sekunden auf dem Empfängergerät wieder verschwinden zu lassen. Mancher Sexter wiegt sich bezüglich der Weiterleitbarkeit seiner Sexts dadurch womöglich in relativer Sicherheit. Dass dies jedoch ein Trugschluss ist, wird in Abschnitt 2.5.1 erläutert.

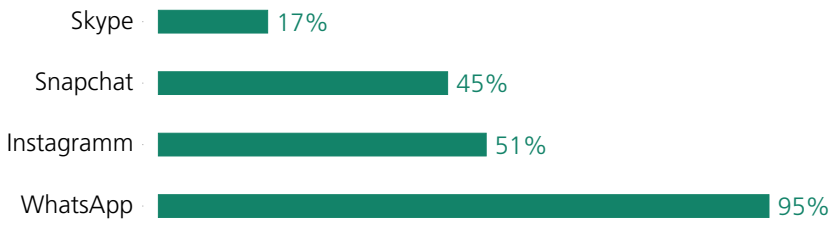


Abbildung 2.2: Anteil jugendlicher Smartphone-Besitzer bei der Nutzung verschiedener Online-Kommunikationsmedien. Quelle: [3]

Der kanadische Messenger »Kik« verspricht dem Benutzer Anonymität, da im Gegensatz zu Diensten wie WhatsApp oder Snapchat nicht dessen Telefonnummer, sondern lediglich ein Benutzername zur Identifikation verwendet wird [11]. Somit kann Kik auch auf nicht GSM-fähigen Geräten, etwa herkömmlichen PCs, Notebooks oder Tablet-Computern verwendet werden. Gerade die Anonymität, die Kik seinen Nutzern bietet, scheint für viele ein Anreiz zu sein, den Messenger zum Sexting mit Fremden oder für den kommerziellen Handel mit selbst hergestellten Sexts zu nutzen. Dementsprechend finden sich im Internet zahlreiche Websites, auf denen gezielt nach Sexting-Interessierten gesucht werden kann. Insgesamt ist Kik in Deutschland jedoch weitaus weniger verbreitet als WhatsApp und Snapchat.

Der Online-Dienst »Instagram«, welcher auf das Veröffentlichen und Teilen von Fotos spezialisiert ist, wird von 51% der Jugendlichen regelmäßig genutzt. Dieser wurde in der jüngeren Vergangenheit ebenfalls mit Sexting in Verbindung gebracht [12].

Abbildung 2.2 zeigt den jeweiligen Anteil von jugendlichen Smartphone-Nutzern, die bestimmte Messenger-Apps und soziale Plattformen nutzen, welche sich zum Sexting eignen.

2.5.1 Datensicherheit

Viele etablierte Messenger-Dienste, wie z. B. WhatsApp oder Skype bieten inzwischen standardmäßig eine Ende-zu-Ende-Verschlüsselung, was zumindest verhindert, dass gesendete Nachrichten (dies gilt auch für Bild-, Audio- und Videodateien) von Dritten unverschlüsselt abgefangen werden können. Zumindest im WhatsApp-Protokoll sind Schwachstellen[3] bekannt, die potenziell für Angriffe ausgenutzt werden können. Der Facebook-Messenger bietet mittlerweile ebenfalls die Option eines verschlüsselten Chats, diese muss jedoch durch beide Chat-Partner explizit eingeschaltet werden [13].

Der Messenger-Dienst Snapchat hat hingegen in der Vergangenheit negative Schlagzeilen in Bezug auf die Sicherheit der versendeten Daten gemacht, da durch die App versprochene Funktionalitäten nicht konsequent genug verfolgt werden und leicht umgangen werden können. Die vom Betreiber »Snap Inc.« zugesicherte Funktionalität zum automatischen Löschen versendeter Bilddateien auf dem Empfängergerät, welche in Abschnitt 2.5 erläutert wurde, findet in Wirklichkeit nicht statt. Das US-amerikanische Unternehmen »Decipher Forensics« veröffentlichte am 23. Januar 2014 einen Artikel [14], der beschreibt, wie man auch nach Ablauf des Zeitlimits versendete oder erhaltene Bilder im Speicher wieder auffinden kann. Das Vorgehen zur Wiederherstellung der Bilder ist auf zahlreichen Websites detailliert beschrieben und selbst durch technische Laien durchführbar. Darüber hinaus besteht auch ohne diese Hintertür immer die Möglichkeit, einen Screenshot eines gerade geöffneten Bildes zu erstellen, welcher dann wiederum beliebig weiterverwendet werden kann.

3 https://www.heise.de/newsticker/meldung/Krypto-Experte-Keine-Backdoor-in-WhatsApp-3596359.html

3 Cybergrooming

Im englischen Sprachgebrauch wird das Wort grooming[1] verwendet, um die Anbahnung eines sexuellen Kontakts durch gezieltes Einschmeicheln, wie etwa durch Gefälligkeiten oder Geschenke, zu bezeichnen. Wird hierbei das Internet als Kommunikationsmittel eingesetzt, spricht man im Englischen von Cybergrooming. Im deutschen Sprachgebrauch bezieht sich der entsprechende Begriff Cybergrooming ausschließlich auf die gezielte Kontaktaufnahme zu Minderjährigen im Internet mit der Absicht eines sexuellen Missbrauchs. Dabei muss der Missbrauch nicht zwingend körperlich stattfinden, sondern kann beispielsweise auch per Webcam über das Internet vollzogen werden [15]. Auch bereits das Zusenden von Textnachrichten mit sexuellem Bezug wird als sexuelle Belästigung gewertet [16] und fällt ebenfalls unter die Bezeichnung Cybergrooming.

Cybergrooming kann auch mit Sexting einhergehen, etwa dann, wenn Cybergrooming mit dem Ziel betrieben wird, Kinder und Jugendliche dazu zu bringen, Nacktaufnahmen (also Sexts) an den Täter (Cyber-Groomer) zu senden. Eine typische Vorgehensweise vieler erwachsener Cyber-Groomer, um sich das Vertrauen des Opfers zu erschleichen und überhaupt den Erstkontakt zu erreichen, ist, sich selbst als Kind bzw. Jugendlicher oder verständnisvoller Gesprächspartner auszugeben, um eine Vertrauensbeziehung aufzubauen, welche anschließend ausgenutzt werden kann [15].

3.1 Rechtslage

Die Strafbarkeit des sexuellen Missbrauchs von Kindern ist in §§ 176[2], 176a[3], 176b[4] StGB geregelt. Diese Paragrafen stellen jegliche sexuelle Handlungen an, mit und vor Kindern unter 14 Jahren unter Strafe. Ebenso macht sich strafbar, wer ein Kind zu sexuellen Handlungen an sich selbst oder anderen veranlasst. Darüber hinaus wird nach § 176 StGB bestraft, wer auf ein Kind »mittels Informations- oder Kommunikationstechnologie einwirkt, um [...] das Kind zu sexuellen Handlungen zu bringen«. Dies korreliert mit der Definition des Begriffs Cybergrooming und macht das Phänomen somit zu einem Straftatbestand. Straflos hingegen bleibt das Cybergrooming, wenn der Täter seinen Chat-Partner irrtümlicherweise für 14 Jahre oder älter hält [15].

FALLBEISPIEL

Das Oberlandesgericht Hamm verurteilte einen 55-jährigen Mann zu einer Freiheitsstrafe von 9 Monaten auf Bewährung, weil er der 9-jährigen Tochter einer Bekannten über den Messenger-Dienst WhatsApp intime Fragen zur Beziehung zu ihrem Freund stellte und fragte, ob sie eine Freundin habe, die mit ihm selbst eine Beziehung eingehen wolle. Die Verurteilung erfolgte mit der Begründung des sexuellen Missbrauchs von Kindern gemäß § 176 Abs. 4 Nr.3 StGB. Das

1 Engl. to groom: pflegen, striegeln, vorbereiten.

2 § 176 StGB: Sexueller Missbrauch von Kindern

3 § 176a StGB: Schwerer sexueller Missbrauch von Kindern

4 § 176b StGB: Sexueller Missbrauch von Kindern mit Todesfolge

> Gericht begründete die Verurteilung insbesondere mit der abstrakten Äußerung des Wunsches, etwas mit dem Kind und dessen Freund »machen« zu wollen, nachdem er zuvor sexuell anzügliche Nachrichten versendet hatte [17].

Kritiker argumentieren, dass das derzeitig geltende Recht zum Thema Cybergrooming lückenhaft sei. So wird gesagt, § 176 StGB decke nicht alle Arten der sexuell motivierten Kontaktaufnahme zu Kindern ab (z. B. muss dem Täter eine gewisse Hartnäckigkeit nachgewiesen werden). Ebenso wurde in der Vergangenheit (insbesondere seitens der Strafverfolgungsbehörden) bereits mehrfach eine Vorratsdatenspeicherung gefordert, um Verbindungsdaten bei Providern abrufen zu können. Deren Wiedereinführung wurde nach mehrfachen Rechtsstreitigkeiten mit der Umsetzung der EU-Richtlinie 2006/24/EG im Jahre 2015 zwar beschlossen, jedoch gestaltet sich die konkrete Umsetzung in der Praxis aufgrund massiven Gegenwinds und mehrerer Rechtsstreitigkeiten bislang schwierig.

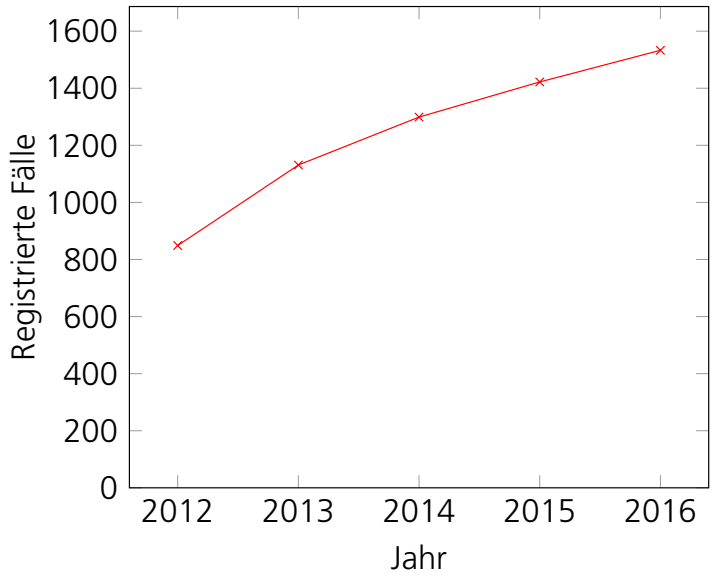


Abbildung 3.1: Registrierte Fälle von sexuellem Missbrauch von Kindern nach §§ 176, 176a, 176b StGB mit dem Tatmittel Internet in den Jahren 2012 bis 2016 (Quelle: Polizeiliche Kriminalstatistik [19]).

Zudem wird kritisiert, dass die Formulierung des Gesetzestextes zu ungenau sei und somit Schlupflöcher offen lasse [18]. Ein anderer Kritikpunkt an geltendem Recht zielt dagegen eher in Richtung des Täterschutzes. So wird moniert, dass mit § 176 StGB bereits die Vorbereitung zum sexuellen Missbrauch, welche selbst noch keinen Versuch darstellt, unter Strafe gestellt ist [15]. Inwieweit die verschiedenen Kritikpunkte am geltenden Recht inhaltlich sowie moralisch gerechtfertigt sind oder nicht, erfordert eine eingehende Auseinandersetzung mit dem Thema aus unterschiedlichen Blickwinkeln und ist nicht Diskussionsgegenstand dieser Studie.

3.2 Verbreitung

Da der Begriff Cybergrooming juristisch nicht definiert ist, existiert dementsprechend auch keine explizite Statistik zu diesem Tatbestand. Jedoch lässt sich eine solche aus Straftatbeständen im Zusammenhang mit den Paragrafen §§ 176, 176a, 176b StGB, die mit dem Tatmittel Internet begangen wurden, ableiten.

Diese werden in der polizeilichen Kriminalstatistik (PKS) des Bundeskriminalamts erfasst und jährlich veröffentlicht. Abbildung 3.1 zeigt die Zahl registrierter Fälle für die Jahre 2012 bis 2016. Demnach wurden im Jahr 2015 in der Bundesrepublik Deutschland insgesamt 1.422 Fälle des sexuellen Missbrauchs von Kindern nach den Paragrafen §§ 176, 176a und 176b StGB mit dem Tatmittel Internet registriert [19]. Im Vergleich zum Jahr 2014 (1.299 registrierte Taten) bedeutet dies einen Anstieg um 9,5%.

Eine Opferbefragung aus dem Jahr 2011 ergab jedoch, dass die Anzeigebereitschaft bei sexuellem Missbrauch insgesamt relativ niedrig ist (zwischen 11,9% und 18,4%) [20]. Geht man im Falle von Cybergrooming von einer ähnlichen Anzeigequote aus, lässt dies eine entsprechend hohe Dunkelziffer erwarten.

Darüber hinaus lässt sich vermuten, dass viele Fälle, bei denen der Kontakt seitens des Opfers bereits vor einem realen Treffen abgebrochen wurde, ebenfalls nicht zur Anzeige gebracht wird, da kein körperlicher Missbrauch stattgefunden hat.

3.3 Täter

Entgegen dem Stereotyp von Männern mittleren und höheren Alters als pädophile Straftäter ist eine auffallend große Zahl der Täter beim Cybergrooming eher jüngeren Alters. So sind 65% der Täter jünger als 30 Jahre. Darüber hinaus ist jeder dritte Tatverdächtige selbst noch ein Kind oder Jugendlicher [21]. Es handelt sich also keineswegs um ein rein auf die Pädophilie oder Pädokriminalität älterer Täter zurückzuführendes Phänomen.

Die vom Bundesministerium für Familie, Senioren, Frauen und Jugend[5] (BMFSFJ) geförderte Mikado-Studie[6] der Universität Regensburg nennt einige Charakteristika zu erwachsenen Cyber-Groomern. So zeichnet diese laut der Studie ein hohes Bildungsniveau und ein junges Alter aus. Bei knapp einem Viertel der im Rahmen der Studie befragten Täter handelte es sich um Frauen [22].

3.3.1 Vorgehensweisen

Laut dem Kriminologen Thomas-Gabriel Rüdiger[7], Cybergrooming-Experte an der Fachhochschule der Polizei des Landes Brandenburg (FHPol Brandenburg), lassen sich Cyber-Groomer je nach Planungsgrad ihrer Vorgehensweise in zwei Gruppen einteilen [23]:

- kurzfristig spontan handelnde Täter
- langfristig strategisch handelnde Täter

5 https://www.bmfsfj.de/

6 http://www.mikado-studie.de/, Juli 2018

7 https://de.linkedin.com/in/tgruediger

Erstgenannte planen die Ansprache eines potentzellen Opfers kaum und offenbaren früh im Gespräch ihre eigentlichen Ziele (z. B. den Wunsch nach Webcam-Sex). Häufig wird in Chat-Foren einfach eine öffentlich lesbare Anfrage nach der Bereitschaft zu sexuellen Handlungen gestellt. Wird darauf eingegangen, versucht der Täter schnell, auf ein sichereres Kommunikationsmedium auszuweichen, wie etwa einen verschlüsselten Messenger (z. B. WhatsApp oder Skype), um den weiteren Chat-Verlauf dem Zugriff durch Dritte zu entziehen. Bei der Auswahl ihrer Opfer unternehmen die spontan agierenden Täter eine Einschätzung ihrer Erfolgsaussichten. Sofern sie diese als hoch genug erachten, versuchen sie, ihre Opfer möglichst rasch zu den gewünschten sexuellen Handlungen zu bringen. Einmal auf die Forderungen des Täters eingegangen (z. B. durch Zusenden eines Nacktfotos), hat der Täter ein Druckmittel, welches er beispielsweise einsetzen kann, um von seinem Opfer weitere Bilder oder Aufnahmen zu erpressen [23].

Im Gegensatz zum spontan agierenden Täter plant der strategisch handelnde Täter seine Aktionen im Vorfeld. Er spricht gezielt potenzielle Opfer an und setzt dabei die Möglichkeiten der Online-Welt gezielt für sich ein (z. B. Kontaktaufnahme über ein Online-Spiel).

Da Nicknamen nicht mehrfach vergeben werden können, ist es üblich, seinen Wunschnamen durch weitere Zeichenketten zu erweitern, bis ein eindeutiger Nickname entsteht. Häufig werden dazu Zahlenkombinationen mit einer bestimmten Bedeutung, beispielsweise das Geburtsjahr oder das Alter des jeweiligen Nutzers verwendet. Hinter dem Nickname »Lisa13« lässt sich somit z. B. ein 13-jähriges Mädchen vermuten. Dadurch ist es für strategisch handelnde Täter oft relativ einfach, Minderjährige anzusprechen, die in ihr Beuteschema passen. Häufig täuschen diese Täter ihre Opfer durch falsche Altersangaben, da diese beim Chat mit einem vermeintlichen Gleichaltrigen meist weniger skeptisch sind. Nach [22] wendet etwa ein Drittel der erwachsenen Cyber-Groomer derartige Täuschungsstrategien an.

FALLBEISPIEL

Im Jahr 2008 kam es zur Verurteilung eines 53-jährigen Mannes vor dem Landgericht Konstanz, der eine 14-jährige unter Angabe eines deutlich jüngeren Alters im Chat dazu brachte, sich mit ihm zu treffen, woraufhin er sie entführte und laut Aussage des Mädchens mehrfach vergewaltigte. Aufgrund pubertärer Auflehnung gegen die Eltern und familiärer Probleme des Mädchens fiel es dem Täter besonders leicht, sich als verständnisvoller Chat-Partner zu inszenieren und so das junge Mädchen dazu zu bringen, ihm besonderes Vertrauen entgegenzubringen [24, 25].

Ziel des strategisch handelnden Täters ist es, eine Vertrauensbeziehung zwischen sich und dem Opfer aufzubauen, welche anschließend zur sexuellen Ausbeutung genutzt werden kann. Da dieses strategische Vorgehen deutlich aufwendiger ist, hat der strategisch planende Täter in der Regel deutlich weniger Opfer, mit denen er kommuniziert, als der kurzfristig spontan handelnde [26].

Wird das Cybergrooming nicht nur für den sexuellen Missbrauch über das Internet genutzt, sondern um ein reales Treffen mit dem Opfer vorzubereiten, verfahren Täter häufig nach einem vierstufigen Verfahren [27]:

1. **Kontaktaufnahme:**
   Zunächst sucht der Täter im Internet gezielt nach Opfern, die seinem Beuteschema entsprechen und spricht diese an.

2. **Vertrauensaufbau:**
   Um das Vertrauen seines Opfers zu gewinnen, spielt der Täter Interesse an dessen Lebenswelt vor oder gibt sich als Gleichaltriger aus. Dabei versucht er, das Opfer in ein Abhängigkeitsverhältnis zu bringen.

3. **Identitätsprüfung:**
   Um sicherzugehen, dass es sich bei seinem Chat-Partner auch tatsächlich um ein Kind des angegebenen Alters handelt, fordert der Täter die Angabe von Profilen in sozialen Netzwerken, Bilder oder Webcam-Übertragungen an.

4. **Übergriff:**
   Der Täter fragt sein Opfer über dessen sexuellen Entwicklungsstand aus und sendet bzw. verlangt pornografisches Material, wie z. B. Nacktaufnahmen. Um einen realen Missbrauch zu begehen, versucht der Täter, sein Opfer zu einem Treffen zu überreden.

Doch auch profane Strategien, wie das Versprechen materieller Zuwendungen oder Geschenke werden von Tätern angewendet, um ihre Opfer zu einem Treffen zu bewegen, wie die folgenden beiden Beispiele zeigen.

FALLBEISPIEL

Ein 45-jähriger Mann wurde vom Landgericht Tübingen zu sieben Jahren Haft verurteilt, weil er über das Internet mehreren 13- bis 16-jährigen Mädchen Bilder seines Geschlechtsteils schickte und sie zu Treffen aufforderte, woraufhin es auch zu erzwungenem sexuellen Kontakt kam [28]. Mit dem Versprechen, ihnen Kleidung, Handys und andere Dinge zu kaufen, konnte der Mann die Mädchen dazu bringen, sich auf ein Treffen mit ihm einzulassen. Der Fall verdeutlicht auch die hohe Dunkelziffer an Cybergrooming-Fällen, da lediglich eines von insgesamt 40 Mädchen, denen der Mann pornografische Selbstaufnahmen geschickt hatte, die Polizei informierte [29].

FALLBEISPIEL

Ein 35-jähriger Administrator der Online-Spielwelt »Minecraft« erschlich sich im Sommer 2016 das Vertrauen eines dort angemeldeten 12-jährigen Jungen, indem er ihm Vergünstigungen für das Spiel gewährte. Schließlich brachte der Mann den in der Schweiz lebenden Jungen in seine Düsseldorfer Wohnung, wo er ihn mehrfach missbrauchte, bis nach 8 Tagen ein Spezialkommando die Wohnung stürmte. Die Eltern des Jungen hatten keine Kenntnis von der Internetbekanntschaft ihres Sohnes. Freunde dagegen wussten von seinem Vorhaben, das Elternhaus zu verlassen [30, 31]. Dieser Fall von Cybergrooming ist insofern besonders brisant, da der Täter als Administrator des Online-Spiels eigentlich zu einer Gruppe von Personen zählt, denen bei der Verhinderung von Cybergrooming eine Schlüsselrolle zukommt (z. B. durch Sperren von entsprechend auffälligen Nutzern).

Opfertyp	Anteil unter viktimisierten Mädchen	Belastung durch Viktimisierung	Interesse an sexuellen Chats
Unauffällige	33%	schwach	nein
Souveräne	30%	schwach	ja
Brave-Schockierte	18%	hoch	nein
Traumatisierte	12%	hoch	kaum
Abenteuerinnen	7%	schwach	ja

Tabelle 3.1: Fünf Opfertypen von Mädchen bei Cybergrooming. Nach: [16]

3.3.2 Statistische Daten

Im Rahmen der Mikado-Studie [22] wurden Erwachsene sowie Kinder in Deutschland und Finnland zu Häufigkeit, Ursachen, Bedingungen sowie Auswirkungen sexuellen Missbrauchs von Kindern und Jugendlichen befragt. Von etwa 581 befragten Erwachsenen mit sexuellen Onlinekontakten gaben 19% an, bereits sexuelle Onlinekontakte zu Jugendlichen gehabt zu haben. Sexuelle Onlinekontakte zu Kindern hatten 5% der Befragten. Bei einem Drittel der sexuellen Onlinekontakte zwischen Erwachsenen und Minderjährigen kam es auch zu realen Treffen, wobei es in 100% der Fälle zu sexuellen Handlungen gekommen sei.

3.4 Opfer

Grundsätzlich wird hier der Fokus auf Kinder unter 14 Jahren als (potenzielle) Opfer von Cybergrooming gelegt, da bei ihnen das Cybergrooming nach § 176 StGB in jedem Fall strafbar ist. Der Kriminologe Thomas-Gabriel Rüdiger geht davon aus, dass jedes Kind, das mit digitalen Medien aufwächst, mindestens einmal in seinem Leben auf einen Cyber-Groomer im virtuellen Raum trifft [21].

3.4.1 Internetzugang

Voraussetzung dafür, Opfer von Cybergrooming zu werden, ist Zugang zu einem Kommunikationsmittel mit Internetverbindung. In heutigen Haushalten sind dies i.d.R. gewöhnliche Desktop-PCs bzw. Laptops, Smartphones bzw. Tablets sowie internetfähige Spielekonsolen. Laut der KIM-Studie 2016 [32] liegt der Anteil an Kindern zwischen 6 und 13 Jahren, in deren Haushalt sich mindestens ein Computer befindet, bei 97%. Ebenso hoch ist der Anteil der Haushalte mit einem Internetzugang. Über ein Handy oder Smartphone verfügten 98%, über eine Spielekonsole 75% der Haushalte. Der Anteil der Kinder, die einen eigenen Computer besitzen, liegt bei 20%. Ein eigenes Smartphone dagegen besitzen 32% der Kinder und Spielekonsolen 44%. 60% der 6- bis 13-Jährigen nutzen laut der Studie regelmäßig einen Computer, eine Spielekonsole oder ein Handy bzw. Smartphone. Über 50% nutzen das Internet mindestens einmal pro Woche, während es unter den 12- bis 13-Jährigen bereits 87% regelmäßig nutzen.

3.4.2 Viktimisierung begünstigende Faktoren

Ein pauschal definierbarer Opfertypus hinsichtlich Cybergrooming lässt sich nicht feststellen. Jedoch beeinflussen gewisse Faktoren, wie z. B. individuelle Charaktereigenschaften oder das Umfeld einer minderjährigen Person das Risiko, Opfer von Cybergrooming zu werden. Hierzu führte Katzer [16] im Jahre 2005 eine Studie unter insgesamt 1.700 Schülern und Schülerinnen in Nordrhein-Westfalen durch (davon 760 Jungen und 940 Mädchen). Die Befragten befanden sich in der 5. bis 11. Schulklasse und setzten sich aus Gymnasiasten, Realschülern, Gesamtschülern und Berufsschülern zusammen. Zu den gefährdenden Faktoren zählen laut Katzer beispielsweise ein generell risikofreudiges Verhalten (z.B. das Aufsuchen zwielichtiger Umgebungen, Drogenkonsum, Delinquenz), eine ausgeprägte sexuelle Neugier sowie ein fehlendes Risikobewusstsein.

Analog zu sexuellem Missbrauch in der physischen Welt werden laut Katzer überwiegend Mädchen Opfer von Cybergrooming. In 58% der in der Studie berichteten Cybergrooming-Fällen kam es zu Formen sogenannter leichter sexueller Viktimisierung (z. B. ungewolltes Fragen nach Aussehen oder Unterhaltung über sexuelle Themen), bei 18% sogar zu schwerer sexueller Viktimisierung (z. B. Erhalt von Pornografie oder Nacktfotos, Aufforderung zu sexuellen Handlungen vor der Webcam). Weiter stellte Katzer fest, dass leichte sexuelle Viktimisierung häufiger ältere Mädchen betraf, während das Alter bei schwerer sexueller Viktimisierung eine geringere Rolle spielte [16].

Katzer ermittelte im Rahmen ihrer Studie fünf Gruppen von typischen weiblichen Cybergrooming-Opfern, welche in Tabelle 3.1 zusammengefasst dargestellt sind. In der Gruppe der Abenteuerinnen war das Risiko einer schweren sexuellen Viktimisierung aufgrund ihres Chat-Verhaltens deutlich höher als in den übrigen Gruppen.

Bei den Opfern von Cybergrooming führt die eigene Scham oft dazu, dass sie sich nicht oder erst spät einer anderen Person wie z.B. Eltern oder Freunden anvertrauen und über etwaige Cybergrooming-Fälle sprechen [33].

3.5 Verwendete Kommunikationskanäle

Zur ersten Kontaktaufnahme dienen Cyber-Groomern häufig Chat-Portale oder Online-Spiele, die über ein relativ unkompliziertes Anmeldeverfahren verfügen. Ist der Kontakt zu einem potenziellen Opfer erst einmal aufgebaut, wird jedoch i. d. R. versucht, die weitere Kommunikation möglichst rasch auf ein sicheres Kommunikationsmedium zu verlagern. Dadurch wird vermieden, dass strafrechtlich relevante Kommunikationsinhalte durch Dritte (z. B. Forenadministrator oder Strafverfolgungsbehörden) abgehört werden. Im Allgemeinen suchen Cyber-Groomer bevorzugt solche Internetplattformen für die Opfersuche auf, die besonders von Kindern und Jugendlichen genutzt werden. Dies sind insbesondere Chat-Foren, Online-Spiele und Messenger-Dienste. Die drei Plattformarten werden nachfolgend im Einzelnen erläutert.

Messenger-Dienst	Registrierung über
WhatsApp	Handynummer
Snapchat	E-Mail
Signal	Handy- / Telefonnummer
kik	Benutzername
Skype	E-Mail
ICQ	E-Mail

Tabelle 3.2: Häufig verwendete Messenger-Dienste

3.5.1 Chat-Foren

Chat-Foren stellen für Cyber-Groomer eine unkomplizierte Möglichkeit dar, anonym und unbeobachtet Kontakt zu Minderjährigen aufzunehmen, um diese im weiteren Verlauf sexuell auszubeuten. Zwar werden die einzelnen Chatrooms bei einigen Foren von Administratoren überwacht, die unangemessene Nachrichten entfernen und Benutzer sperren können, jedoch handelt es sich hierbei aus Kostengründen um eine kleine Anzahl von Mitarbeitern des Forenbetreibers, für welche es unmöglich ist, jederzeit die gesamte Kommunikation aller Chat-Teilnehmer mitzuverfolgen.

Außerdem bieten die meisten Forenbetreiber gleichzeitig auch einen Privat-Chat an. In solchen Privat-Chats findet die Kommunikation zwischen zwei Chat-Partnern unbeobachtet statt, einige Betreiber blenden jedoch Warnhinweise vor potenzieller Belästigung ein und bieten eine Funktionalität, unangemessene Äußerungen des Chat-Partners an die Administratoren des Forums zu melden.

In vielen Fällen wird dann das entsprechende Nutzerkonto oder der Nickname für das Forum gesperrt. Doch mit der bloßen Sperrung von Nutzerkonten oder Nicknamen ist dem Phänomen Cybergrooming nicht effizient beizukommen, da eine solche Sperrung durch den Foren-Betreiber einen Täter nicht daran hindert, sich unter einer neuen Identität anzumelden und sich erneut anonym in Chatrooms zu bewegen. Laut der Initiative »jugendschutz.net« erfolgen sexuelle Übergriffe gegen Minderjährige im Internet bevorzugt über Privatnachrichten in Chat-Foren oder Messengern [34].

Knuddels: Knuddels[8] ist ein im deutschsprachigen Raum verbreitetes Chat-Forum, welches seit 1999 existiert. Laut eigenen Angaben ist es mit mehr als einer Million registrierten Mitgliedern das größte Chat-Forum Deutschlands und engagiert sich u. a. durch Aufklärungsarbeit, ein Moderatorensystem für öffentliche Chat-Räume und einen »Notrufbutton« aktiv für die Sicherheit seiner Chat-Teilnehmer. Für die Anmeldung fordert Knuddels zwar ein Mindestalter von 14 Jahren, jedoch wird die Altersangabe nicht überprüft, sodass hier falsche Angaben gemacht werden können und sich jeder unabhängig von seinem tatsächlichen Alter anmelden kann. Gespräche, die im Rahmen dieser Studie mit verschiedenen Landeskriminalämtern geführt wurden, ergaben, dass Knuddels bereits seit längerer Zeit im Fokus von Ermittlern steht, da Cybergrooming dort ein häufig auftretendes Phänomen ist.

8 https://www.knuddels.de/

3.5.2 Online-Spiele

Unter der großen Zahl an verfügbaren Online-Spielen existieren zahlreiche, die speziell auf Kinder und Jugendliche zugeschnitten sind. Solche Spielwelten sind bei Cyber-Groomern sehr beliebt, da sie dort auf eine große Menge potenzieller Opfer treffen, an die sie über das gemeinsame Spielen leicht und unverdächtig herantreten können. Im Folgenden werden exemplarisch einige Online-Spiele, die bereits im Zusammenhang mit Cybergrooming aufgefallen sind, kurz porträtiert.

Habbo Hotel: In dem Online-Spiel »Habbo Hotel«[9] können Spieler virtuelle Hotelzimmer mit Möbelstücken ausstatten. Das Spiel ist besonders auf die Interaktion zwischen den Spielern ausgelegt. Jeder Spieler tritt über einen Avatar in den Spielräumen auf und kann mit beliebigen anderen Spielern via Chat-Funktion kommunizieren. Er hat hierbei die Wahl zwischen einem Privat-Chat mit einem einzelnen anderen Spieler oder mit einer Gruppe von Spielern. Neben einer Webbrowser-Version ist »Habbo Hotel« auch als App für iOS- und Android-basierte Smartphones verfügbar. Das Spiel verfügt über folgende Sicherheitsmechanismen [35], die gegen Cybergrooming gerichtet sind:

- einen Wortfilter, der rassistische, beleidigende und sexuell anzügliche Wörter durch das harmlose Fantasiewort »bobba« ersetzt. Der Mechanismus, auf dem der Wortfilter basiert, ist ein simpler Abgleich mit einer Liste durch den Forenbetreiber verbotener Wörter. Da die Wörter lediglich bei exakter Übereinstimmung gefiltert werden, ist der Wortfilter leicht zu umgehen. Da beispielsweise auch die Namen einschlägiger Messenger-Dienste wie WhatsApp oder Kik aufgrund ihrer Beliebtheit bei Cyber-Groomern gefiltert werden, stellen diese ihnen einen Präfix voran (z.B. »xskype« statt »skype« oder »xkik« statt »kik«), sodass sie von dem Wortfilter zwar nicht als verboten erkannt werden, vom jeweiligen Chat-Partner jedoch noch verstanden werden können.

- eine Schaltfläche zum Blockieren der Privatnachrichten bestimmter Spieler,

- eine Raummoderation, die es dem Spieler u. a. erlaubt, zu entscheiden, welche anderen Spieler sich in den eigenen virtuellen Räumen befinden dürfen,

- anonyme Benutzerprofile der Spieler.

Durch die Anonymität der Benutzerprofile haben allerdings auch pädokriminelle Täter leichtes Spiel, da sie sich nicht dem Risiko aussetzen müssen, anhand ihrer E-Mail-Adresse, Telefonnummer, o.ä. enttarnt zu werden. Habbo Hotel stand in der Vergangenheit mehrfach in der Kritik, da es dort vermehrt zu Cybergrooming gekommen war [36].

9 https://www.habbo.de/

Clash of Clans: »Clash of Clans«[10] ist ein Strategiepiel für iOS- oder Android-basierte mobile Endgeräte. Ziel ist es, zusammen mit anderen Mitspielern eine Gemeinschaft aus mehreren Spielfiguren sowohl wirtschaftlich als auch militärisch gegen andere Gemeinschaften zu behaupten. Ähnlich wie bei »Habbo Hotel« existiert auch bei »Clash of Clans« eine Chat-Funktion, jedoch beschränkt diese sich auf Gruppengespräche und lässt keine private Kommunikation zwischen zwei Einzelpersonen zu. Allerdings wird in dem englischsprachigen Blog »Kids Privacy«[11] von vereinzelten Aufforderungen zum Kommunikationswechsel auf den für Cybergrooming bekannten Messenger »Kik« (s. Abschnitt 3.5.3) berichtet. Ob diese Aufforderungen tatsächlich im Zusammenhang mit Cyber- Grooming stehen, wird jedoch nicht berichtet, da die Autorin vermutlich nicht auf die Aufforderungen eingegangen war. Zusätzlich existiert ebenso wie bei »Habbo Hotel« eine Funktion zum Blockieren der Nachrichten anderer Spieler sowie unangemessene Beiträge zu melden [37]. Es sind zwei Fälle bekannt, in denen über das Spiel »Clash of Clans« Kontakt zwischen Cyber-Groomern und ihren Opfern zustande kam, woraufhin es in beiden Fällen auch zu tätlichem sexuellem Missbrauch kam [38].

Minecraft: »Minecraft« ist ebenfalls ein Strategiespiel und mit 144 Millionen verkauften Exemplaren eines der meistverkauften Spiele weltweit [39]. Im Mehrspielermodus können einander fremde Spieler online miteinander in Kontakt treten. Wie im Beispielkasten auf Seite 15 geschildert, kam es auch in Minecraft zu Cybergrooming, das schließlich zur Entführung und dem sexuellen Missbrauch eines 12-jährigen Jungen führte.

3.5.3 Messenger

Im Gegensatz zu Online-Foren und Online-Spielen werden Messenger-Dienste im Zusammenhang mit Cybergrooming meist erst nach dem Erstkontakt verwendet. Da die meisten Messenger mittlerweile über Ende-zu-Ende-Verschlüsselungsmechanismen[12] verfügen, bieten sie dem Täter beim Cybergrooming eine hohe Sicherheit vor dem Zugriff durch Strafverfolgungsbehörden oder die Dienstanbieter. Während die Verschlüsselung von Nachrichteninhalten im Hinblick auf Sexting ein gewisses Maß an Sicherheit bietet, da die Bilder zunächst nur den beiden Kommunikationspartnern zugänglich sind, stellt sie bei der Erkennung von Cybergrooming dagegen eine große Hürde dar. Der strafrechtlich relevante Teil der Kommunikation wird daher bevorzugt über verschlüsselte Messenger-Nachrichten vollzogen. Des Weiteren bieten solche Messenger in der Regel auch eine Video-Chat-Funktion, welche sowohl zur Überprüfung der Identität des Opfers als auch zum Austausch erotischen oder pornografischen Bildmaterials genutzt wird. Besonders sicher aus Täterperspektive sind Messenger, bei deren Registrierung keine Handynummer erforderlich ist, wie z.B. der kanadische Messenger »kik« (s. Abschnitt 2.5). Tabelle 3.2 listet die bekanntesten Messenger-Dienste auf.

10 https://clashofclans.com/

11 https://kidsprivacy.net/, Juli 2018

12 Ende-zu-Ende-Verschlüsselung bedeutet, dass der Inhalt einer Nachricht auf dem Gerät des Senders verschlüsselt wird und erst auf dem Gerät des Empfängers wieder entschlüsselt werden kann. Sofern die Verschlüsselung an sich sicher ist und niemand außer dem Empfänger über den entsprechenden Entschlüsselungsschlüssel verfügt, bedeutet dies, dass die Nachricht auf ihrem Weg zwischen Sender und Empfänger von keiner dritten Partei gelesen werden kann.

3.6 Feldstudie

Um ein Verständnis für die Verbreitung von Cybergrooming in Internet-Foren und die Vorgehensweise von Tätern zu bekommen, wurde eine eigene stichprobenartige Feldstudie durchgeführt. Hierzu wurde sich auf verschiedenen Chat-Foren mit Nicknamen angemeldet, welche aus einem weiblichen Vornamen gefolgt von der Zahl 13 bestanden (z.B. »jana13«, »bianca13«). Da an den Namen angehängte Zahlen häufig als Altersangabe interpretiert werden, wurde vermutet, dass Personen mit derartigen Nicknamen von Cyber-Groomern in Foren bevorzugt angesprochen werden. In einigen Foren (insbesondere Knuddels) dauerte es in der Regel nur wenige Sekunden, bis Anfragen nach Privat-Chats eingingen. Diese Erkenntnis deckt sich mit den Erfahrungen anderer Rechercheure [40]. Die Anfragen häuften sich teils sehr stark und gingen in sämtlichen Fällen von männlichen[13] Nutzern aus. Laut eigener Angaben dieser Cyber-Groomer variierte deren Alter zwischen 25 und 45 Jahren. Die Recherchen wurden grundsätzlich passiv durchgeführt, d.h. es wurden keine Forennutzer aktiv angesprochen, sondern nur auf eingehende Chat-Anfragen reagiert. Des Weiteren wurden Fragen in einer möglichst knappen Form beantwortet und während des Gesprächsverlaufs explizit auf das angebliche Alter von 13 Jahren hingewiesen.

In sämtlichen Fällen kam es zu eindeutiger sexueller Belästigung bzw. Fragen nach einem Treffen durch den Chat-Partner. Während einige der Chat-Partner die Kontaktaufnahme sehr direkt gestalteten und bereits zu Beginn nach der Bereitschaft zu sexuellen Handlungen fragten, begannen die meisten zunächst ein unverfängliches Gespräch, um kurz darauf ebenfalls sexuelle Absichten zu offenbaren. Diese Beobachtungen decken sich also mit den in Abschnitt 3.3 geschilderten Tätertypen.

Folgende Schlussfolgerungen konnten aus den durchgeführten Recherchen abgeleitet werden:

- 100% der sexuellen Anbahnungsversuche wurden über Privat-Chats durchgeführt.
- 100% der Anfragen nach Privat-Chats wurden mit dem Ziel Cybergrooming gestellt.
- Einige der untersuchten Plattformen wiesen eine hohe Dichte von Cyber-Groomern unter den männlichen Forennutzern auf
- Viele Cyber-Groomer waren nach eigenen Angaben relativ jung (zwischen 20 und 30 Jahren). Zwar lassen sich diese Altersangaben nicht verifizieren, allerdings ließ die Verwendung jugendsprachlicher Ausdrücke ebenfalls auf ein eher junges Alter schließen.

Inwieweit es sich bei den vielen Benutzerkonten, über die die Cyber-Groomer Kontaktversuche unternahmen, um mehrere *Nicknamen* einer einzigen oder einiger weniger Personen handelt, konnte nicht untersucht werden. Es besteht also durchaus die Möglichkeit, dass die festgestellte Menge von Cyber-Groomern auf den untersuchten Plattformen tatsächlich niedriger ist und durch multiple gleichzeitig geführte Benutzerkonten verzerrt wurde.

13 Laut Profilinformationen.

4 Technische Grundlagen

In diesem Kapitel werden wichtige Konzepte sowie technische Grundlagen von Verfahren vorgestellt, die im Rahmen der automatisierten Erkennung von Sexting und Cybergrooming sinnvoll eingesetzt werden können. Das Kapitel dient somit als Basis für einige Lösungsansätze, die im weiteren Verlauf der Studie beschrieben werden.

4.1 Farbräume

Ein digitales Bild liegt zunächst lediglich in Binärdaten vor, wobei jedem Pixel ein bestimmter Wert (bei Grauwertbildern) bzw. drei Werte (bei Farbbildern) zugeordnet sind. Diese Pixelwerte sind sequenziell gespeichert. Um ein Bild korrekt darstellen zu können, muss das entsprechende Bildbetrachtungsprogramm die Pixelwerte adäquat interpretieren. Hierzu wird auf standardisierte Farbräume zurückgegriffen, welche genau vorschreiben, welcher Pixelwert welchem Farbwert zugeordnet wird. Transformationen zwischen den einzelnen Farbräumen sind jederzeit möglich. Nachfolgend sind einige im Bereich der Hauterkennung relevante und häufig diskutierte Farbräume beschrieben.

- **RGB** bezeichnet eigentlich ein Farb-Modell, welches durch eine Vielzahl von Farbräumen implementiert wird. RGB ist einer der meistgenutzten Standards zum Speichern und Bearbeiten von Bildern [41] und gliedert sich in drei Farbkanäle für die Grundfarben **R**ot, **G**rün und **B**lau, welche sowohl die visuelle Wahrnehmung des menschlichen Auges als auch die Anzeige auf Bildschirmen bestimmen. Farbe und Intensität (Helligkeit) eines Pixels ergeben sich aus der Zusammensetzung der drei Farbwerte.

- **Normalisierter RGB:** Durch eine Normalisierung der RGB-Werte kann man den Einfluss der Intensität verringern, was insbesondere bei der Identifikation von Hautpixeln von Vorteil ist, da auf diese Weise unterschiedliche Belichtungsverhältnisse kompensiert werden können.

- **HSV** ist ein Farbraum, der an die menschliche Farbwahrnehmung angelehnt ist. Eine Farbe im HSV-Farbraum wird über die drei Komponenten Farbwert (engl. **H**ue), Farbsättigung (engl. **S**aturation) und Helligkeitswert (engl. **V**alue) beschrieben. Um wie im Fall des normalisierten RGB-Farbraums die Intensität zu ignorieren, kann für die Hauterkennung der Helligkeitswert **V** verworfen werden.

- **YCbCr** gliedert die Pixelwerte eines Bildes in Grundhelligkeit (**Y**) und die zwei Farbkanäle blau-gelb (**Cb**) und rot-grün (**Cr**). Hier kann im Rahmen einer Hauterkennung analog zu HSV der Wert **Y** für die Grundhelligkeit weggelassen werden, um Verfälschungen durch unterschiedliche Belichtungsverhältnisse in den Bildern zu vermeiden.

4.2 Maschinelles Lernen

Maschinelles Lernen (ML) ist ein Kerngebiet der künstlichen Intelligenz (KI), welches dem Bereich der angewandten Informatik zuzuordnen ist. Vereinfacht gesprochen beschäftigt sich ML mit der Generierung von Wissen aus Daten und lässt sich in die folgenden vier Teilgebiete einteilen: **überwachtes, semiüberwachtes, unüberwachtes** sowie **bestärkendes Lernen**. Die Kategorie »überwachtes Lernen« spielt, hinsichtlich der diskutierten Lösungsstrategien in dieser Studie, eine wichtige Rolle. Sie stellt das populärste Teilgebiet des MLs dar, zu dem seit Jahrzehnten intensiv geforscht wird, und ist Gegenstand einer Vielzahl von Lösungen, die in der Industrie und Wirtschaft branchenunabhängig eingesetzt werden. Zu den wichtigsten Vertretern des überwachten Lernens zählen die **Klassifikation, Regression** und **Anomalieerkennung**. **Deep Learning** hingegen repräsentiert einen Spezialfall des MLs, welcher sich allen vier Teilgebieten zuordnen lässt.

INFO

Maschinelles Lernen ist heute eine vielbeachtetete Technologie. Wir gehen im Anhang für den interessierten Leser auf weitere Details ein, die für das Verständnis der späteren Kapitel hilfreich sind.

Allgemein strebt maschinelles Lernen an, mit speziellen Algorithmen Entscheidungen auf Basis bekannter Daten (Trainingsdaten) zu treffen. Durch eine Analyse der Trainingsdaten finden derartige Algorithmen Zusammenhänge zwischen Merkmalen (Features) der einzelnen Dateneinheiten, die auf eine bestimmte Eigenschaft schließen lassen. So wird ein Modell gelernt, das die Zuordnung einer bisher ungesehenen Dateneinheit gleichen Typs zu einer der bekannten Klassen erlaubt. Das sogenannte überwachte Lernen spielt im Hinblick auf die in dieser Studie diskutierten Lösungsstrategien eine große Rolle. Bei dieser Art von ML-Algorithmen ist die Klassenzugehörigkeit jedes Beispiels der Trainingsmenge vorab bekannt und wird von dem Algorithmus genutzt, um eine Abbildung von den übrigen Merkmalen der Beispiele auf die entsprechende Klasse zu finden. Dabei soll das gelernte Modell möglichst viele Entitäten korrekt klassifizieren, darf aber andererseits nicht zu eng an der Trainingsmenge orientiert sein, da sonst die Gefahr einer Überanpassung (Overfitting) an die Trainingsdaten besteht. Ein solches Overfitting führt letztlich dazu, dass das Modell zu sehr die Eigenheiten der Trainingsdaten widerspiegelt und somit oft nicht verallgemeinerbar auf neue Datensätze ist.

Ist das Lernen auf einer Trainingsmenge (Training) abgeschlossen, muss das gelernte Modell auf einer Testmenge evaluiert werden, um zu überprüfen ob es auf unbekannte Daten verallgemeinerbar ist und um seine Vorhersagegenauigkeit zu messen. Je nach Anwendungsdomäne existieren dazu unterschiedliche Maße, wie z. B. »precision/recall« oder »accuracy«. Wichtig ist, dass sowohl die Trainings- als auch die Testmenge (möglichst gleich viele) Beispiele aller Klassen enthalten. Geht es bei einer Klassifikation darum, Aussagen über das Vorhandensein einer einzigen Eigenschaft zu machen, wird der Einfachheit halber häufig von positiven und negativen Trainingsbeispielen gesprochen. Zur Klasse der positiven gehören dabei diejenigen Beispiele, die die gesuchte Eigenschaft aufweisen, zur Klasse der negativen zählen alle übrigen. Eine solche Klassifikation wird auch als binär bezeichnet, da zwischen zwei Klassen (positiv und negativ) unterschieden wird.

5 Stand der Technik

In diesem Kapitel werden Methoden vorgestellt, welche für die automatische Erkennung von Sexting und Cybergrooming geeignet sind. Es wird auch erörtert, wie fortgeschritten und zuverlässig diese nach aktuellem Stand der Technik bereits sind. Dabei werden ausschließlich technische Ansätze berücksichtigt. Im Bereich der automatischen Sexting-Erkennung sind dies in erster Linie Verfahren zur Detektion von Nacktbildern bzw. sogenannten NSFW[1]-Bildern. Unter der Bezeichnung NSFW werden im Netzjargon meist Nacktbilder verstanden. Zur Erkennung von Cybergrooming spielen hauptsächlich textbasierte Analyseverfahren eine Rolle. Diese wiederum gliedern sich in Verfahren zur Chat-Analyse und solche zur Identifizierung von Personen anhand ihres Schreibstils.

5.1 Erkennung von Sexting

Um eine automatische Sexting-Erkennung zu realisieren, bedarf es eines Algorithmus, der in der Lage ist, Bilder anhand der dargestellten Inhalte zu klassifizieren. Diese reichen von in aufreizenden Posen dargestellten Personen bis hin zu Pornografie. Bilder, auf denen der Algorithmus derartige Inhalte entdeckt, können je nach Policy mit einem Warnhinweis versehen oder direkt für das Versenden bzw. Speichern auf dem Gerät blockiert werden. Hinsichtlich der Umsetzung der entsprechenden Bildklassifikation gibt es verschiedene Herangehensweisen:

- Binäre Klassifikation: Betrachtet man die Sexting-Erkennung als binäres Entscheidungsproblem, werden zu prüfende Bilder einer von zwei Klassen (z. B. »Sext« und »kein Sext«) zugeordnet. Dies erfordert eine gute Trennbarkeit zwischen den beiden Klassen, da es sonst zu einer hohen Zahl von Fehlklassifikationen kommt.

- Multiklassen-Klassifikation: Als ordinales Entscheidungsproblem modelliert, lässt sich eine abgestufte Klassifikation mit mehreren Klassen realisieren (z. B. von »kein erotischer Inhalt« über »leicht erotischer Inhalt« bis hin zu »pornografischer Inhalt«). Durch diese Abstufung können Bilder, die Grenzfälle darstellen, mittleren Klassen zugeordnet werden, während eindeutige Bilder den äußeren Klassen zugeordnet werden. Wählt man eine ungerade Anzahl von Klassen, dient die mittlere als neutrale Klasse (d. h. für ein Bild, das dieser Klasse zugeordnet wird, trifft der Algorithmus keine Aussage). Bei einer geraden Anzahl von Klassen dagegen existiert immer eine Tendenz in die eine oder andere Richtung.

- Regression: Ebenfalls möglich ist die Modellierung als Regressionsproblem. Dabei wird im Gegensatz zur Klassifikation mit diskreten Klassen die Zuordnung zu einem numerischen Wert auf einer kontinuierlichen Werteskala vorgenommen. So können Bilder z. B. mit Werten zwischen 0 (»definitiv kein Sext«) und 1 (»definitiv Sext«) versehen werden. Wird ein Bild dem Wert 0,5 zugeordnet, bedeutet dies, dass der Algorithmus für dieses Bild keine eindeutige Aussage treffen kann.

1 Abk. für »Not Safe For Work«

Zwar existieren derzeit keine Verfahren, die explizit auf die Erkennung von Sexting ausgerichtet sind, jedoch kann man sich Methoden zunutze machen, welche eine Lösung zu verwandten Problemstellungen bieten. Diese werden im Folgenden beschrieben.

Da Sexts nackte Körperpartien enthalten, können beispielsweise Algorithmen zur Nacktbilderkennung eingesetzt werden, um abzuschätzen, ob ein gegebenes Bild nackte Personen darstellt oder nicht. Diese Algorithmen wiederum greifen zumeist auf Verfahren zur Hauterkennung zurück. Eine intensiv beforschte Disziplin, welche starke Überschneidungen zur reinen Nacktbilderkennung aufweist, ist die Erkennung von pornografischem Bildmaterial. Nacktbild- und Pornografieerkennung werden häufig auch unter dem allgemeineren Oberbegriff NSFW-Erkennung geführt. Aufgrund einiger länderspezifischer Richtlinien zur Beschränkung der Verfügbarkeit von erotischem Bildmaterial im Internet sowie Unternehmenspolitik, den eigenen Mitarbeitern am Arbeitsplatz den Zugriff auf solches Material zu verwehren, ist der Forschungsanreiz im Bereich der NSFW-Erkennung sehr hoch. Es existieren daher bereits einige zuverlässige Verfahren, welche zum Teil auch als kommerzielle Lösungen (meist in Form eines Webservice) angeboten werden.

In den genannten Disziplinen kommen, je nach Lösungsansatz, sowohl verschiedene Farbraummodelle als auch Algorithmen aus dem maschinellen Lernen (s. Abschnitt 4.1 bzw. C) zum Einsatz. Ebenso ist eine Kombination aus Algorithmen beider Bereiche möglich.

5.1.1 Farbwertbasierte Hauterkennung

Eine verbreitete und intuitive Methode, Nacktbilder automatisiert zu erkennen, basiert auf der Idee, in einem zu untersuchenden Bild Pixel zu identifizieren, deren Farbe mit bekannten Hautfarben übereinstimmen. Diese werden Hautpixel genannt. Bei vielen Verfahren wird die Summe der Hautpixel eines Bildes in das Verhältnis zur Gesamtmenge der Bildpixel gesetzt. Überschreitet der Anteil von Hautpixeln im Bild einen zuvor bestimmten Schwellwert, wird das Bild als Nacktbild klassifiziert. Allerdings stoßen derartige Algorithmen zur Nacktbilderkennung schnell an ihre Grenzen, wodurch sie für eine zuverlässige Sexting-Erkennung allein kaum ausreichend sind. Die folgenden drei Szenarien sollen die Probleme der farbwertbasierten Hauterkennung verdeutlichen:

- Nacktheit im Sinne eines Sexts muss nicht mit einer großen Menge sichtbarer Haut einhergehen. Das Bild einer entblößten weiblichen Brust ist sicher als Sext anzusehen, enthielte bei gleicher Kameradistanz aber i.d.R. eine geringere Menge von Hautpixeln als beispielsweise das Bild einer Frau in kurzem Sommerkleid. Dementsprechend ergeben sich in der Praxis hohe Falsch-Positiv-Raten (FPR) und Falsch-Negativ-Raten (FNR).

- Hautfarben sind nicht eindeutig bestimmbar. Beleuchtung und Sättigung eines Fotos beeinflussen die Farben der Pixel stark. Selbst, wenn diese Faktoren herausgerechnet werden können und nur die Farbtöne selbst betrachtet werden, bleibt das Farbspektrum, das sämtliche menschlichen Hauttöne abdeckt, sehr breit. Dies führt wiederum zur Überschneidung mit den Farbwerten vieler anderer Materialien wie z. B. Holz. Eine Auswahl von Samples unterschiedlicher Hauttöne im RGB-Farbraum ist in Abbildung 5.1 zu sehen. Abbildung 5.2 zeigt die gleichen Samples im HSV-Farbraum und die jeweilige Reduktion auf die Grundfarbe.

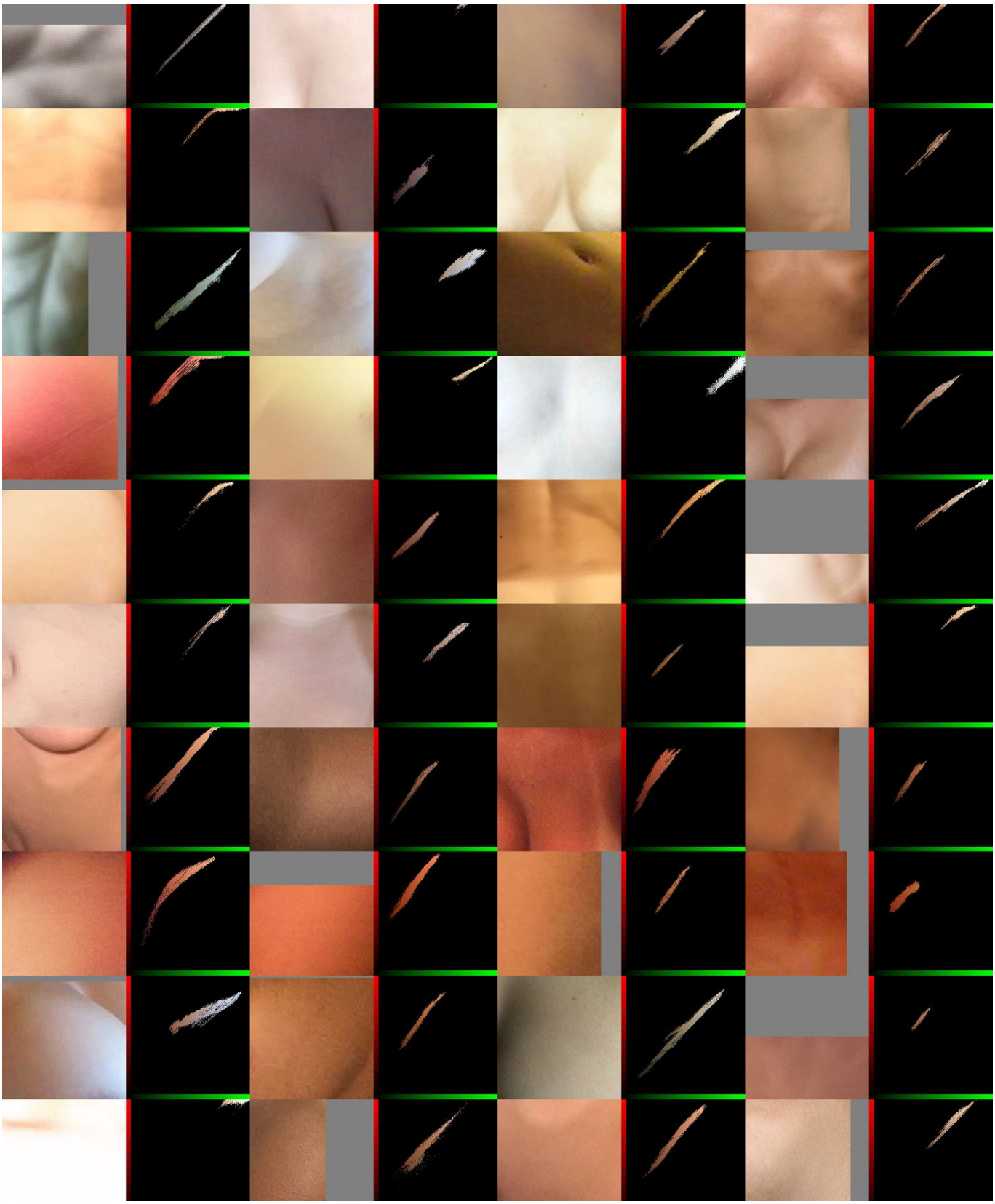


Abbildung 5.1:

Eine Sammlung von 20 Hautpartien im RGB-Farbraum. Es ist deutlich zu erkennen, wie vielfältig die Hautfarben sind. Links ist jeweils das Haut-Sample, rechts dessen Position im Farbspektrum aufgezeigt. Dabei stellt die Y-Achse den Rotanteil, die X-Achse den Grünanteil dar.

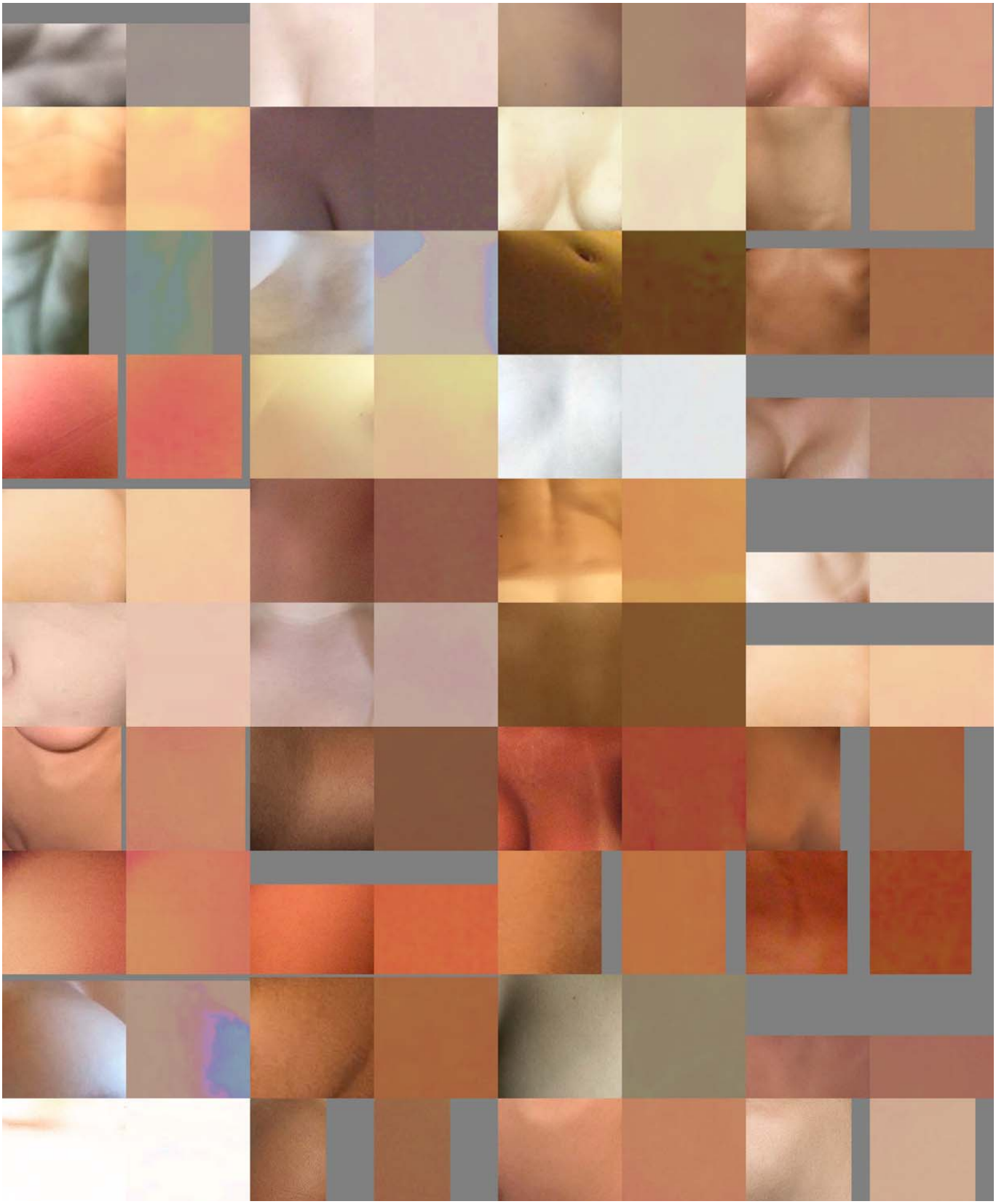


Abbildung 5.2:

Einfacher wird die Handhabe, wenn man im HSV-Farbraum arbeitet. Der Wert »H« stellt hier die Grundfarbe dar. Gezeigt sind die gleichen 20 Haut-Samples wie in Abbildung 5.1 und daneben ihre Reduktion auf die Grundfarbe. Helligkeit und Sättigung sind jeweils auf einen Durchschnittswert gesetzt und somit neutral.

Abbildung 5.3:
Links: Ein Beispiel für eine erfolgreiche Gesichtserkennung. Der grüne Rahmen umschließt die Nasenpartie, die als Referenz für Hautpixel herangezogen wird. Mitte: Ergebnis einer Filterung nach den so definierten Hautpixeln: Der Bereich um die Nase hat auch Pixel beinhaltet, welche Haarfarbe und Hintergrund umfassen. Rechts: Manuelle Definition von Hautfarben, wie im grünen Rahmen: Hier wird tatsächlich nur der Hautbereich nicht gefiltert

- Schwarz-Weiß-Bilder enthalten keinerlei Farbinformationen, weshalb farbwertbasierte Methoden für sie nicht anwendbar sind.

Beide Faktoren führen zu dem Problem, dass allgemeine Schwellwerte sehr anfällig gegen Falsch-Positiv- und Falsch-Negativ-Klassifikationen sind. Entweder wird ein großes Farbspektrum als Hautfarbe erkannt, dadurch werden aber andere Objekte schnell fälschlicherweise als Hautpartien angesehen, oder es werden nur sehr enge Farbgrenzen gesetzt, wodurch helle bzw. stark gebräunte Haut nicht mehr erkannt wird. Ebenso problematisch ist der Schwellwert des Pixelanteils; ein harmloses Sommerfoto in leichter Bekleidung kann hier nur schwer von einem Sext unterschieden werden.

Eine Alternative zu allgemein definierten Hautfarben bieten Verfahren, die für ein Bild die Hautfarbe individuell bestimmen. Verbreitet ist hier das Konzept der Hautfarbenbestimmung über eine Gesichtserkennung. Zuerst wird geprüft, ob auf einem Foto ein Gesicht erkannt werden kann. In diesem Gesicht wird dann ein Bereich als Definition für die aktuelle Hautfarbe verwendet, der üblicherweise tatsächlich ungeschminkte Haut zeigt; beispielsweise die Nasenpartie.

Die Pixel des Nasenbereichs dienen dann als Referenz für die übrigen Pixel des Bildes; stimmen ihre Farben überein, geht man von Hautbereichen aus.

Hier treten allerdings auch wieder Herausforderungen auf:

- Eine Erkennung von Gesichtern und darüber hinaus Gesichtsbereichen ist nicht immer zuverlässig. Es werden Gesichter nicht oder an völlig falschen Stellen erkannt. Dies führt zu nicht vorhandenen oder falschen Referenzpixeln.

- Die Beleuchtung eines Fotos kann im Gesichtsbereich anders als in anderen Hautpartien sein. Beleuchtung, Schatten oder Bräunungsgrad können einen starken Einfluss auf die Farbwerte der jeweiligen Pixel haben.

Abbildung 5.4:
Links: Eine gescheiterte Gesichtserkennung. Auf dem Bild wurden verschiedene Bereiche falsch als Gesichtspartien erkannt, dementsprechend liegt der grüne Rahmen mit der vermeintlichen Nasenpartie in einem zufälligen Bereich. Rechts: Die Filterung scheitert hier dementsprechend, es werden nur wenige Teile des Bildes herausgefiltert.

Ausgewählte Verfahren: Im Folgenden werden einige bereits existierende Verfahren zur Erkennung von Nacktbildern beschrieben:

- Jones und Rehg vom damaligen US-amerikanischen Computerhersteller Compaq beschrieben bereits in einem 1998 erschienenen Fachbericht ein farbwertbasiertes Verfahren zur Hauterkennung [42]. Mittels Histogrammanalyse ermittelten sie Farbspektren, anhand derer sich Hautpixel und Nicht-Hautpixel möglichst gut voneinander unterscheiden lassen. Auf einer selbst erhobenen Bildermenge erzielte das Verfahren eine TPR von 80% und eine FPR von 8,5%.

- Ap-Apid stellte im Jahr 2005 ein Verfahren zur Nacktbilderkennung vor, welches auf einer Menge von 1056 Bildern (421 Nacktbilder, 635 neutrale Bilder) eine TPR von knapp 95% und eine FPR von knapp über 5% erzielte [41]. Das Verfahren basiert auf der Beobachtung, dass Nacktbilder i.d.R. überproportional viele Hautpixel aufweisen und zusammenhängende Hautregionen nah beieinander liegen. In einem ersten Schritt werden die Hautpixel des Bildes klassifiziert. Dazu werden zunächst zwei unterschiedliche Farbraumdarstellungen des Bildes generiert: Eine im normalisierten RGB- und eine im HSV-Farbraum. Für beide Farbräume wurden zuvor Modelle für Hautpixel auf Trainingsbildern mit unterschiedlichen Hauttypen trainiert. Laut Ap-Apid hat sich ein Anteil von weniger als 15% Hautpixeln als guter Indikator dafür erwiesen, dass es sich bei dem geprüften Bild nicht um ein Nacktbild handelt. Solche Bilder werden durch den Algorithmus als Nicht-Nacktbilder klassifiziert. Bei einem Hautpixelanteil von 15% oder mehr werden weitere Merkmale wie z.B. die Größe der erkannten zusammenhängenden Hautregionen untersucht. Für diese wurden ebenfalls Schwellwerte festgelegt. Werden

sämtliche Schwellwerte überschritten, wird das entsprechende Bild als Nacktbild klassifiziert. Auf Basis dieses Verfahrens wurde eine freie Softwarelösung entwickelt, welche in den beiden Programmiersprachen Python (»nude.py«[2]) und JavaScript (»nude.js«[3]) verfügbar ist.

- Santos et al. entwickelten 2007 ein Verfahren zur Nacktbilderkennung, welches auf Farbwerten und Textur beruht [43]. Eine Einteilung des Bildes in Regionen unterschiedlicher Größe verbessert das Verfahren zusätzlich. Zunächst wird eine Normalisierung der Bilder auf 256 x 256 Pixel vorgenommen. Sämtliche Bilder werden, sofern dies nicht bereits der Fall ist, in das JPEG-Format konvertiert. Nun wird das Bild in den YCbCr-Farbraum überführt und Hautpixel mittels Histogrammanalyse klassifiziert. Anschließend wird eine Binarisierung der Pixelfarbwerte vorgenommen, d. h. Hautpixel werden weiß, Nicht-Hautpixel schwarz eingefärbt.

  Santos et al. folgen der Annahme, dass das Hauptaugenmerk eines Bildes in der Regel in dessen Mitte liegt. Auch für Nacktbilder wurde diese Beobachtung auf einer Menge von 508 Bildern gemacht [44]. Diese Annahme veranlasste Santos et el. dazu, konzentrische rechteckige Bildausschnitte verschiedener Größe zu betrachten. Für jedes dieser Rechtecke wurden farb- und texturbasierte Bildmerkmale extrahiert, welche wiederum dazu genutzt wurden, einen ML-Algorithmus zu trainieren. Die zum Training verwendeten Merkmale sind:

  - Anzahl verbundener Hautpixel
  - Verhältnis zwischen Hautpixeln zur Gesamtpixelzahl
  - Grauwertkontrast zwischen benachbarten Pixeln
  - Grauwertkorrelation zwischen benachbarten Pixeln
  - Anzahl der Vorkommen von Pixelpaaren mit gleichen Grauwerten
  - Homogenität der Grauwerte aller Pixel

  Mittels Kreuzvalidierung wurde der ML-Algorithmus auf 2.680 Bildern (davon 1.340 Nacktbilder) trainiert und das gelernte Modell auf einer Menge von 2.080 Bildern (davon 910 Nacktbilder) getestet. Während TPR und TNR bei Betrachtung des gesamten Bildes bei 84,9% bzw. 92% lagen, brachte die Beschränkung auf den kleinsten rechteckigen Bildausschnitt eine Erhöhung auf 96% bzw. 96,6%. Da TPR und TNR für größere Bildausschnitte niedriger waren, bestätigt sich hier die Vermutung, dass die relevanten Bildinhalte vor allem auf die Mitte konzentriert sind.

  Wurden alle Bildmerkmale aus sämtlichen Bildausschnitten berücksichtigt, konnten TPR und TNR weiter auf 97,4% bzw. 98,4% gesteigert werden. Da die Klassifikation für ein Bild im

2 https://github.com/hhatto/nude.py

3 https://www.patrick-wied.at/static/nudejs

Durchschnitt 30 Millisekunden dauerte, ist das Verfahren gut für eine Echtzeitanwendung geeignet. Santos et al. bemerkten zudem, dass Bilder, die fälschlicherweise als Nacktbilder klassifiziert wurden, einen hohen Anteil rot- und gelbfarbiger Pixel enthielten und z.B. Sand oder Gestein abbildeten.

Ein 2014 von Platzer et al. entwickeltes Verfahren zur Erkennung von Nacktbildern bzw. Pornografie ist »Skin Sheriff« [45]. Hierbei kommt eine farbwertbasierte Hauterkennung zum Einsatz, deren Vorhersagegenauigkeit durch ML-Algorithmen verstärkt werden soll. Der Ansatz ist in zwei Stufen unterteilt: (1) ein Hautpixelerkennungsverfahren und (2) ein Pornografieerkennungsverfahren.

Da Bilder in unterschiedlicher Qualität und Größe vorliegen können, werden diese vor der Hautpixelerkennung zunächst einem Vorverarbeitungsschritt unterzogen. Dieser stellt sicher, dass alle Bilder unter möglichst gleichen Ausgangsbedingungen verarbeitet werden. Zunächst wird das Eingabebild auf eine Breite von maximal 1.000 Pixeln skaliert, anschließend wird mit einem sogenannten Autokontrast eine Kontrastkorrektur vorgenommen. Dieser Schritt dient dazu, die Farbwerte auf stark über- bzw. unterbelichteten Bildern zu normalisieren.

Zur Hautpixelerkennung wird das Bild sowohl im RGB- als auch im HSV-Farbraum dargestellt. In den zwei Farbräumen klassifiziert der Algorithmus unabhängig voneinander die Hautpixel anhand zuvor bestimmter Schwellwerte. Die Schnittmenge der Pixel, die in beiden Farbräumen als Hautpixel klassifiziert wurden, bildet eine sogenannte »Skinmap«. Verrauschte Hautbereiche (d. h. Bereiche, in denen zusammenhängende Hautbereiche durch Nicht-Hautpixel unterbrochen werden) werden homogenisiert.

Um die FPR des Verfahrens zu reduzieren, wird im Anschluss ein zweiter Algorithmus angewendet. Dafür werden zunächst die zusammenhängenden Hautbereiche aus der Skinmap extrahiert. Regionen, die keine typisch menschlichen Formen aufweisen (z.B. rechteckige oder elliptische) werden verworfen. Aufgrund der Annahme, dass das Objekt des Interesses in der Bildmitte dargestellt ist, wird die verbleibende Skinmap in ein Kachelmuster aus 3x3 Quadraten eingeteilt. Enthält das mittlere Quadrat weniger als 29% Hautpixel oder wird auf dem Bild ein Gesicht identifiziert, welches mehr als 38% aller Hautpixel enthält, wird das entsprechende Bild als Nicht-Nacktbild klassifiziert. Im letzten Schritt klassifiziert ein ML-Algorithmus das Bild anhand folgender Merkmale:

- Hautpixelanteil im Bild
- Anteil der größten Hautregion an der Menge aller Hautpixel
- Hautpixelanteil in einem die drei größten Hautregionen umspannenden Polygon
- Verhältnis zwischen Fläche und Durchmesser der größten Hautregion
- Verhältnis der Fläche der größten Hautregion zu dem kleinsten sie einschließenden

Rechteck (Bounding Box)

- Anteil der zweitgrößten Hautregion an der Menge aller Hautpixel
- Verhältnis der Fläche der zweitgrößten Hautregion zu ihrer Bounding Box

Der Ansatz von Platzer et al. wurde auf dem Compaq-Bilddatensatz mit insgesamt 13.633 Bildern evaluiert. Während die Nacktheitserkennung eine TPR von 82,3% und eine FPR von 11,4% erzielte, konnte mit dem Pornografiedetektor eine Accuracy von 91,9% erreicht werden.

- Im Jahr 2017 erzielte Roheda eine Accuracy von 95,6% bei der Erkennung von Hautpixeln mit einem Klassifikationsverfahren, bei welchem die Farb- und Sättigungswerte benachbarter Pixel mit berücksichtigt wurden [46]. Die Bilder wurden dazu im HSV-Farbraum dargestellt. Um dieses Ergebnis zu erreichen, wurden etwa 20.000 Bildausschnitte als Trainingsmaterial verwendet. Im Vergleich zu bisherigen farbwertbasierten Hauterkennungsverfahren ist dies eine beachtliche Klassifikationsgenauigkeit.

  Jedoch sollte hier angemerkt werden, dass die zur Evaluation verwendete Bilddatenbank »color FERET Database«[4], welche zur Evaluation von Gesichtserkennungsverfahren erstellt wurde, Aufnahmen von Personen vor einem hellen und homogenen Hintergrund enthält, sodass eine Abgrenzung von Hautpixeln zu Nicht-Hautpixeln deutlich leichter zu erreichen ist als bei Aufnahmen unter Alltagsbedingungen, etwa mit verschiedenen Hintergrundobjekten, die den Klassifikationsprozess in die Irre führen können.

5.1.2 Deep-Learning-Modelle

Lange Zeit galt im Bereich der Computer Vision, also der automatischen Bilderkennung durch Algorithmen, die klassische Herangehensweise mit fest definierten Objektmerkmalen als Standard. Seit jedoch im Jahre 2012 der Bildklassifikationswettbewerb »ILSVRC«[5] (ImageNet Large Scale Visual Recognition Competition) erstmals durch ein Convolutional Neural Network (CNN) gewonnen wurde [47], intensivierte sich die Forschungsarbeit im Bereich CNNs bzw. Deep Learning schlagartig. Nur drei Jahre später, im Jahre 2015, verkündete »Microsoft«, mit einem CNN erstmals die menschliche Erkennungsrate von Objekten auf Bildern übertroffen zu haben [48].

Da die Erkennung von nackten Körpern mit denselben Prinzipien realisierbar ist, wie die gewöhnliche Bilderkennung, ließen erste Lösungen zur Nacktbilderkennung auf Basis von CNNs nicht lange auf sich warten. Aufgrund der großen Nachfrage nach Maßnahmen zur Filterung von pornografischem Material im Internet existieren bereits einige freie sowie kommerziell angebotene Verfahren zur Nacktbild- bzw. Pornografieerkennung, die mithilfe von CNNs sehr hohe Trefferraten erzielen.

4 http://sit4.me/colorferet, Juli 2018

5 http://sit4.me/LSVRC, Juli 2018

Diese Lösungen lassen sich als Black Boxen verwenden und sind untereinander austauschbar. Die entsprechenden Modelle führen für ein gegebenes Bild eine binäre Klassifikation durch, wobei die positive Klasse einem Nacktbild und die negative Klasse einem Nicht-Nacktbild entspricht. Anstatt allerdings das Bild fest einer der beiden Klassen zuzuordnen, werden lediglich die Wahrscheinlichkeitswerte ausgegeben, die das Modell für die Klassen errechnet hat. So könnte die Ausgabe zu einem Bild z.B. lauten: NSFW: 0,87; SFW: 0,13[6]. Das Modell würde das Bild also mit 87%iger Sicherheit als Nacktbild klassifizieren, während es die Wahrscheinlichkeit, dass es sich um ein Nicht-Nacktbild handelt, mit 13% beziffert. Der Wahrscheinlichkeitswert für die eine Klasse ist stets die inverse Wahrscheinlichkeit der anderen Klasse, sodass die Summe der beiden stets den Wert 1 (bzw. eine Wahrscheinlichkeit von 100%) ergibt. Üblicherweise wird das Bild dann derjenigen Klasse mit dem höchsten Wahrscheinlichkeitswert (d. h. größer als 0,5) zugeordnet. Diese strikte Zuordnung wird von den Anbietern jedoch bewusst nicht vorgenommen, um den Anwendern ihrer Lösungen die Wahl eines geeigneten Schwellwerts selbst zu überlassen. Dadurch lassen sich Fehlertoleranzen für Falsch-Positiv- und Falsch-Negativ-Klassifizierungen individuell einstellen. In manchen Anwendungsfällen, etwa Jugendschutzfilter, kann es beispielsweise besonders wichtig sein, dass keine Nacktbilder versehentlich als Nicht-Nacktbilder durchgehen, weshalb hier ein niedrigerer Schwellwert gewählt werden kann. Verfügbare auf Deep Learning bzw. CNNs basierende Modelle zur Nacktbilderkennung sind nachfolgend aufgelistet:

Name	**Anbieter**	**kommerziell**
Sightengine Nudity Detection API	Sightengine	ja
Clarifai NSFW Model	Clarifai	ja
Nude Detect	Netspark	ja
Yahoo Open NSFW Model	Yahoo	nein

Tabelle 5.1: Verfügbare Nacktbilderkennungsverfahren basierend auf Deep Learning.

Während es sich bei Sightengine[7], Clarifai[8] und Netspark[9] um kommerzielle Anbieter handelt, ist das Yahoo Open NSFW Model die bislang einzige quelloffene professionelle Lösung. Letzteres kann direkt in die eigene Software integriert und unbegrenzt zur Bildklassifikation genutzt werden.

5.1.3 Verschlagwortung von Bildern

Im Gegensatz zur bloßen Erkennung von Nacktbildern sind CNNs auch dazu geeignet, deutlich feinere Klassifikationsmodelle zu lernen. Mit entsprechenden Trainingsdaten lassen sich so auch Modelle trainieren, die beispielsweise erkennen, ob auf einem Nacktbild ein erkennbares Gesicht zu sehen ist oder ob die abgebildete Person minderjährig zu sein scheint. Werden einem Eingabebild mehrere Eigenschaften zugewiesen, spricht man häufig auch von Verschlagwortung. Bei diesen Eigenschaften kann es sich sowohl um abgebildete Objekte handeln als auch um abstrakte Dinge wie eine Stimmung, die das Bild vermitteln soll. Üblicherweise wird die Verschlagwortung

6 Im englischen Sprachgebrauch sind für Nacktbilder bzw. Nicht-Nacktbilder die Termini NSFW (Not Safe For Work) und SFW (Safe For Work) gebräuchlich.

7 https://sightengine.com

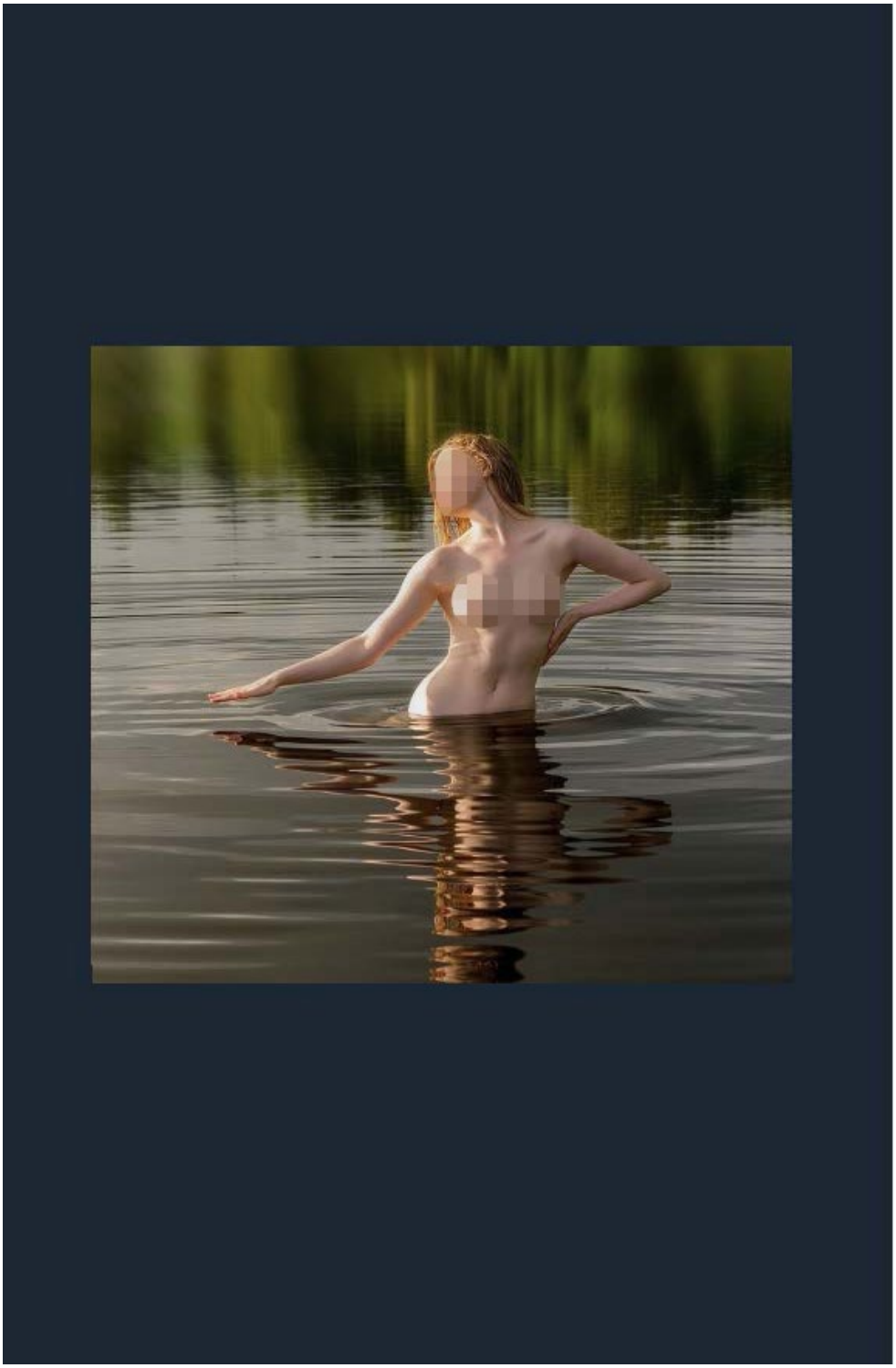
8 https://clarifai.com

9 http://www.nudedetect.com

von Bildern in der Bildersuche (»Image Retrieval«) eingesetzt. Gibt ein Nutzer einen Begriff, wie z.B. »Sonnenuntergang« ein, bekommt er sämtliche Bilder einer Datenbank angezeigt, welche zuvor mit dem entsprechenden Begriff verschlagwortet worden sind. Sie könnte jedoch auch bei der Sexting-Erkennung eingesetzt werden, etwa um nur bestimmte Arten von Sexts zu filtern.

Das von Saito und Matsui entwickelte CNN »Illustration2Vec« (I2V) [49] wurde mit knapp 1,3 Mio. aus dem Internet gecrawlten Bildern trainiert. Klassenlabels wurden aus entsprechenden Beschreibungstexten bzw. Metadaten zu den Bildern entnommen. Da eine große Zahl der Bilder Pornografie enthielt, ist das gelernte Modell nicht nur in der Lage, Nacktbilder zu erkennen, sondern darüber hinaus auch entblößte Körperregionen, Aufnahmeperspektiven und sexuelle Handlungen. Obwohl es sich bei der Trainingsmenge vielfach um Animes, d. h. Bilder aus japanischen Zeichentrickfilmen, handelte, lässt es sich auch auf Fotos anwenden. Auch die zuvor in Abschnitt 5.1.2 erwähnten Anbieter Sightengine und Clarifai stellen neben der reinen Nacktbilderkennung eine Vielzahl weiterer Modelle zur Verfügung, mit Hilfe derer sich verschiedene Bildinhalte automatisch identifizieren lassen. Die Klassifizierung eines Beispielbildes durch die Demo-Version von Clarifai ist in Abbildung 5.5 zu sehen. Es werden lediglich Klassenzugehörigkeiten ab einem bestimmten Wahrscheinlichkeitswert angezeigt, von denen die meisten eindeutig zutreffen, wie z.B. »water« oder »nude«.

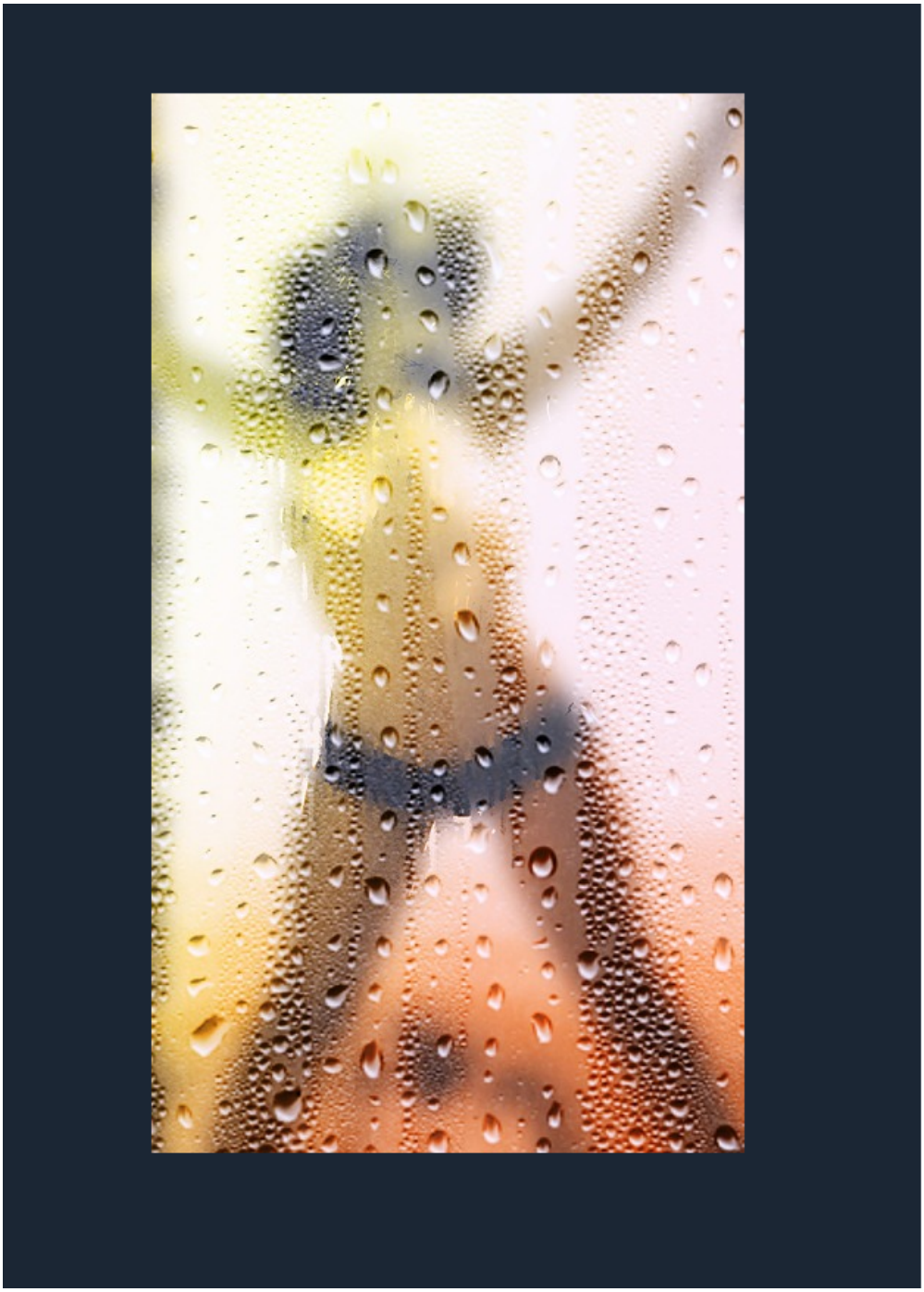
Das Trainieren von Modellen zur Erkennung bestimmter Bildinhalte stellt heutzutage – ausreichendes Bildmaterial sowie Hardware-Ressourcen vorausgesetzt – kein großes Problem mehr dar. Jedoch ist die künstliche Intelligenz solcher Modelle keineswegs mit der eines Menschen vergleichbar. Im Gegensatz zum Menschen, der aufgrund seines Hintergrundwissens auch in sehr verrauschten oder abstrakten Daten noch Muster erkennen kann, müssen Modelle des maschinellen Lernens stets auf ein fest definiertes und eng abgestecktes Anwendungsszenario hin trainiert werden. Konfrontiert man ein solches Modell jedoch mit Daten aus einem fremden Kontext oder stark verrauschten Daten, so ist es meist nicht in der Lage, eine sinnvolle Klassifizierung durchzuführen. Abbildung 5.6 zeigt einen solchen Fall, in dem ein unscharfes Bild klassifiziert wird. Dass es sich um eine nackte Frau hinter einer kondenswasserbehafteten Glaswand handelt, ist für den menschlichen Betrachter klar erkennbar. Das Clarifai-Modell hingegen erkennt in dem Bild einen völlig falschen Kontext, was zur Fehlklassifikation führt.



PREDICTED CONCEPT	PROBABILITY
water	0.998
lake	0.986
girl	0.983
recreation	0.975
nature	0.974
river	0.969
summer	0.968
leisure	0.953
reflection	0.940
wet	0.935
pool	0.932
nude	0.931
outdoors	0.909
child	0.904
beach	0.903
fun	0.902
fair weather	0.890
woman	0.887
one	0.870
swimming	0.854

Abbildung 5.5:

Klassifizierung eines Beispielbildes durch die Clarifai Demo-Abfrage. Auf der rechten Seite sind die Klassen, denen das Modell das Bild mit hohem Wahrscheinlichkeitswert zuordnet. Das Bild wird der Klasse »nude« mit dem Wert 0,931 zugeordnet, d. h. mit 93,1%iger Wahrscheinlichkeit als Nacktbild klassifiziert. Quelle Bild: CC0 Pixabay.



PREDICTED CONCEPT	PROBABILITY
text	0.983
business	0.981
symbol	0.971
illustration	0.971
internet	0.963
designing	0.962
signalise	0.958
communication	0.953
achievement	0.941
motivation	0.932
education	0.922
strategy	0.921
creativity	0.920
technology	0.918
leadership	0.914
management	0.911
solution	0.898
control	0.898
teamwork	0.891
cloud	0.891

Abbildung 5.6:

Klassifizierung eines Beispielbildes durch die Clarifai Demo-Abfrage. Auf der rechten Seite sind die Klassen, denen das Modell das Bild mit hohem Wahrscheinlichkeitswert zuordnet. Das Clarifai-Modell erkennt in dem Bild einen völlig falschen Kontext, was zur Fehlklassifikation führt. Quelle Bild: CC0 Pixabay.

5.1.4 Automatische Altersbestimmung

Die bisher besprochenen Technologien können zwar dazu genutzt werden, um pornografisches oder erotisches Bildmaterial automatisiert zu erkennen, jedoch spielt das Alter der abgebildeten Personen dabei keine Rolle. Bei der Implementierung einer automatisierten Sexting-Erkennung, die den Kinder- und Jugendschutz zum Ziel hat und ein unachtsames Versenden von Nacktaufnahmen minderjähriger Internetnutzer verhindern soll, könnte eine entsprechende Software zur Nacktbilderkennung um einen Mechanismus zur Altersbestimmung erweitert werden.

Im Gegensatz zur Personen- oder Nacktbilderkennung handelt es sich bei der automatischen Altersbestimmung von Personen um eine deutlich schwierigere Aufgabe, da das Alter kennzeichnende visuelle Merkmale einer Person von einer Vielzahl unterschiedlicher Faktoren (z.B. Genetik, Ernährung, Drogenkonsum, Kleidung) abhängen [50]. Hinzu kommen die im Bereich Computer Vision generell geltenden Einflussfaktoren wie Belichtung oder Perspektive. Nichtsdestotrotz ist auch dieses Thema Gegenstand aktueller Forschungsarbeiten. Es sind dabei zwei Bereiche zu unterscheiden: Während ein Forschungszweig die Bestimmung des realen Alters (»real age estimation«) zum Ziel hat, beschäftigt sich ein anderer mit der Bestimmung des wahrgenommenen Alters (»apparent age estimation«), also dem Alter, das die meisten menschlichen Betrachter der abgebildeten Person ebenfalls zuschreiben würden. Letztgenannter Forschungszweig ist für die Erkennung von Sexts Minderjähriger folglich irrelevant. Die meisten Forschungsansätze setzen CNNs oder andere ML-Algorithmen ein, welche das Alter anhand von Gesichtsmerkmalen bestimmen. Die Bestimmung des Alters kann auf zwei Arten durchgeführt werden:

- Klassifikation nach Altersgruppen: Anstatt das exakte Alter einer Person vorherzusagen, liefert der Algorithmus ein Intervall (z.B. »15 bis 20 Jahre«). Um Sexts von Minderjährigen zu erkennen, ist diese Art der Altersbestimmung ebenfalls ungeeignet, da die Altersintervalle i.d.R. nicht der Einteilung in minder- und volljährig entsprechen.

- Regression: Die Regression erlaubt eine jahresgenaue Altersbestimmung. Im Gegensatz zur Klassifikation mit voneinander unabhängigen Klassen lernt ein Regressionsverfahren eine kontinuierliche Funktion. Handelt es sich bei der Zielgröße wie im Falle der Altersbestimmung um einen einfachen Zahlenwert, ist die Regression i.d.R. geeigneter als die Klassifikation. Da es anders als bei der Klassifikation keine Einteilung in korrekte und falsche Zuordnungen gibt, wird die Güte von Regressionsverfahren üblicherweise durch den »mittleren absoluten Fehler« (»Mean Absolute Error«, kurz: MAE) angegeben.

Rothe et al. verwendeten ein »Deep EXpectation« (DEX) genanntes Regressionsverfahren, mit dem sie auf unterschiedlichen Testmengen einen MAE zwischen 2,68 und 4,79 Jahren erreichten [51]. In einem Vorverarbeitungsschritt wird dabei die Position und Ausrichtung des Gesichts normalisiert und der Hintergrund entfernt, sodass Eingabebilder in einer einheitlichen Form vorliegen. Die Regression zur Altersbestimmung wird dann durch ein CNN durchgeführt. In [52] stellten Ranjan et al. ebenfalls ein CNN-basiertes Regressionsverfahren vor, das einen MAE von 2,0 Jahren erzielte.

5.1.5 Vergleich der vorgestellten Verfahren

Die zuvor erläuterten Verfahren, welche im Rahmen einer Sexting-Erkennung genutzt werden können, basieren auf unterschiedlichen Prinzipien und weisen somit auch unterschiedliche Vor- und Nachteile auf. Diese werden nachfolgend noch einmal zusammenfassend einander gegenübergestellt.

Farbwertbasierte Hauterkennung: Schneiden farbwertbasierte Verfahren zur Hauterkennung auf geschlossenen, homogenen Bildmengen mit einer hohen farblichen Differenz zwischen Haut- und Nicht-Hautpixeln gut ab, so offenbaren sich ihre Schwächen bei der Anwendung auf eine große Menge von heterogenen Bildern. Überschneiden sich die Farbspektren von Hautpixeln und Nicht-Hautpixeln, kann eine saubere Trennung zwischen beiden anhand der Farbwerte allein nicht mehr gelingen. So kann es beispielsweise dazu kommen, dass Sonnenuntergangsbilder als Nacktbilder klassifiziert werden, da sie eine hohe Anzahl von Pixeln im rötlichen Farbspektrum enthalten, in dem sich auch viele Hauttöne wiederfinden [44]. Diese Verwechslungsgefahr besteht ebenso bei Bildern, die Holz, Sand und andere Materialien mit hautähnlicher Farbe enthalten [44].

Da die menschliche Haut je nach Ethnie, Hauttyp und Bräunungsgrad ein breites Spektrum im rot-gelben Farbspektrum einnimmt, ist eine klare Abgrenzung von Hautpixeln zu Nicht-Hautpixeln unmöglich. Unterschiedliche Belichtungsverhältnisse und Kameraeigenschaften vergrößern dieses Spektrum sogar noch und erschweren die Erkennung zusätzlich. Handelt es sich hingegen um Schwarz-Weiß-Bilder, ist eine Erkennung von Hautpixeln nur noch über den Grauwert möglich. Hier ist der Informationsverlust bereits so groß, dass eine vernünftige Erkennung quasi unmöglich ist, weshalb sich die Hauterkennung prinzipiell nur für Farbbilder eignet. Eine weitere Schwierigkeit besteht in der Tatsache, dass ein hoher Hautpixelanteil nicht ausschließlich bei Nacktbildern vorliegt. Während ein Urlaubsfoto am Strand durchaus viel nackte Haut enthalten kann, kann auf der anderen Seite ein Foto mit entblößten Brüsten deutlich weniger Hautpixel aufweisen.

Die genannten Probleme machen die Hauterkennung – zumindest als alleinstehende Lösung – zu keiner geeigneten Methode für die Sexting-Erkennung. Nichtsdestotrotz ist ein zuverlässiges Hauterkennungsverfahren im Rahmen der Sexting-Erkennung als Filter in einem Vorverarbeitungsschritt denkbar. Ein solcher Filter kann die Entscheidungen nachfolgender Verfahren zuverlässiger machen, etwa indem irrelevante Bildinhalte entfernt werden, welche einen ML-Algorithmus (z. B. ein CNN) zu Fehlinterpretationen verleiten könnten.

Deep-Learning-Modelle: Im Gegensatz zu den in Abschnitt 5.1.1 vorgestellten, auf der statistischen Verteilung von Farbwerten beruhenden Verfahren erfordern CNNs sehr viele Trainingsbeispiele (mehrere tausend bis Millionen). Das macht das Training von Deep-Learning-Modellen einerseits sehr rechenintensiv und erfordert geeignete Hardware[10]. Andererseits sind CNNs in der Lage, aus der Vielzahl verschiedener Trainingsbeispiele Merkmale der zu erkennenden Inhalte zu erlernen. Ist die Trainingsmenge ausreichend groß und divers, kann ein Overfitting des resultierenden Klassifikationsmodells vermieden werden, sodass die gelernten Merkmale auf Test- und Produktivdaten verallgemeinerbar sind.

10 Zum Training von NNs werden üblicherweise leistungsstarke Grafikkarten eingesetzt

Ein weiterer Vorteil ist, dass die Merkmale nicht wie bei klassischen ML-Algorithmen durch den Entwickler vorgegeben werden müssen, sondern von dem CNN während des Trainingsprozesses selbstständig erlernt werden. Der Programmieraufwand für ein CNN-basiertes Verfahren zur Sexting-Erkennung wäre dadurch nicht wesentlich höher als für ein farbwertbasiertes Verfahren. Einen relativ hohen Zeitaufwand dagegen stellen Datengewinnung und -bereinigung sowie Trainings- und Optimierungsprozesse dar.

5.2 Erkennung von Cybergrooming

Auch im Rahmen der Aufdeckung von Cybergrooming in Chat-Foren und sozialen Netzwerken kann auf ML-Verfahren zurückgegriffen werden. Hierbei spielen insbesondere Methoden aus dem Bereich der Autorschaftsanalyse eine tragende Rolle, welche ein interdisziplinäres Forschungsfeld darstellt und sich aus den Bereichen (Computer-)Linguistik bzw. Philologie, Psychologie, Mathematik sowie Informatik zusammensetzt. Zwei Unterdisziplinen innerhalb der Autorschaftsanalyse, die im Rahmen von Cybergrooming Anwendung finden, wären z.B. Autorenprofiling sowie die Autorschaftsverifikation, die in diesem Abschnitt vorgestellt werden.

Autorschaftsanalyse: Das Internet bietet eine Reihe von Möglichkeiten der Anonymität an, sodass sich Nutzer mit krimineller Energie oftmals sicher fühlen, wenn sie Straftaten begehen. Soll etwa die eigene IP-Adresse verborgen werden, bieten sich kostenlose Anonymizer an, die anonyme Verbindungen über Proxyserver ermöglichen. Soll dagegen das Abhören einer Verbindung zwischen zwei Parteien verhindert werden, so können entsprechende Messenger wie z.B. Threema oder Signal verwendet werden. Diese verfügen über Verschlüsselungsprotokolle, die das Abhören zwar nicht unmöglich machen, aber zumindest erschweren. Wollen Nutzer dagegen öffentlich kommunizieren (beispielsweise über soziale Netzwerke, Foren oder Chats), gleichzeitig jedoch anonym bleiben, so bieten sich Pseudonyme an, um die eigene Identität in öffentlichen Räumen zu verschleiern.

Im Gegensatz zur Verschlüsselung haben Pseudonyme allerdings einen Haken, sie schützen zwar Vor- und Nachnamen des Nutzers, verhindern aber nicht, dass aus unverschlüsselten Texten (z.B. E-Mails, Facebook-Kommentaren oder Tweets) Spuren bestimmt werden können, die eine Deanonymisierung der jeweiligen Verfasser erlauben. Ein Forschungsgebiet, welches sich mit der Untersuchung anonymer Verfasser bzw. Autoren von Texten auseinandersetzt, ist die **Autorschaftsanalyse**. Unter diesem Oberbegriff wird ein Verbund verschiedener Unterdisziplinen verstanden, die sich mit der Attribution[11], Verifikation oder Charakterisierung von Autoren beschäftigen. Die enorme Menge digitaler Texte macht jedoch eine manuelle Autorschaftsanalyse (etwa durch einen Sachverständigen) oftmals unpraktikabel, da diese mit sehr viel Aufwand und Kosten verbunden ist. So ist tatsächlich in den letzten Jahren zu beobachten, dass Forensik-Unternehmen, polizeiliche Institutionen und große Konzerne (z. B. Wirtschaftsprüfungsgesellschaften) auf das Potenzial der (semi-)automatisierten Autorschaftsanalyse setzen, um den Analyseprozess zu beschleunigen. Die automatische Autorschaftsanalyse im heutigen Zeitalter basiert nahezu ausschließlich auf ML-Konzepten und -Verfahren, um Schreibstile oder generell Sprachmuster von Autoren zu modellieren.

11 In der Literatur auch als Identifikation, Erkennung oder Zuordnung bezeichnet.

Im Kontext von Cybergrooming spielen hauptsächlich die beiden Disziplinen »Autorenprofiling« und »Autorschaftsverifikation« eine wichtige Rolle. In den folgenden zwei Unterabschnitten wird ausführlich auf beide Disziplinen eingegangen und der Bezug zu Sexting hergestellt.

5.2.1 Autorenprofiling

Autorenprofiling (kurz **AP**) ist eine der bekanntesten Unterdisziplinen der Autorschaftsanalyse, bei der es darum geht, Texte hinsichtlich spezifischer Eigenschaften von Autoren zu klassifizieren. Die Spanne der möglichen Autoreigenschaften, die AP fokussiert, erstreckt sich von »Alter«, »Berufstätigkeit«, »Bildungsniveau«, »Geschlecht«, »Händigkeit (Linkshändigkeit / Rechtshändigkeit)«, »Muttersprachlichkeit« bis hin zu »Persönlichkeit«. Aus der Sicht des MLs repräsentiert AP somit ein n-äres Klassifikationsproblem, wobei die Ausprägungen innerhalb jeder einzelnen Autoreneigenschaft die Klassen darstellen. So hat etwa die Eigenschaft »Geschlecht« beispielsweise zwei[12] Ausprägungen (männlich und weiblich), während bei der Eigenschaft »Bildungsniveau« mehrere Ausprägungen (z.B. ohne, niedrige, mittlere oder hohe Schulbildung) möglich sind.

Bezug zu Cybergrooming

Im Rahmen von Cybergrooming spielt insbesondere die Autoreigenschaft »Alter« eine wichtige Rolle. Cyber-Groomer tarnen sich in entsprechenden Plattformen oftmals als Kinder oder Teenager, um so mit ihren Opfern ins Gespräch zu kommen. Mithilfe eines AP-Systems könnte man nun das ungefähre Alter eines Cyber-Groomers X bestimmen und entsprechend Alarm schlagen, wenn das System für X vorhersagt, dass dieser weder in den Altersbereich eines Kindes noch in den eines Teenagers fällt. Wenn darüber hinaus X sein falsches Alter (beispielsweise in einem Chat-Verlauf) preisgibt, könnte ein Abgleich des vorhergesagten Altes mit dem genannten Alter durchgeführt werden, um die Vorhersage des Systems zusätzlich zu stützen.

Während das Alter einer Person im Kontext von Cybergrooming von Bedeutung ist, erscheint die Eigenschaft »Geschlecht« weniger relevant. Hintergrund ist, dass die Gruppe der Cyber-Groomer nahezu ausschließlich aus männlichen Personen besteht, sodass kein Mehrwert zu erwarten ist.

Jenseits von Cybergrooming bietet sich AP für eine Reihe von anderen Anwendungsmöglichkeiten wie z. B. Marktsegmentierung auf Basis der ermittelten Autoreigenschaften an. Darüber hinaus kann AP als ein Filterungsschritt für andere Unterdisziplinen der Autorschaftsanalyse (z.B. die Autorschaftsattribution) dienen, sodass die initiale Menge potenzieller Autoren auf Eigenschaften reduziert werden, die mit den Autoreigenschaften des anonymen Dokuments übereinstimmen. Dadurch kann mitunter wertvolle Laufzeit eingespart werden.

12 Forschungsarbeiten im Kontext von AP haben bisher (nach unserem Kenntnisstand) nur die beiden Ausprägungen betrachtet. In der heutigen Zeit jedoch sind weitere Geschlechtsidentitäten etabliert. Facebook listet z.B. 60 verschiedene Geschlechtsidentitäten auf (siehe http://www.faz.net/aktuell/gesellschaft/facebook-60-auswahlmoeglichkeiten-fuer-geschlecht-13135140.html).

Arbeitsweise von Autorenprofiling-Verfahren

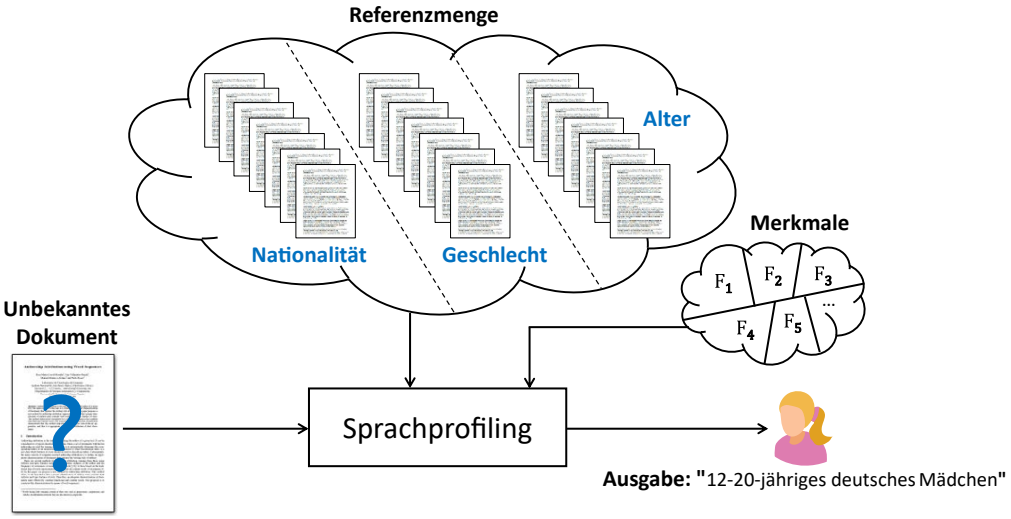


Abbildung 5.7: Schematische Darstellung eines Autorenprofiling-Systems.

Um eine Aussage über eine Autoreigenschaft hinsichtlich eines Textes zu treffen, muss zunächst eine geeignete Referenzmenge[13] mit Beispieltexten vorliegen. Die darin befindlichen Texte müssen entsprechend den Ausprägungen der betrachteten Autoreigenschaft gekennzeichnet sein (siehe Abbildung 5.7). Der Vorgang Texte (bzw. generell Daten) zu kennzeichnen, wird im Fachjargon »labeln« genannt und stellt einen Prozess dar, der i. d. R. von Menschen durchgeführt wird. Liegt eine Referenzmenge vor, so geht es im nächsten Schritt darum, eine geeignete Datenrepräsentation zu wählen, um die Texte maschinell weiterzuverarbeiten. Eine im AP häufig anzutreffende Form der Repräsentation stellen sogenannte »Vector Space Models« (kurz VSM) dar. Die Grundidee von VSMs ist es, zunächst verschiedene Features aus den Texten zu extrahieren, mit deren Hilfe sich die Ausprägungen der Autoreigenschaft beschreiben lassen. Hierbei kann auf eine Vielzahl von Features zurückgegriffen werden (siehe Tabelle 5.2)

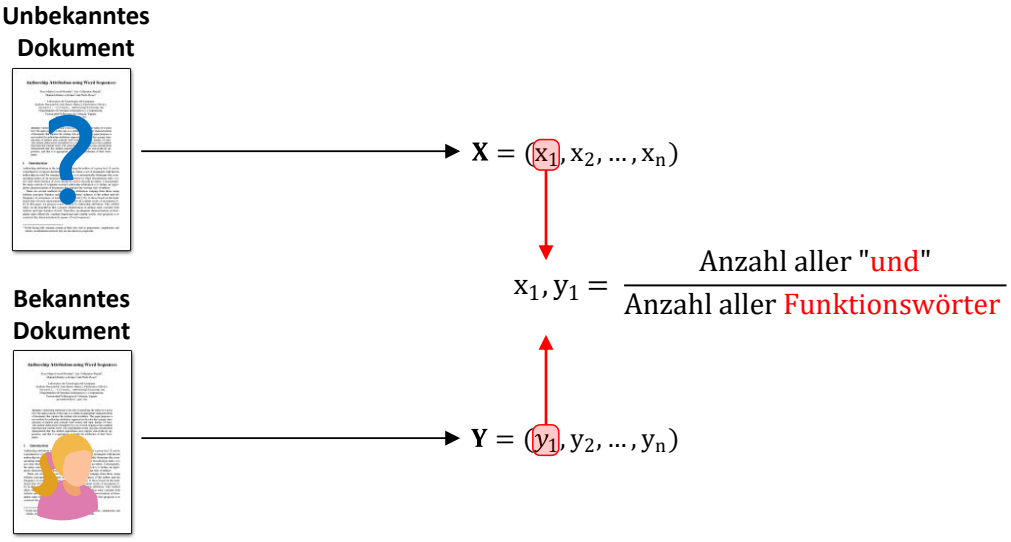


Abbildung 5.8: Konstruktion eines Feature-Vektors.

13 Im Rahmen von Klassifikationsproblemen, die Texte betrachten, werden diese Referenzmengen auch »Korpora« genannt.

Nachdem eine Auswahl von Features getroffen wurde, gilt es im nächsten Schritt, die Texte als sogenannte Feature-Vektoren numerisch auszudrücken. Abbildung 5.8 zeigt, wie aus zwei Texten zwei reellwertige Feature-Vektoren X und Y konstruiert werden. Anschließend können diese in einem hochdimensionalen, metrischen Vektorraum repräsentiert und hinsichtlich ihrer Ähnlichkeiten zueinander in Bezug gebracht werden. Dies geschieht oftmals mithilfe von Ähnlichkeitsfunktionen[14]. Ein bekannter Vertreter dieser Funktionen ist etwa die Kosinus-Ähnlichkeit cos(α), die wie folgt definiert ist:

$$\cos(\alpha) = \frac{X \cdot Y}{||X|| \cdot ||Y||} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (x_i)^2}}$$

Hierbei bezeichnet α den Winkel zwischen den beiden Vektoren X und Y. Die resultierenden Werte liegen im Bereich zwischen 0 und 1, wobei Werte nahe 1 auf eine hohe Ähnlichkeit und Werte nahe 0 auf eine hohe Unähnlichkeit deuten. Neben solchen »distanzbasierten Verfahren« werden in AP oftmals auch klassische ML-Klassifikationsverfahren angewendet. Bekannte Vertreter sind hier SVMs, Naive Bayes, Logistic Regression oder auch Random Forests. Unabhängig davon, welches ML-Verfahren angewendet wird, stellen VSMs i. d. R. die erste (und einfachste) Wahl dar.

Erfolgsaussichten seitens der Forschung: Auch wenn in den letzten Jahren eine starke Zunahme an AP-Verfahren seitens der Forschung beobachtet werden kann, ergab unsere Suche nach wissenschaftlichen Artikeln in Bezug auf AP im deutschsprachigen Raum nahezu keine Treffer[15]. Demgegenüber fanden sich sehr viele Fachartikel zu AP in anderen Sprachen, zumeist Englisch. Im Folgenden erläutern wir zwei Ansätze für die englische und einen Ansatz für die niederländische Sprache, die zumindest Deutsch ähnelt.

Ein aktueller AP-Ansatz, der neben dem Alter einer Person auch die Autoreigenschaften »Geschlecht« und »Berufszweig« vorhersagen konnte, wurde z.B. von Jiang et al. in [55] vorgestellt. Ihr Ansatz basierte auf einer Kombination von zwei NN-Architekturen (CNNs und LSTMs) sowie einem Wahrscheinlichkeitsmodell namens »Latent Dirichlet allocation«. Hinsichtlich der Repräsentation der Texte wählten die Forscher zeichen- und wortbasierte Merkmale, wobei sogenannte Einbettungen (engl. »embeddings«) zum Einsatz kamen, in denen jedes einzelne Feature als ein hochdimensionaler Embedding-Vektor dargestellt wird. Nach der Konstruktion der Einbettungen wurden die einzelnen Vektoren sequenziell in das zusammengesetzte NN eingegeben, um dieses zu trainieren. Bei dieser Vorgehensweise wird die Reihenfolge der einzelnen Merkmale berücksichtigt. Dadurch können syntaktische Muster[16] in den Texten identifiziert werden, mit denen etwa die Schreibkompetenz des Verfassers bewertet und damit dessen Alter geschätzt werden kann. Hinsichtlich der Autoreigenschaft Alter erzielten Jiang et al. eine Erkennungsgenauigkeit von 79.6%, wobei sie als Datensatz den »Blog Authorship Corpus« verwendeten.

14 Ein umfassender Überblick dazu findet sich in [54].

15 Es fanden sich lediglich einige Webseiten wie z. B. Blogs, die das Thema aufgriffen, sowie einige Bücher (darunter eins von einem bekannten Profiler aus Deutschland).

16 Beispielsweise solche Features, die für die Konstruktion von Sätzen benötigt werden.

Feature-Kategorie	Kurzbeschreibung / Beispiele
Interpunktionszeichen	(,), [,], !, ?, ;, :, . . .
Buchstaben	A-Z, Ä, Ö, Ü, a-z, ä, ö, ü, ß
Buchstaben n-Gramme	Textbeispiel → {Te, ex, xt, tb, be, . . .}; n = 2
Präfixe	Textbeispiel (Vorsilbe)
Infixe	Textbeispiel (innerer Wortbestandteil)
Suffixe	Textbeispiel (Nachsilbe)
Funktionswörter	Artikel (der, das, einer, eines, . . .), Konjunktionen (und, oder, . . .), . . .
Anglizismen	Wortentlehnungen (z.B. Mail, Newsletter, Chat, Meeting, Update, . . .)
Neologismen	Kunstwörter (z.B. Abmahnwelle, Nerd, googeln, verschlimmbessern, . . .)
Wort n-Gramme	Ein kleines Textbeispiel → {(Ein kleines), (kleines Textbeispiel)}; n = 2
Kollokationen	Häufig vorkommende Wortverbindungen (z. B. starker Tobak)
Wortarten	Adjektive, Interjektion, Numerale, Substantive, . . .
Wortart n-Gramme	(Artikel-Adjektiv-Nomen), (Pronomen-Nomen-Artikel), . . .
Phrasen/Redewendungen	Redensarten (z.B. aus dem Nähkästchen plaudern)
Satz-Anfänge/Endungen	Satzanfang(Nomen), Satzende(finites Verb), . . .
Wort-Komplexität	Wörter bestimmter Länge, Wörter mit x Vokalen
Satz-Komplexität	Sätze bestimmter Länge, Vorfeld/Mittelfeld/Nachfeld-Komplexitäten, . . .
Text-Komplexität	Funktionswort-Dichte, Koreferenzketten, . . .
Verständlichkeitsindizes	Gunning Fog Readability Index, Flesch-Kincaid Reading Ease, . . .
Grammatikalische Fehler	Falsche Verwendung von Genus, Kasus, Kommata, . . .

Tabelle 5.2: Eine Auswahl von 20 Feature-Kategorien, die sich für verschiedene Unterdisziplinen im Rahmen der Autorschaftsanalyse eignen. Entnommen aus [53].

Schler et al. [56] sind Forscher, die den »Blog Authorship Corpus« kompiliert und Forschern weltweit zur Verfügung[17] gestellt haben. Der Korpus, der insgesamt 681.288 Postings von 19.320 Bloggern umfasst, gilt als De-facto-Standard auf dem Gebiet des APs und wurde auch in anderen Unterdisziplinen der Autorschaftsanalyse häufig verwendet (z.B. [57, 58, 59]). Basierend auf diesem Korpus untersuchten Schler et al. in ihrer Studie [56] die Erfolgsaussichten bzgl. der beiden Autoreigenschaften »Geschlecht« und »Alter«. Dabei verwendeten sie ein ML-Verfahren namens Multi-Class Real Winnow, um die Blog-Postings entsprechend zu klassifizieren. Im Gegensatz zu dem NN-basierten Ansatz von Jiang et al., verfolgten Schler et al. jedoch einen anderen, sogenannten **Bag-of-Features**-Ansatz, bei dem Features unabhängig voneinander betrachtet werden. Trotz dessen Nachteils, dass einzelne Merkmale manuell definiert werden müssen, stellt Bag-of-Features, innerhalb, aber auch jenseits von AP den am häufigsten verwendeten Ansatz der Datenrepräsentation dar. Schler et al. fokussierten sich auf die folgenden Merkmale: Fremdwörter (Wörter, die nicht im Wörterbuch vorkommen), Wortklassen (engl. »part of speech«), Funktionswörter sowie Hyperlinks, kombiniert mit Wort-Unigrammen, die den höchsten Informationsgehalt aufweisen. Hinsichtlich ihrer Evaluierung betrachteten die Autoren drei mögliche Altersklassen Age_{13-17} (13-17 Jahre), Age_{23-27} (23-27 Jahre) und Age_{33-42} (33-42 Jahre) und erreichten eine Gesamtgenauigkeit von 76,2%. Die Altersklasse Age_{13-17} konnte von Age_{23-27} mit 87,3% und von

17 Abrufbar unter http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm

$Age_{33–42}$ mit über 96% Genauigkeit unterschieden werden. Allerdings wurden viele Beispiele aus $Age_{33–42}$ irrtümlich als $Age_{23–27}$ klassifiziert.

Ein weiterer AP-Ansatz, der statt Englisch die niederländische Sprache fokussiert, wurde von Nguyen et al. in [60] präsentiert. Hierbei wurden ausschließlich Twitter-Tweets von insgesamt 3.000 Autoren untersucht, wobei lineare Modelle (Logistic und Linear Regression) als ML-Verfahren dienten. Ähnlich zu Schler et al. [56] betrachteten Nguyen et al. ebenfalls Altersklassen {$Age_{<20}$, $Age_{20–40}$, $Age_{>40}$} sowie diskrete Lebensabschnitte wie {Schüler, Student, berufstätig}. Zusätzlich dazu wurde das Alter auch als kontinuierlicher[18] Wert vorhergesagt. Als Merkmale verwendeten Nguyen et al. Wörter, die mindestens zehnmal in den Trainingsdaten vorkamen. Hinsichtlich der diskreten Altersklassen und Lebensabschnitte erzielten die Autoren einen hohen F1-Wert[19] von 0.863. Bzgl. der Vorhersage des Alters als kontinuierliche Zahl wurde eine Abweichung von weniger als 4 Jahren erreicht. Laut Nguyen et al. [60] erzielte ihr Ansatz bessere Ergebnisse als 17 menschliche Rater, welche für das Experiment herangezogen wurden. Neben ihren Ergebnissen haben die Autoren auch einige Beobachtungen festgehalten. Ihrer Meinung nach sollte die Eigenschaft Geschlecht parallel zum Alter mit betrachtet werden, da diese einen starken Einfluss bzgl. dem Alter einer Person hat. Weiterhin haben sie festgestellt, dass sprachliche Änderungen hauptsächlich bei jüngeren Personen vorkommen und dass bei einem Alter jenseits von ca. 30 Jahren die betrachteten Features kaum eine Varianz aufweisen, um die Autoren voneinander zu trennen.

Erfolgsaussichten seitens der Wirtschaft: Jenseits der Forschung sind in jüngster Zeit erste AP-Anwendungen in der Wirtschaft zu beobachten, die das Thema Cybergrooming aufgreifen. Eine der wenigen von ihnen ist die Smartphone-App Privalino, auf die im Folgenden detailliert eingegangen wird.

Privalino ist eine telegramkompatible Messenger-App, die Kinder zwischen 6 und 10 Jahren online vor Cybergrooming, Sextortion, Sexting und Cyber-Mobbing schützen soll[20]. Während das Kind online kommuniziert, überprüft die App den Sprachstil aller Nachrichten des Chat-Partners nach Auffälligkeiten. Hinweise, dass der Chat-Partner ein Erwachsener ist, liefern Satzkomplexität, Rechtschreibung und Wortschatz. Bei Gefahr durch Mobbing, Cybergrooming oder Sexting fragt Privalino das Kind, ob es den Chat-Partner persönlich kennt. Gleichzeitig werden die Eltern des Kindes per E-Mail benachrichtigt, damit diese den verdächtigen Chat-Partner ggf. blockieren und melden können.

Privalino erkennt Cyber-Groomer u. a. am Schreibstil, indem es beispielsweise die Lesbarkeit des Textes formal bestimmt. Hierfür wendet die App sogenannte Lesbarkeitsformeln wie etwa Flesch, Smog, Coleman Liau und Kincaid an. Der Lesbarkeitsindex ist ein Verfahren, mit dem die Textverständlichkeit bestimmt werden kann. Dieser sagt allerdings nichts über den Inhalt eines Textes aus. Generell sind alle Lesbarkeitsformeln sprach- und textgenrespezifisch. Das bedeutet, sie lassen sich nicht in unveränderbare Form auf deutschsprachige Texte anwenden und müssen vorher neu geeicht werden.

18 Analog zu unserem vorgestellten Ansatz.

19 Das F1-Maß kann Werte zwischen 0 und 1 annehmen.

20 https://www.privalino.de (Aufruf: 30.10.2018)

Der Flesch-Lesbarkeitsindex stammt von Rudolf Flesch [61] und misst, wie leicht ein Text aufgrund seiner Wort- und Satzstruktur lesbar und verständlich ist. Der Index ergibt in der Regel eine Zahl zwischen 0 und 100. Je höher der Wert, desto leichter verständlich ist der Text. Ein gut verständlicher Text weist Werte zwischen 60 und 70 auf. Die Flesch-Formel ist bei der Berechnung auf die englische Sprache abgestimmt und berechnet sich wie folgt:

$$FL = 206,835 - 84,6 * ASW - 1,015 * ASL$$

$$mitASW^{21} = \frac{\text{Silbenanzahl gesamter Text}}{\text{Anzahl Wörter im Text}}$$

$$undASL^{22} = \frac{\text{Anzahl Wörter im Text}}{\text{Anzahl Sätze}}$$

Da die Formel für die englische Sprache entwickelt wurde, eignet sich diese nicht für die deutsche Sprache, da deutsche Wörter im Schnitt länger sind als englische. Die Satzlänge ist in beiden Sprachen in etwa gleich lang. Auch die Bewertungsskala der Schweregrade ist nicht für die deutsche Sprache geeignet. Toni Amstad [62] hat die Formel auf die deutsche Sprache wie folgt übertragen.

$$FL = 180 - 58,5 * ASW \check{} ASL$$

Ein gut lesbarer deutscher Text hat einen Wert zwischen 40 und 60, »Allgemeine Geschäftsbedingungen« haben beispielsweise einen Wert von etwa 15 und sind somit schwieriger lesbar.

Der Kincaid-Lesbarkeitsindex [63] (Flesch-Kincaid-Grade-Level, kurz: FKGL) drückt die Lesbarkeit eines Textes in der Anzahl der Schuljahre aus, die ein Leser absolviert haben muss, um den Text zu verstehen. Eine Bewertung von 7 bedeutet z.B., dass ein Siebtklässler das Dokument verstehen kann. Ein verständlicher Text hat einen durchschnittlichen Wert von etwa 7,0 bis 8,0. Kincaid orientiert sich am US-amerikanischen Schulsystem und wird wie folgt berechnet:

$$FKGL = (0,39 * ASL) + (11,8 * ASW) \check{} 15,59$$

Die Satzlänge hat bei Kincaid einen größeren Einfluss auf den Index als beim Flesch-Lesbarkeitsindex. Bei beiden Indizes dominiert jedoch die Wortlänge, was auch die begrenzte Anwendbarkeit auf die deutsche Sprache mit ihren Komposita (Wortzusammensetzungen) erklärt.

Die SMOG-Lesbarkeitsformel wurde 1969 von G. Harry McLaughlin [64] veröffentlicht. Das Akronym SMOG steht für Simple Measure of Gobbledygook. Der errechnete SMOG-Indexwert entspricht ungefähr dem Alter, das benötigt wird, um den Text zu verstehen. Eine Punktzahl von 7,4 bedeutet beispielsweise, dass der Text von einem durchschnittlichen Schüler der 7. Klasse verstanden werden kann. Nachdem 30 Sätze aus einem Text ausgewählt wurden, kann die Formel wie folgt berechnet werden:

21 ASW = Average Number of Syllables per Word

22 ASL = Average Sentence Length

$$SMOG = 1,0430\sqrt{\text{Anzahl Wörter} \geq 3 \text{ Silben} \frac{30}{\text{Anzahl Sätze}}} + 3,1291$$

Der SMOG-Index kann nach Bamberger und Vanecek wie folgt für die deutsche Sprache verwendet werden:

$$SMOG = \sqrt{\text{Anzahl Wörter} \geq 3 \text{ Silben}} - 2$$

Der 1975 von Meri Coleman und T. L. Liau entwickelte Coleman-Liau-Index [65] ist eine der am häufigsten verwendeten Lesbarkeitsformeln. Eine Besonderheit des Coleman-Liau-Index ist jedoch, dass die Formel keine Zählung von Silben beinhaltet. Stattdessen orientiert sich der Index an der Anzahl von Buchstaben im Text und wird für die englische Sprache mit der folgenden Formel berechnet:

$$CLI = 0,0588L - 0,296S - 15,8$$

L ist die durchschnittliche Anzahl der Buchstaben pro 100 Wörter. S ist die durchschnittliche Anzahl der Sätze pro 100 Wörter. Das Ergebnis der Formel ist eine Note. Zum Beispiel bedeutet 10,6, dass der Text für einen Schüler der Klasse 10-11 geeignet ist. Auch diese Lesbarkeitsformel ist für die deutsche Sprache nicht geeignet, da die Wörter im Deutschen deutlich länger sind als im Englischen. Werden die ersten 700 Wörter des Kinderbuches »Alice im Wunderland« analysiert, zeigt sich ein Lesbarkeitsindex für die englische Sprache von 5,8 (der Text kann von einem durchschnittlichen Schüler der 5. bis 6. Klasse verstanden werden). Für die deutsche Fassung des Kindertextes liegt der Lesbarkeitsindex bei 9,6. Zwar ist die Anzahl von Sätzen gleich, aber die längeren Wörter im Deutschen suggerieren laut dem Coleman-Liau-Index eine höhere Satzkomplexität. Die durchschnittliche Anzahl von Zeichen pro Wort beim deutschen Text beträgt 4,6. Die Wortlänge im englischen Originaltext ist dagegen durchschnittlich 3,9 Zeichen lang.

Lesbarkeitsformeln bilden eine gute Grundlage, um die Verständlichkeit von Texten zu bestimmen. Sie untersuchen Texte auf objektiv fassbare und zählbare Merkmale, wie etwa Wort- und Satzlänge bzw. wie sich diese auf die Lesbarkeit von Texten auswirken. Je länger beispielsweise die Wörter im Satz, desto komplexer und schwieriger lesbar ist dieser. Die Formeln reichen allerdings nicht aus, um einen Text effizient zu bewerten, da zu wenige Parameter gemessen werden. Außerdem sagt die Formel nichts über den Inhalt einer Nachricht aus. Wie bereits erwähnt, wurden die meisten Lesbarkeitsindices für die englische Sprache konzipiert. Falls diese nicht für die deutsche Sprache geeicht werden, sind sie irreführend, da deutsche Wörter generell länger sind als englische. Der Index würde grundsätzlich eine höhere Textkomplexität suggerieren. Weiterhin ist zu beachten, dass Menschen dazu tendieren, dieselbe Sprache innerhalb einer Gruppe zu sprechen, um Zugehörigkeitsgefühl zu erzeugen bzw. zu verstärken [66]. Cyber-Groomer können ihren Schreibstil in Kinderchats anpassen, um eine höhere Gruppenakzeptanz zu erwirken, sodass Lesbarkeitsformeln hier nicht mehr greifen können. Außerdem müssen Cyber-Groomer nicht per se Erwachsene sein.

Neben den Lesbarkeitsformeln berücksichtigt Privalino auch die Anzahl der Wörter sowie die

Wort- und Satzlänge als Features. Kinder schreiben nicht nur kürzere Wörter, ihr Wortschatz ist auch kleiner als der eines Erwachsenen. So wie die Lesbarkeitsformeln, schließen auch diese Textmerkmale die Grammatik und Semantik, d.h. den eigentlichen Inhalt der Kommunikation, aus. Die Grammatik ist in dem Sinne relevant, da Texte von Kindern beispielsweise mehr Rechtschreibfehler aufweisen. Auch verwenden Kinder weniger Nebensätze, sodass ihre Sätze deutlich kürzer sind[23]. Die Satzlänge ist diesbezüglich ein guter Indikator für das ungefähre Alter des Textverfassers. Allerdings zeigt das folgende Beispiel, dass Messenger für Kurznachrichten prädestiniert sind und die genannten Textmerkmale (Satzlänge, Lesbarkeitsindex etc.) keine zufriedenstellenden Ergebnisse liefern, um Cyber-Groomer zu identifizieren.

Username	**Nachrichtenverlauf**
<...an32>	hast du smartphone?
<...maus12>	ja, hab ich
<...an32>	wollen wir geile fotos tauschen?
<...maus12>	was für?
<...an32>	willst du mein schwanz sehen?
<...maus12>	was soll ich dann zurückschicken?
<...an32>	ein nacktbild von deinem arsch oder <zensiert>

Tabelle 5.3: Beispiel Cybergrooming

Auf der lexikalischen Ebene analysiert Privalino häufig vorkommende Wörter (z.B. »Schule« und »langweilig«). Diese werden mit Wahrscheinlichkeiten versehen und in die Klassen »Cybergrooming« und »ungefährlich« eingeteilt.

Das Bag-of-Words-Modell stellt das wohl bekannteste und gängigste Verfahren zur Konstruktion von Feature-Räumen dar. Dabei werden Wörter unabhängig voneinander und ohne Kontext, in dem sie vorkommen, betrachtet. Es gilt:

$w_1, w_2, \ldots, w_j \in T$ ist die Anzahl eines Wortes w_j in dem Textdokument T. Alle Vorkommen eines Wortes w_j in T werden in einem Feature-Raum zusammengefasst. Wenn nun in der Onlinekommunikation ein Wort aus der gelabelten Liste vorkommt, ist dies ein Indiz dafür, zu welcher Klasse der Text gehört. Wörter wie »Nacktbild« oder »ausziehen« werden in die Klasse »Cybergrooming« eingeordnet und »Schule« und »langweilig« in die Kategorie »ungefährlich«. Da eine solche Wortliste nie vollständig sein kann, sind solche Features fehleranfällig, da Wörter, die nicht in der Liste vorkommen, nicht erkannt werden. Privalino gibt zusätzlich an, dass Trainingsdaten täglich aktualisiert werden, um sich den neuen Maschen von Cyber-Groomern anzupassen.

Auf der semantischen Ebene wird der Bag-of-Words-Ansatz ebenfalls verfolgt, um Rückschlüsse über die Stimmung beider Chatpartner zu ziehen. Hier verwendet Privalino Listen mit positiven und negativen Wörtern. Dieser lexikalische Ansatz kann helfen, die Stimmung im Text zu erkennen. Das Wort »gut« ist beispielsweise positiv konnotiert, das Wort »schlecht« negativ. Dazwischen existieren allerdings Stimmungswörter, die mehrdeutig sind und je nach Kontext eine andere Bedeutung haben können (z.B. »Schwanz«). Ein weiterer Nachteil des Bag-of-Words-Ansatzes ist, dass alle Flexionen eines Wortes berücksichtigt werden müssen. Es sei denn, die Wörter werden auf ihren Stamm zurückgeführt. Da Kinder und Jugendliche besonders in Chats nicht auf die

23 https://www.privalino.de/merkmale, Juli 2018

Rechtschreibung achten, werden falsch geschriebene Wörter vom System nicht erkannt. Zudem können Wörter auch absichtlich falsch geschrieben sein wie z.B. »ich will s** mit dir haben«.

Auf der semantischen Ebene verwendet Privalino für ihren Algorithmus auch Pronomen (wie »ich«, »du«, »er/sie/es«). Das Unternehmen hat herausgefunden, dass die Verwendung von Pronomen Rückschlüsse auf die Identität des Autors zulässt. Die »ich-bezogenen« Personalpronomen »ich« oder »mein« werden verstärkt von Kindern verwendet. Das folgende Beispiel zeigt die Schwierigkeiten auf, wenn auf (Schlüssel-)Wortbasis Nachrichten analysiert werden und der Kontext nicht mitberücksichtigt wird. Der Satz »Ich zeig dir meins – Du zeigst mir deins« ist weder negativ konnotiert, noch weist er Schlüsselwörter auf, die auf einen Cyber-Groomer hindeuten. Die Wort- und Satzlänge sind beide kurzgehalten und der Satz beinhaltet zudem Personalpronomen, die »ich-bezogen« sind, was auf das Kommunikationsverhalten eines Kindes hindeutet. Nur im Kontext betrachtet, ergibt dieser Satz einen Sinn, dass es sich um Sexting handelt bzw. Cybergrooming. Privalino gibt auf seiner Homepage bekannt, dass sie kontinuierlich neue Algorithmen und Merkmale testen. Ansätze auf Basis von neuronalen Netzen (Deep Learning genannt) sind geplant. Mithilfe dieser Verfahren erhofft sich Privalino, irrtümliche Warnungen zu minimieren.

5.2.2 Autorschaftsverifikation

Die Autorschaftsverifikation (kurz AV) stellt neben AP eine weitere Unterdisziplin der Autorschaftsanalyse dar und beschäftigt sich mit der zentralen Fragestellung, ob ein Text eines anonymen Autors U tatsächlich von einem bekannten Autor A stammt, wobei von diesem Vergleichstexte vorliegen. Formal ausgedrückt soll also geklärt werden, ob U = A (übereinstimmende Autorschaft) oder U /= A (andere Autorschaft) gilt.

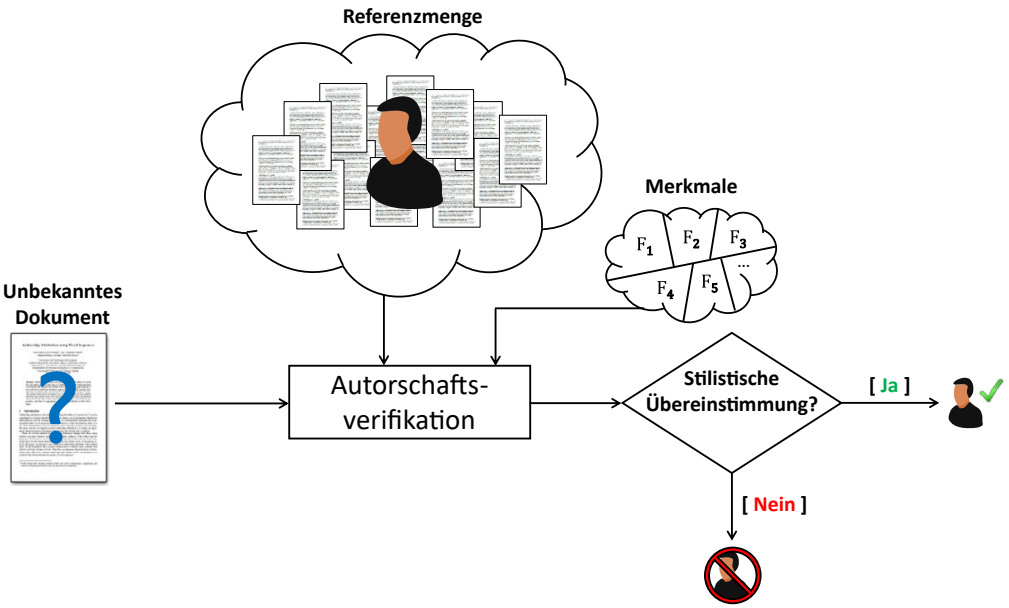


Abbildung 5.9: Schematische Darstellung eines Autorschaftsverifikationssystems.

Im Gegensatz zu AP existiert in AV nur eine Klasse (siehe Abbildung 5.9) aus der eine Entscheidungsgrenze bestimmt werden soll. Aus Sicht des MLs fällt AV daher in die Kategorie der unären Klassifikationsprobleme.

Bezug zu Cybergrooming

Manche Cyber-Groomer gelten als Wiederholungstäter. Es ist somit damit zu rechnen, dass sie selbst nach einer Verurteilung weiterhin die Nähe zu Kindern suchen. Hierfür bieten sich für Cyber-Groomer, wie eingangs erwähnt, Pseudonyme an, die eine anonyme Kommunikation im Netz erlauben. Um hinter Pseudonymen getarnte Cyber-Groomer wiederzuerkennen, die sich in Chat-Portalen und sozialen Netzwerken bewegen, kann AV als ein ermittlungstechnisches Werkzeug behilflich sein. Allerdings ist zu beachten, dass das nur funktionieren kann, sofern Referenztexte (z.B. Chat-Verläufe) von den bekannten Tätern zur Verfügung stehen. Abseits von Cybergrooming existieren für die AV zahlreiche Anwendungsmöglichkeiten, von denen seitens der Forschung bereits einige umgesetzt wurden. Im Bereich Cybersecurity setzten beispielsweise Neal et al. AV für die kontinuierliche Authentifizierung von Benutzern auf Basis der verwendeten Sprache ein. Anstelle von Passwörtern oder biometrischen Charakteristika wie etwa Fingerabdrücke, Handvenenstruktur oder Stimme wird so der Schreibstil einer Person auf dessen Einzigartigkeit überprüft. Barbon et al. verwendeten AV dagegen, um kompromittierte Social-Media-Accounts zu identifizieren. Im Bereich des Information-Retrievals wurde AV zudem genutzt, bestehende Systeme zu erweitern (beispielsweise von den Forschern Rexha et al. [69]). Dadurch wird die Suche nicht nur auf der thematischen, sondern auch auf der stilistischen Ebene der Texte ermöglicht. Auch im Rahmen der Fake-News- Erkennung wurde bereits auf AV zurückgegriffen [70] sowie für die Erkennung von Alzheimer im frühen Stadium [71]. Beim Letzteren diente AV als ein Instrument, um die abnehmende Qualität der natürlichen Sprache betroffener Alzheimer-Patienten zu messen.

Arbeitsweise von Autorschaftsverifikationsverfahren

Die Arbeitsweise von AV-Verfahren ähnelt sehr der von AP-Verfahren. Auch hier wird eine Referenzmenge mit Beispieltexten benötigt, die jedoch anders konstruiert wird und eine andere Semantik aufweist. Während in AP Beispiele vorliegen, die Autoreigenschaften wie Alter, Geschlecht oder Nationalität repräsentieren, enthält in AV die Referenzmenge D_{A} lediglich Beispieltexte einer bekannten Person A. Diese Texte dienen dazu, den Schreibstil von A bestmöglich zu modellieren. Um dies zu erreichen, muss (ähnlich zu AP) auch hier eine geeignete Datenrepräsentation gewählt werden, die sich (anders als in AP) ausschließlich aus stilistischen Merkmalen zusammensetzt. Hintergrund dabei ist, dass inhaltsbasierte Wörter (die in AP oftmals eine tragende Rolle spielen) den Schreibstil einer Person nicht zuverlässig, und vor allem nicht generalisierend, widerspiegeln können. Stilistische Merkmale (wie z.B. Funktionswörter oder Wortarten, siehe Tabelle 5.2) sind dagegen **Textsorten-** und weitgehend **Thema**-unabhängig und eignen sich daher, Schreibstile von Autoren zu beschreiben.

Mithilfe solcher Merkmale wird ein stilistisches Modell konstruiert, welches es anschließend in einem unbekannten Dokument D_{U} wiederzufinden gilt. Dabei ist es offensichtlich, dass nicht alle Merkmale tatsächlich vorgefunden werden. Um eine erfolgreiche Verifikation des Schreibstils zu ermöglichen, wird daher ein Akzeptanzkriterium benötigt, welches in der Lage ist, fehlende Stilmerkmale in D_{U} zu tolerieren.

Erfolgsaussichten seitens der Forschung: Seit mehr als zwei Jahrzehnten wird auf dem Gebiet der AV intensiv geforscht, wobei zahlreiche Verfahren bereits vorgeschlagen wurden. Eines der bekanntesten Verfahren ist die **Impostors-Methode** (kurz IM), welche von Koppel und Winter im Jahre 2013 [72] vorgestellt wurde. Um das unäre Klassifikationsproblem, das die AV darstellt, in ein binäres Klassifikationsproblem umzuwandeln, behilft sich IM eines Tricks. Hierfür werden mithilfe einer Suchmaschine Dokumente ausfindig gemacht, die thematisch zu dem anonymen Dokument D_U, als auch zu den Referenztexten des bekannten Autors A passen. Hintergrund dieser sogenannten Impostors-Dokumente ist eine Gegenklasse $\neg A$ zu konstruieren, sodass das zugrunde liegende Klassifikationsverfahren bei seiner Entscheidung $\{A, \neg A\}$ (gleicher/anderer Autor) hinsichtlich D_U nicht ausschließlich auf Referenztexten von A basiert. AV-Methoden, die auf externe Dokumente angewiesen sind, werden im Fachjargon als »extrinsisch« bezeichnet.

IM stützt sich auf ein distanzbasiertes Klassifikationsverfahren (ähnlich wie die in Unterabschnitt 5.2.1 aufgeführte Kosinus-Ähnlichkeit), das die stilistische Ähnlichkeit von D_U zu den Referenztexten von A als auch zu den Impostors-Dokumenten misst. Die Klasse desjenigen Dokuments, dessen Schreibstil am ähnlichsten zu dem von D_U ist, wird als Ergebnis vorhergesagt. IM lieferte den Grundstein für viele weitere Ansätze (z.B. [73, 74, 75, 76, 77]) und zählt zu den erfolgreichsten Verfahren auf dem Gebiet der AV. Im Rahmen des internationalen jährlichen AV-Wettbewerbs **PAN**[24] gewann IM (bzw. IM-basierte Verfahren) zweimal in Folge den ersten Platz [78, 79].

Ein weiterer erfolgreicher Ansatz, der zu den Meilensteinen in der AV gehört, ist die **Unmasking-Methode**, welche von Koppel und Schler [80] im Jahre 2004 vorgestellt wurde. Anstatt die Gemeinsamkeiten von Texten zu untersuchen, betrachteten die Autoren, wie gut sich diese diskriminieren lassen. Grundgedanke hierbei ist, dass sich Werke vom selben Autor oftmals nur anhand weniger, essenzieller Merkmale unterscheiden, während der grundlegende Schreibstil gleich bleibt. Eine mögliche Ursache hierfür ist beispielsweise das Genre oder die Thematik des Textes sowie altersbedingte Änderungen im Schreibstil. Wenn diese Features entfernt werden, ist die Differenzierung der Texte wesentlich schwieriger. Bei Werken von unterschiedlichen Autoren, die sich zusätzlich im grundlegenden Schreibstil unterscheiden, lassen sich diese selbst nach Entfernung von essenziellen Features gut unterscheiden.

Nach diesem Prinzip schlagen Koppel und Schler eine Methode vor, bei welcher iterativ die stärksten und schwächsten Features entfernt und die Texte nach jeder Iteration anhand der übrig geblieben Features verglichen werden. Mit den stärksten und schwächsten Features sind jene gemeint, die zur Unterscheidung der Texte am meisten bzw. wenigsten beitragen. Das Resultat sind sogenannte Degradationskurven (siehe Abbildung 5.10).

24 https://pan.webis.de/clef15/pan15-web/author-identification.html, Juli 2018

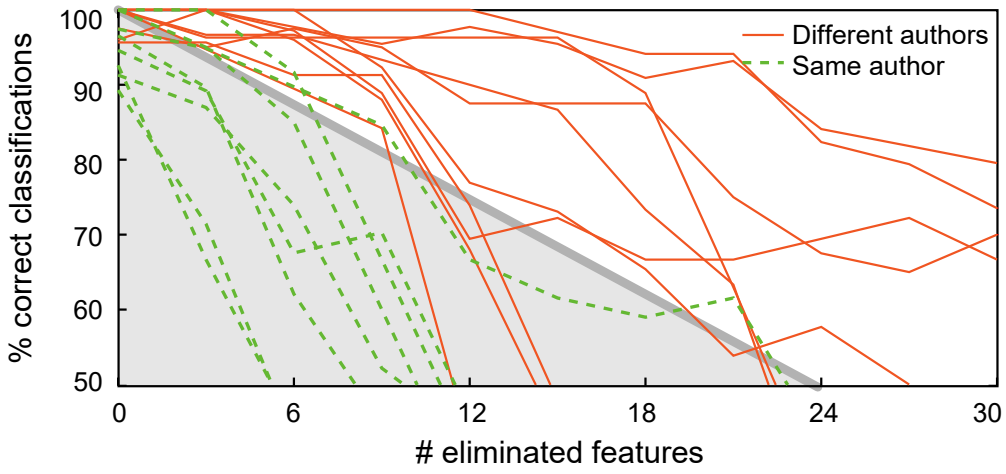


Abbildung 5.10: Unmasking in Aktion: Jede Kurve repräsentiert einen Vergleich zwischen je zwei Texten (das unbekannte Dokument und der Referenztext). Textpaare mit übereinstimmender/fremder Autorschaft sind grün bzw. orange gefärbt. Abbildung entnommen aus [81].

Jeder Text-zu-Text-Vergleich bildet jeweils eine Degradationskurve. Kurven, die stark abfallen, bedeuten, dass die Texte nach der iterativen Entfernung von Features wesentlich schlechter unterschieden werden können (und dementsprechend vom selben Autor stammen) als Kurven, die kaum bis gar nicht abfallen. In ihren Experimenten erzielten Koppel und Schler [80] eine Erkennungsgenauigkeit von 95,7% bzgl. 209 Verifikationsfällen, wobei Unmasking aus 189 Nein- sowie 20 Ja-Fällen 181 bzw. 19 Autorschaften korrekt klassifizieren konnte.

6 Eignungsprüfung

In diesem Kapitel werden die Anforderungen an technische Lösungen zur Erkennung von Sexting und Cybergrooming formuliert sowie darüber hinaus einige der in Kapitel 10 beschriebenen Umsetzungsmöglichkeiten evaluiert.

6.1 Evaluation bestehender Technologien zur Sexting-Erkennung

Zur Evaluierung der in Kapitel 10 vorgestellten Verfahren im Rahmen einer Sexting-Erkennungssoftware wurde zunächst eine Testmenge aus insgesamt 2.000 Bildern erstellt, die mithilfe eines Web-Crawlers von Seiten der Internet-Plattform Reddit heruntergeladen wurden. Insgesamt enthielten 1.000 Bilder Nacktaufnahmen von Personen in eindeutig aufreizenden Posen oder pornografische Inhalte. Diese wurden als Positivbeispiele für die Klassifikation markiert. Die übrigen 1.000 Bilder enthielten gewöhnliche Aufnahmen von Personen, wobei deren Kleidung von Alltagskleidung bis hin zu Bademode reichte. Diese wurden als Negativbeispiele markiert. Die meisten Bilder entsprachen in ihrem Aufnahmewinkel typischen Selbstporträts mit der Handykamera (»Selfies«).

Anhand der so generierten Testmenge wurden verschiedene Verfahren zur Nacktheits- bzw. Pornografieerkennung getestet. Für schwellwertbasierte Verfahren konnte hierbei zusätzlich ein geeigneter Schwellwert zur Unterscheidung zwischen positiven und negativen Beispielbildern ermittelt werden. Um die Verfahren hinsichtlich ihrer Eignung für eine Sexting-Erkennung untereinander vergleichbar zu machen, wurde einheitlich mit den Metriken TPR (True-Positive-Rate bzw. Richtig-Positiv-Rate) und FPR (False-Positive-Rate bzw. Falsch-Positiv-Rate) gemessen.

Von den in Kapitel 10 vorgestellten Software-Lösungen wurden folgende mit der im Rahmen dieser Studie generierten Testmenge evaluiert:

- Nudepy
- I2V
- Clarifai
- Yahoo Open NSFW

Nude Detect und Sightengine wurden nicht getestet, da die kostenfreien Zugänge zu den entsprechenden Web-Services zu geringe Kontingente freier Bildklassifikationen erlauben. Um diese Anbieter also ausgiebig zu testen, wäre ein kostenpflichtiger Zugang erforderlich. Jedoch zeigten Ergebnisse auf einer reduzierten Testmenge, dass für beide Klassifikatoren keine höhere Genauigkeit als bei Clarifai und Yahoo Open NSFW zu erwarten ist, weshalb auf die Nutzung der kostenpflichtigen Variante verzichtet wurde.

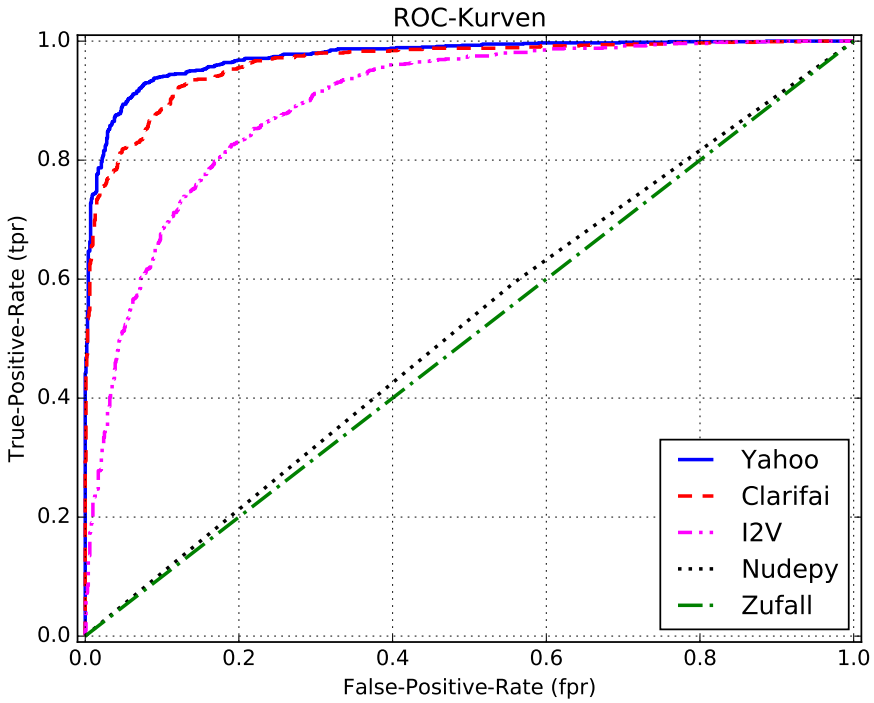


Abbildung 6.1: Evaluationsergebnisse verschiedener Verfahren zur Nacktheits-/Pornografieerkennung auf insgesamt 2.000 Bildern. Die einzelnen ROC-Kurven zeigen die True-Positive-Rate (TPR) in Abhängigkeit der entsprechenden False-Positive-Rate (FPR) des jeweiligen Verfahrens.

Abbildung 6.1 zeigt die Evaluationsergebnisse in Form von ROC[1]-Kurven. Eine ROC-Kurve gibt das Verhältnis von TPR zu FPR für verschiedene Schwellwerte des Klassifikationsverfahrens an. Je größer die unter der daraus resultierenden Kurve liegende Fläche ist (diese kann beliebige Werte zwischen 0 und 1 annehmen), desto besser wird das Verfahren bewertet. Ein (in Bezug auf die Testdaten) fehlerfreies Verfahren würde somit einer ROC-Kurve entsprechen, die senkrecht auf der Ordinatenachse (vertikale Koordinatenachse) bis zum Wert 1 ansteigt und anschließend horizontal verläuft. Wie in Abbildung 6.1 ersichtlich, liegen die Verfahren von Yahoo und Clarifai nah an diesem Idealfall. Sie sind dementsprechend gut in der Lage, Nacktbilder und Pornografie zu erkennen, was sie zu geeigneten Verfahren für eine Sexting-Erkennung macht. I2V liefert ebenfalls relativ gute Ergebnisse, schneidet gegenüber Yahoo und Clarifai jedoch deutlich schlechter ab. Ein Verfahren, dessen TPR und FPR unabhängig vom Schwellwert stets ausgeglichen sind, würde der Hauptdiagonalen des Koordinatensystems folgen. Dies entspräche der Performanz einer zufallsbasierten Klassifikation. Da die ROC-Kurve von Nudepy sehr nah an der einer zufallsbasierten Klassifikation liegt, erweist sich dieses Verfahren als ungeeignet für die Sexting-Erkennung, da es bei der Klassifikation von Nacktbildern zu viele Fehlentscheidungen trifft.

Mit dem Youden-Index lässt sich zu einer gegebenen ROC-Kurve der optimale Schwellwert für ein entsprechendes binäres Klassifikationsverfahren bestimmen. Dieser wird für jeden Schwellwert über die Formel TPR - FPR berechnet. Der Schwellwert mit dem höchsten Youden-Index sollte gewählt werden, um ein bestmögliches Klassifikationsergebnis zu erzielen.

1 Abk. für Receiver Operating Characteristic

Die sogenannte J-Statistik nach Youden wird wie folgt berechnet:

$$J = sensitivity + specificity - 1$$
$$= \text{TPR - FPR},$$

Das Maximum des J-Wertes bestimmt die optimale Schwelle für jeden Klassifikator. Zusätzlich zu den AUC-Werten zeigt die Tabelle 6.1 die optimalen Schwellenwerte für Clarifai, Yahoo und I2V zusammen mit der entsprechenden True Positive Rate (TPR) und False Positive Rate (FPR).

Während die optimalen Schwellwerte für Yahoo und Clarifai ungefähr in der Mitte zwischen 0 und 1 liegen, liegt der optimale Schwellwert von I2V nahe bei 0. Da nude.py einen festen Schwellwert verwendet und nur True/False-Labels zurückgibt, erlaubt es keine Schwellwertoptimierung.

Klassifikationsverfahren	AUC	Schwellwert	TPR	FPR
Yahoo	0,975	0,384	0,928	0,076
Clarifai	0,963	0,682	0,922	0,121
I2V	0,896	0,090	0,826	0,190
nude.py	0,518	-	0,594	0,558
coin toss	0,500	-	0,500	0,500

Tabelle 6.1: AUC, optimale Schwellwerte und zugehörige TPR und FPR der evaluierten NSFW-Klassifikationsverfahren.

6.1.1 Kombination von farbwertbasierten Verfahren und Deep-Learning-Modellen

Um zu evaluieren, ob durch Vorfilter die Leistung existierender Netze noch gesteigert werden kann, haben wir das 'Open NSFW CNN' von Yahoo um verschiedene Filter erweitert, die Bilder vor der Prüfung durch das Netz vorverarbeiten.

Die Idee hierbei ist zu prüfen, ob mit herkömmlichen Methoden der Bildverarbeitung bzw. Objekterkennung eine Filterung von störenden Einflüssen auf die darauf folgende Prüfung durch ein neuronales Netz erfolgen kann. Nehmen wir hierzu einen vereinfachten Fall an: Personen können mit einer Objekterkennung zuverlässig erkannt werden. Diese Erkennung ist eventuell sogar zuverlässiger als die Erkennung eines menschlichen Umrisses durch ein neuronales Netz. Die Entscheidung, ob die Person nackt ist, ist aber mit dem neuronalen Netz zuverlässiger. Nun wird erst mit einer Objekterkennung die Person identifiziert, der Rest des Bildes entfernt und nun nur die Person dem Netz übergeben, das nun nur noch entscheiden muss, ob die Person nackt ist oder nicht. Im Idealfall werden so die Stärken beider Verfahren kombiniert, da die jeweiligen Fehlerquellen vermieden werden.

Die Abbildungen 6.2 und 6.4 zeigen zwei Beispiele für die Vorgehensweise bei der Filterung von Bildern durch eine Bounding Box. Die dazugehörigen Tabellen zeigen, dass eine Steigerung der Erkennung der Bilder um ein Vielfaches möglich ist, wenn die richtige Strategie gewählt wird. So wird im ersten Beispiel die Erkennungsrate von 0,12 auf 0,92 erhöht, wenn die gezeigte Person umrandet wird, der Hintergrund innerhalb der Bounding Box aber unverändert bleibt.

INFO

Eine Bounding Box (deutsch auch »Hüllkörper«) ist bei einem zweidimensionalen Objekt der kleinste Kasten, der dieses Objekt umschließt. Üblicherweise wird hier nicht auf eine geometrische Ausrichtung geachtet, der Kasten verläuft also parallel zu den Bildkanten).

Basierend auf der Tatsache, dass es sich bei Neuronalen Netzen um eine Art Blackbox-Algorithmus handelt, musste zuerst eine Analyse zu dem Vorgehen des Open NSFW CNN von Yahoo durchgeführt werden. Die Ergebnisse der ersten Versuchsreihe zeigten bereits eine Detektionsrate von 76,2% für richtig positive Bilder sowie eine Detektionsrate von 83,7% für richtig negative Bilder. Durch weitere Versuche mit Bildern von Sonnenuntergängen (ein typisches Beispiel für Bilder, in denen viele Pixel mit vermeintlichen Hautfarben vorkommen) sowie Bildern in Graustufen wurde der Einfluss von Farben und Konturen untersucht und festgestellt, dass die Konturen einen stärkeren Einfluss auf die Klassifizierung haben als die Farbe. Zusätzlich wurde eine Auswahl von künstlichen Bildern, welche von dem Neuronalen Netzwerk als NSFW erkannt werden sollten, analysiert. Resultierend aus den Erkenntnissen wurde geprüft, ob sich das Entfernen von Hintergrundpixeln positiv auf die Detektion des Netzes auswirkt.

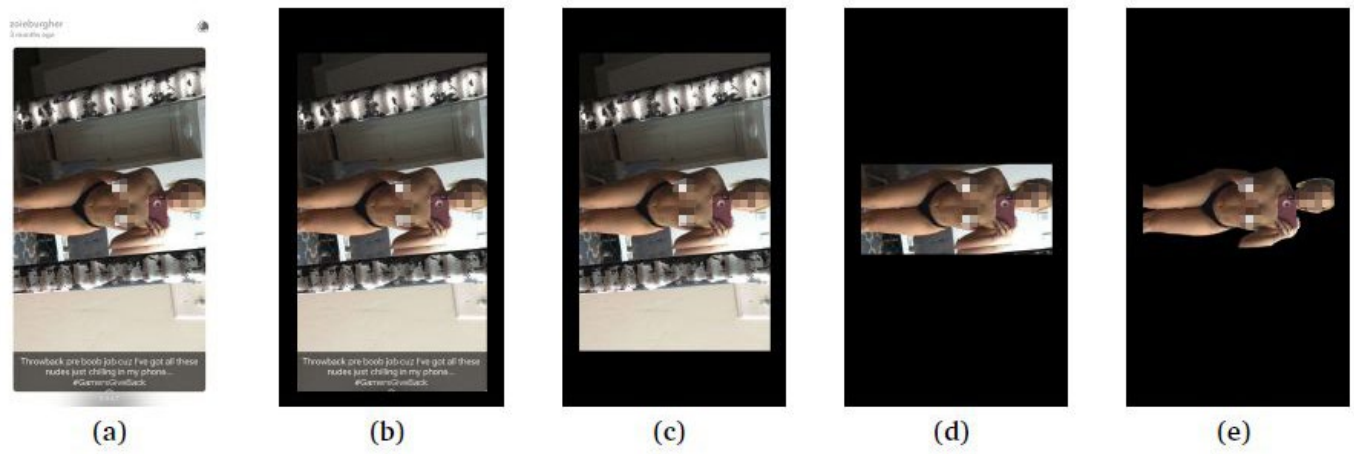


Abbildung 6.2:
Beispiel für unveränderten (a) bis stark eingegrenzten (d) und entfernten Hintergrund (e)

Bild	a	b	c	d	e
NSFW-Wert	0,1234	0,3366	0,3529	0,9283	0,2625

Tabelle 6.2:
NSFW-Bewertung Beispielbild Abbildung 6.2

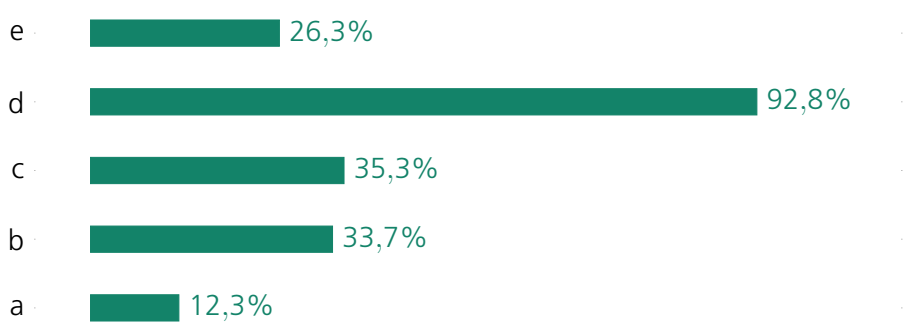


Abbildung 6.3:
NSFW-Erkennungsraten abhängig von Hilfsmethoden

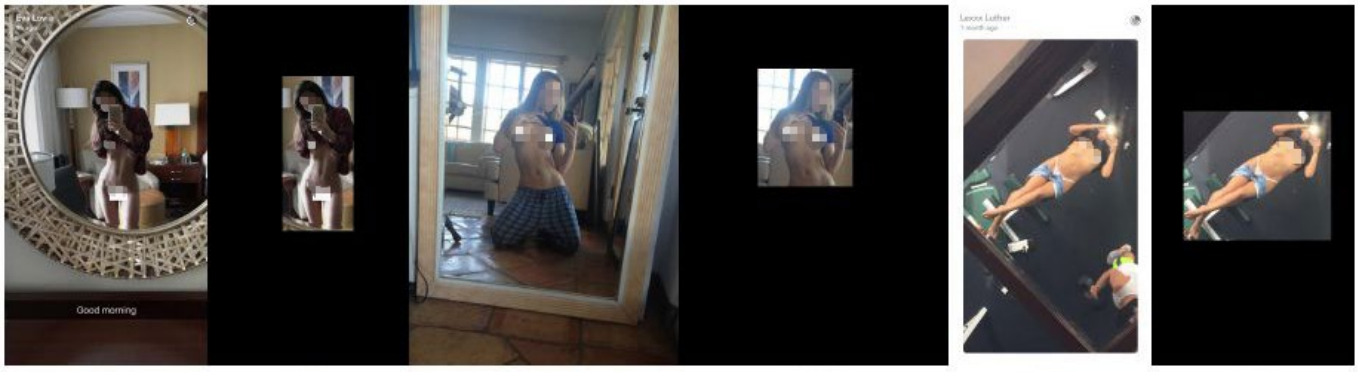


Bild	a	b	c	d	e
NSFW-Wert	0,1234	0,3366	0,3529	0,9283	0,2625

Abbildung 6.4: 3 Beispiele für den Einsatz einer Bounding Box Tabelle 6.3: NSFW Bewertung Beispiele Abbildung 6.4

Tabelle 6.3: NSFW Bewertung Beispiele Abbildung 6.4

Entsprechend den Ergebnissen der Analyse konnte ein zweistufiger Vorfilter konzipiert werden. Dieser besteht aus der Detektion von Hautpixeln sowie aus Methoden zur Abdeckung von Nicht-Hautpixeln. Bei den Detektoren über die Farbräume RGB, normalisierter RGB und HSV erwies sich eine Kombination, bestehend aus diesen drei Detektoren, als effektivste Lösung. Zur Abdeckung der Nicht-Hautpixel erfolgte ein Vergleich über die direkte Abdeckung mittels schwarzer Pixel, die Abdeckung über Convex Hull sowie die Abdeckung über Bounding Box. Dabei erzielte die Abdeckung über Bounding Box das beste Ergebnis. Alle Abdeckungsalgorithmen überschrieben die nicht relevanten Pixel mit schwarzer Farbe.

INFO

Die Convex Hull (deutsch: konvexe Hülle) ist eine geometrische Form, die ein Objekt so umschließt, dass jede Linie zweier Punkte in dem Objekt auch innerhalb der konvexen Hülle liegt. Sie ist üblicherweise aufwendiger zu berechnen und hat eine kleinere Fläche als die Bounding Box desselben Objektes.

Daraus resultierten Kanten, welche die Auswertung durch das Neuronale Netzwerk negativ beeinflussen können. Aus diesem Grund wurde der Bounding-Box-Algorithmus mit einem Gaußfilter kombiniert. Ein Gaußfilter wird in der Bildverarbeitung zum Weichzeichnen verwendet und glättet harte Übergänge bei Kanten. Diese Kombination schnitt im Vergleich zu dem Algorithmus ohne Gaußfilter besser ab und erwies sich als die geeignetste Variante eines Vorfilters. Um die Leistung des Vorfilters zu evaluieren, wurden dessen Detektionsraten mit denen ohne Filter verglichen. Einige Ergebnisse sind in Abbildung 6.4 zu sehen. Die Detektionsraten wurden in diesen drei Beispielen jeweils grob verdoppelt

6.2 Evaluation eigener Technologien zur Cybergrooming-Erkennung

Abgesehen von wissenschaftlichen Experimental-Werkzeugen, existieren (unserer Recherche nach) keine fertigen, öffentlich zugänglichen und kostenlosen Software-Lösungen zur automatisierten Erkennung von Cybergrooming. Dies trifft sowohl auf Autorschaftsverifikation (AV) als auch auf Autorenprofiling (AP) zu. Aus diesem Grunde haben wir bzgl. beider Disziplinen eigene Lösungen entwickelt, um die Praxistauglichkeit hinsichtlich realistischer Daten bewerten zu können. Im Folgenden gehen wir zunächst auf ein einfaches AP-Verfahren ein, welches im Rahmen dieser Studie entwickelt wurde. Anschließend erläutern wir ein von uns bereits entwickeltes AV-Verfahren.

6.2.1 Autorenprofiling: Erkennung von Alter – Datengrundlage

Aufgrund der Tatsache, dass es uns nicht möglich war, an deutschsprachige Cybergrooming-Texte zu gelangen, entschieden wir uns, englischsprachige Texte aus einer öffentlich zugänglichen amerikanischen Webseite namens »Perverted Justice«[2] zu betrachten, die realistische Cybergrooming-Szenarien widerspiegelt. Konkret handelt es sich bei diesen Texten um Konversationen zwischen Cyber-Groomern verschiedener Altersklassen und ihren (primär) minderjährigen Opfern. Abbildung 6.5 zeigt eine beispielhafte Konversation.

vamale_692005 (8:00:44 PM): did you have an orgasm from it?
sweet_erin78 (8:00:51 PM): yea
vamale_692005 (8:00:55 PM): cool
vamale_692005 (8:01:10 PM): so besides taking a cock in your pussy...woudl you try taking it anywhere else?
sweet_erin78 (8:01:17 PM): like wher
vamale_692005 (8:01:25 PM): **would you ever try anal?**
sweet_erin78 (8:01:34 PM): ouch that sound like it hurt
vamale_692005 (8:01:51 PM): not if it is done right..you have to be slow and gentle with that
vamale_692005 (8:01:53 PM): would you try it?
sweet_erin78 (8:02:07 PM): umm i dunno. mebbie
vamale_692005 (8:02:14 PM): okay..just wondering
vamale_692005 (8:02:28 PM): so have you ever had anyone put whip cream or honey on you and lick it off?
sweet_erin78 (8:02:34 PM): no
vamale_692005 (8:02:47 PM): you would love it..that is if you would want to try that
sweet_erin78 (8:02:54 PM): yea that sound way cool!
vamale_692005 (8:02:57 PM): okay cool
vamale_692005 (8:03:06 PM): so have you ever tasted your own pussy?

Abbildung 6.5: Perverted Justice: Konversationen zwischen einem Cyber-Groomer und dessen Opfer.

Die erste Hürde bestand darin, die entsprechenden Dialoge aus den zahlreichen unübersichtlichen Webseiten zu crawlen, zu parsen und entsprechend aufzubereiten. Letzteres war dabei insbesondere wichtig, da das Verfahren nur Textabschnitte des Cyber-Groomers und nicht die des Opfers betrachten soll. Eine weitere Schwierigkeit war, dass die Texte überwiegend mit Slang-Ausdrücken behaftet sind, sodass anspruchsvolle linguistische Verfahren (z.B. POS-Tagger, Chunker oder NER-Erkenner) von vornherein aufgrund der Fehleranfälligkeit ausgeschlossen wurden, um an spezifische Merkmale zu gelangen. Zwar existieren Möglichkeiten, Texte hinsichtlich der Sprache zu normalisieren, jedoch geht dies mit dem Verlust wichtiger stilistischer Merkmale einher, die zu der Vorhersage der Autoreigenschaft »Alter« positiv beitragen können. Aus diesem Grund entschieden wir uns gegen eine gesonderte Normalisierung.

2 http://www.perverted-justice.com

Nach Aufbereitung der Daten haben wir für jeden Cyber-Groomer Xi sämtliche Textabschnitte zusammengefügt, sodass für Xi ein einziger langer Text vorlag. Anschließend haben wir die gesamte Datenmenge (bestehend aus insgesamt 585 Cyber-Groomer-Texten X1, X2, . . . , X585) in eine Trainings- und Testmenge aufgeteilt. Die Trainingsmenge bestand aus 284 und die Testmenge aus 301 Cyber-Groomern und ihren entsprechenden Konversationen. Hinsichtlich der Textlängen variieren die einzelnen Texte stark (von 1 bis 300 KByte). Zwar lassen sich die Texte bzgl. ihrer Längen problemlos vereinheitlichen, jedoch wollten wir ein möglichst realistisches Szenario wiedergeben, in dem ein Cyber-Groomer mehr schreibt als der andere.

Eigenes AP-Verfahren

Als ein beispielhaftes AP-Verfahren entschieden wir uns für ein einfaches neuronales Netz (NN) mit einer Feedforward-Architektur (siehe Abbildung 6.6).

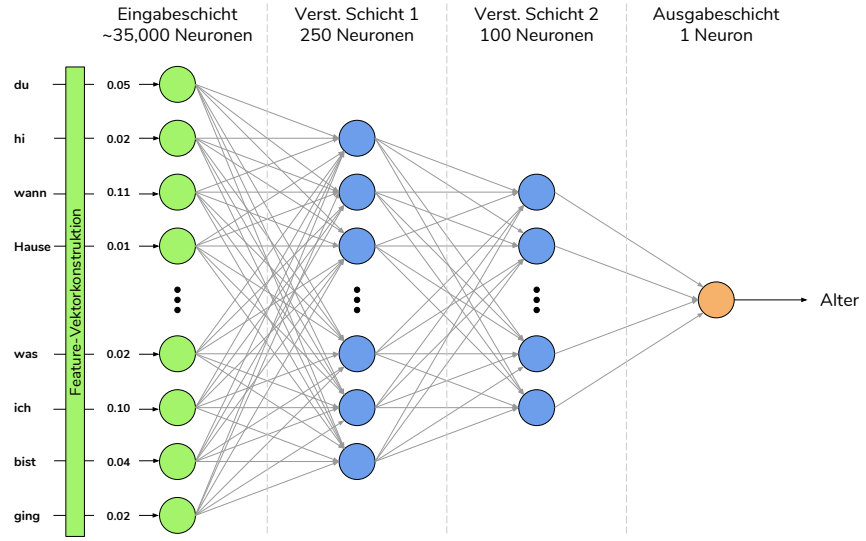


Abbildung 6.6:
Eigene AP-Lösung: Ein einfaches NN mit zwei versteckten Schichten und einem Neuron in der Ausgabeschicht, welches das Alter des Autors als kontinuierlichen Wert vorhersagt.

Das Verfahren nimmt als Eingabe zunächst einen aufbereiteten Trainingskorpus und extrahiert daraus die 2000 am meisten verwendeten Wörter, die das Grundvokabular für das NN bilden. Anschließend konstruierten wir für jeden Cybergrooming-Text Xi einen korrespondierenden Feature-Vektor. Jedes einzelne Feature in einem Vektor ist eine relative Häufigkeit eines Wortes, das sowohl im Grundvokabular als auch im Text Xi vorkommt. Im Rahmen des Trainings werden die generierten Feature-Vektoren nacheinander in das NN eingegeben und für jede Iteration der Fehler gemessen.

Da wir statt diskreter Altersklassen das genaue Alter (eine kontinuierliche Zahl) vorhersagen, betrachten wir hinsichtlich des gemessenen Fehlers nicht die Accuracy sondern das RMSE-Maß (siehe Kapitel C.3.2). Der Fehler drückt dabei aus, wie weit wir bzgl. unserer Vorhersage von dem tatsächlichen Alter abweichen.

Nach Abschluss des Trainings haben wir das gelernte Modell auf der Testmenge evaluiert. Die Spanne der gemessenen Fehler reicht von 0,02 bis hin zu 28,48 Jahren, mit einem Median von

4,61 Jahren. Rundet man diese Zahl auf, dann lässt sich der Wert so interpretieren, dass unser Verfahren sich bzgl. des tatsächlichen Alters einer Person im Schnitt um (plus/minus) 5 Jahre verschätzt. In Anbetracht der Länge der Dokumente und der verwendeten (Slang-)Sprache kann das Ergebnis mit anderen Verfahren aus der aktuellen Forschung mithalten.

6.2.2 Autorschaftsverifikation: Erkennung von selben Autoren – Datengrundlage

Als Datengrundlage für die AV verwendeten wir dieselbe Datenmenge, wie bei dem oben genannten AP-Verfahren, jedoch mit einer anderen Struktur. Da hier Autoren gegeneinander verifiziert werden müssen (statt Autoreigenschaften vorherzusagen) galt es ein Korpus mit entsprechenden Verifikationsfällen zu konstruieren. Alle Cyber-Groomer-Texte wurden hierfür in zwei Hälften aufgeteilt, sodass die eine Hälfte für Fälle mit übereinstimmenden Autorschaften vorgesehen war, während die andere Hälfte für nicht-übereinstimmende Autorschaftsfälle diente. Da allerdings einige Texte nicht über eine ausreichende Textlänge von mindestens 1000 Zeichen verfügten, wurden diese ausgeschlossen. Insgesamt entstand ein ausbalancierter Korpus mit 1158 Verifikationsfällen (mit jeweils 579 Ja- und 579 Nein-Fällen), der schließlich in eine Trainings- und Testmenge aufgeteilt wurde. Die Trainingsmenge umfasste dabei 346 und die Testmenge 812 Verifikationsfälle.

Eigenes AV-Verfahren

Als AV-Verfahren diente unser bereits existierender Ansatz COAV [82] (siehe Abbildung 6.7), welchen wir auf der ARES-Konferenz im Jahre 2017 vorgestellt haben.

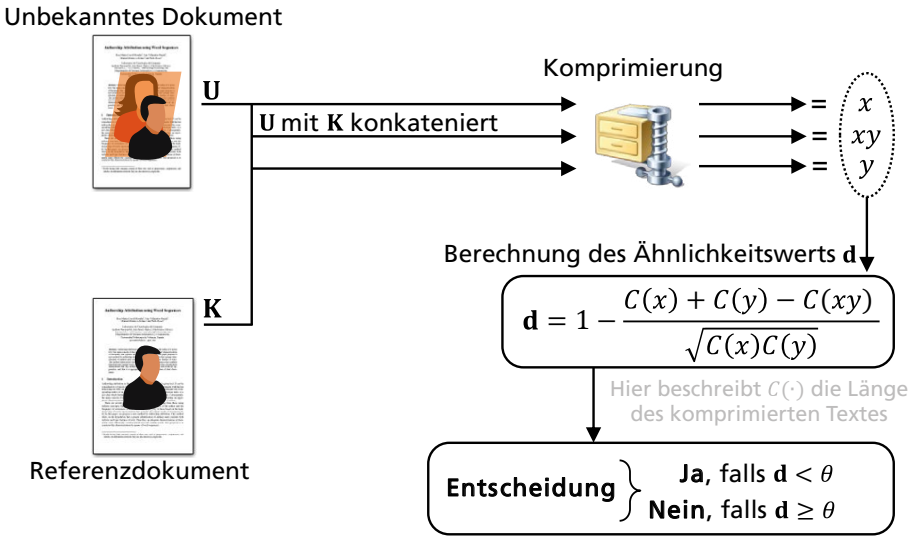


Abbildung 6.7: COAV: Ein auf Kompression basiertes AV-Verfahren.

Das Verfahren erzielt Ergebnisse, die ähnlich oder besser sind als aktuelle State-of-the-Art AV-Ansätze und wurde von uns umfangreich getestet. Die Idee hinter COAV ist es, den manuellen Schritt der Merkmalsextraktion an einen Kompressionsalgorithmus zu delegieren und die Ähnlichkeitsberechnung zwischen den zu untersuchenden Texten auf Basis der Längen ihrer komprimierten Darstellung durchzuführen. Die genaue Arbeitsweise wird im Folgenden erläutert.

Ausgehend von einem unbekannten Dokument U sowie einem Referenzdokument K eines bekannten Autors werden beide Texte samt eines dritten Textes (der aus der Konkatenation von U und K entsteht) komprimiert. Dadurch entstehen drei Dateien x, y und xy. Im nächsten Schritt wird mithilfe einer einfachen Formel ein Ähnlichkeitswert d bzgl. der Dateien x, y und xy berechnet. Unterschreitet d einen zuvor festgelegten Schwellwert θ, so wird als Ergebnis eine übereinstimmende (andernfalls eine fremde) Autorschaft vorhergesagt. Um θ bestimmen zu können, wird der oben erwähnte Trainingskorpus benötigt. Konkret ergibt sich θ aus der Gleichfehlerrate (engl. Equal Error Rate, kurz EER). Visuell entspricht die EER demjenigen Punkt, an dem die Hauptdiagonale die ROC-Kurve schneidet (siehe Abbildung 6.8). Die Bestimmung des Schwellwerts θ ist im Wesentlichen der einzige Trainingsschritt von COAV.

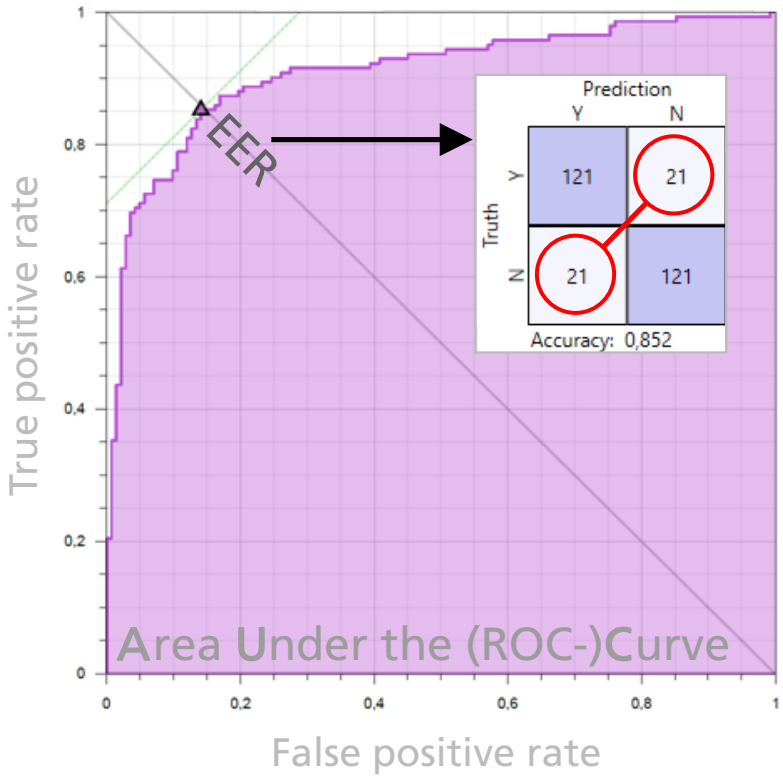


Abbildung 6.8:
COAV: Bestimmung des Schwellwerts auf Basis der Gleichfehlerrate.

Nachdem θ auf Basis des Trainingskorpus bestimmt wurde, haben wir COAV auf der Testmenge evaluiert. Hierbei erzielte COAV aus den 812 Verifikationsfällen 757 (= 384 TP + 373 TN) korrekte und 55 (= 33 FP + 33 FN) falsche Vorhersagen, was einer Erkennungsgenauigkeit von 93,2% entspricht.

Um zu überprüfen, ob COAV nicht irrtümlicherweise die Thematik der Texte klassifiziert, anstatt den Schreibstil zu betrachten, haben wir die Texte mithilfe des sogenannten Text-Distortion-Verfahrens (nach [83]) modifiziert. Ziel dieses Verfahrens ist es, inhaltsbasierte Texteinheiten zu maskieren, sodass am Ende hauptsächlich grammatikalische Strukturen in den modifizierten Texten verbleiben (siehe Abbildung 6.9).

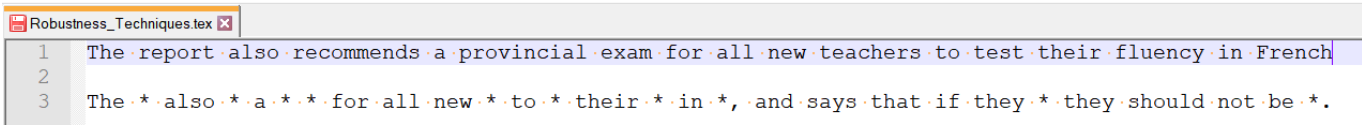


Abbildung 6.9:
COAV: Modifikation der Texte anhand von Text-Distortion.

Nachdem sämtliche Cyber-Groomer-Texte modifiziert wurden, haben wir COAV erneut trainiert und anschließend evaluiert. Hierbei erzielte COAV aus den 812 Verifikationsfällen 654 (= 293 TP + 361 TN) korrekte sowie 158 (= 45 FP + 113 FN) falsche Vorhersagen, was einer Erkennungsgenauigkeit von immer noch 80,54% entspricht. Somit kann aufgefasst werden, dass COAV tatsächlich stilistische Elemente für seine Entscheidung betrachtet und sich nicht primär auf Inhaltswörter stützt.

7 Umsetzbarkeit

In diesem Kapitel fassen wir unsere Erkenntnisse hinsichtlich der technischen Umsetzung verschiedener Aspekte der Studie zusammen. Ziel ist es, eine Grundlage für die darauf folgenden Handlungsempfehlungen zu schaffen. Wir schlagen hier die Brücke zwischen der wissenschaftlichen Betrachtung der Basistechnologien und den Anwendungen, welche mit diesen Technologien erweitert werden können.

7.1 Sexting-Erkennung auf Smartphones

Die in dieser Studie vorgeschlagene Lösung zur Erkennung von Sexts basiert auf einem neuronalen Netz. Um sie in der Praxis einsetzen zu können, muss sie unserer Meinung nach lokal auf dem Smartphone ausgeführt werden. Alternativ wäre auch denkbar, einen Server zu verwenden, welcher Bilder bewerten kann. Diese müssten dann vom Smartphone gesendet werden. In Abbildung 7.1 werden beide Konzepte gegenübergestellt.

Problematisch bei dem Einsatz eines Servers sind unserer Meinung nach allerdings mehrere Punkte:

- Die Privatsphäre des Nutzers würde durch das Versenden an den Server potenziell gefährdet werden.
- Ohne Netz kann entweder die Kamera nicht genutzt werden oder die Sexting-Erkennung ist inaktiv.
- Da alle Fotos (und natürlich auch Videos) an den Server geschickt würden, könnte eine hohe Datenlast entstehen, die schnell die Kontingente schneller Datenvolumen verbraucht.

Um auf dem Smartphone direkt eingesetzt zu werden, muss das Prüfen durch das NN annähernd in Echtzeit geschehen, um die Nutzerakzeptanz nicht zu gefährden. Um dies zu verifizieren, haben wir eine Demonstrator-App erstellt, welche auf einem Smartphone aufgenommene Bilder sowohl online (also direkt aus der Kamera heraus) als auch offline (aus einer gespeicherten Datei) prüfen kann.

Wichtig hierbei ist eine grundsätzliche Anmerkung: Auf dem Smartphone geschieht nur die Prüfung von Bildern durch das neuronale Netz. Das Training dieses Netzes ist deutlich aufwendiger als seine Anwendung; hierfür werden üblicherweise Cluster von Grafikkarten eingesetzt. Das Ergebnis des Trainings, das neuronale Netz, wird dann auf das Smartphone übertragen. Ein Training auf dem Smartphone selbst ist unserer Meinung nach derzeit nicht sinnvoll.

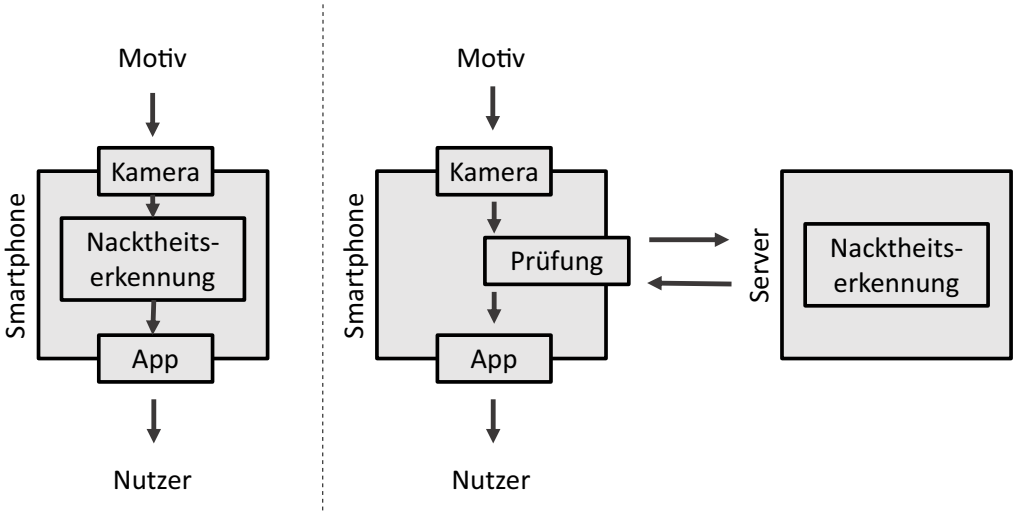


Abbildung 7.1:
Die Prüfung auf Nacktheit kann entweder lokal auf dem Smartphone (links) oder über einen Server (rechts) erfolgen.

7.1.1 Demonstrator-App

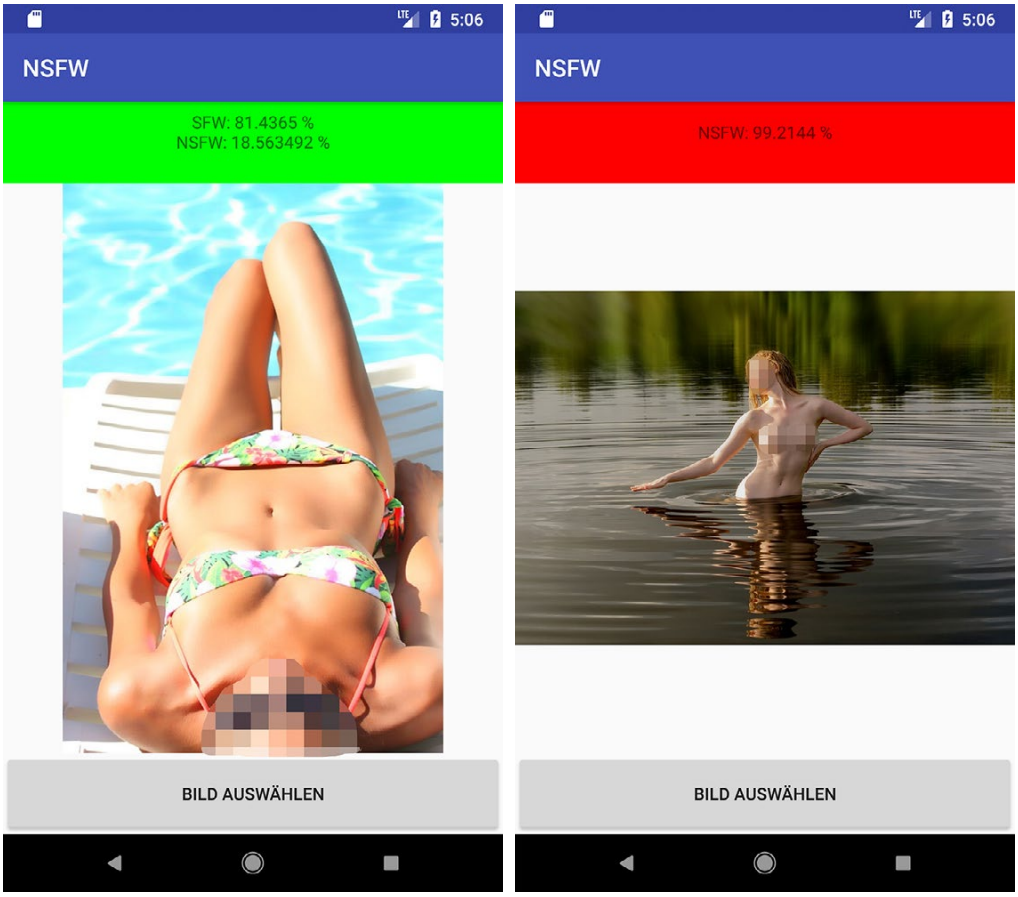
Um die Umsetzbarkeit der Sexting-Erkennung auf einem Smartphone zu veranschaulichen, wurde eine Android-basierte Demonstrator-App entwickelt. Diese verwendet ein CNN zur Erkennung von Nacktbildern. Bei Tests mit einem Huawei Mate 9 mit Android-Betriebssystem konnte die App Bilder eines Video-Streams in Echtzeit klassifizieren. Die Klassifizierung von Einzelbildern nahm nur wenige Millisekunden in Anspruch und verursachte keine wahrnehmbare Latenzzeit. Abbildung 7.2 zeigt Screenshots der App in einer Emulation eines Android-Betriebssystems auf einem PC. Die verwendeten Bilder stammen von Pixabay und sind frei verwendbar (CC0-Lizenz).

Einige Erfahrungen, die wir beim Erstellen der App gesammelt haben, sind auch für eine weitere Entwicklung von Bedeutung.

- Der Trainingsaufwand für die Erkennung kann reduziert werden, wenn als Grundlage ein bereits trainiertes Netz verwendet wird. Das eigene Training eines kompletten Netzes würde erfordern, Millionen von Bildern als Grundlage heranzuziehen, z. B. ImageNet.

- Zum Spezialisieren des Netzes auf die Aufgabe, Sexts zu erkennen, werden trotzdem noch tausende Trainingsbilder benötigt. Diese Trainingsbilder müssen sehr gut gelabelt sein. Fehler bei der Zuordnung und Beschriftung schlagen sich in Fehlern bei der Klassifizierung nieder.

- Auch für die verhältnismäßig geringe Menge von Bildern ist die Verwendung von Highend-GPUs dem Einsatz von Standard-CPUs vorzuziehen, da hier Geschwindigkeitssteigerungen um das zehn- bis zwanzigfache erreicht werden konnten. Der Beschaffungsaufwand lag in unserem Fall bei rund 2000 Euro in 2017.



(a) Klassifizierung als Nicht-Nacktbild (b) Klassifizierung als Nacktbild

Abbildung 7.2: Klassifizierung auf der emulierten App.

Die größte Herausforderung für eine Sexting-Erkennung ist eine Minimierung der Falsch-Positiven. Eine Kamera-App, die jedes zweite Katzenbild oder Strandfoto sofort blockiert, wird Akzeptanzprobleme haben. Ebenso ist aber auch eine zuverlässige Erkennung von Sexts unabhängig von der Umgebung wichtig. Bei dem Umgang mit der App auf dem Smartphone zeigte sich schnell, dass Schattenwurf, Gegenlicht und weitere Faktoren die Erkennung beeinflussen. Man kann also davon ausgehen, dass ein Minderjähriger, der ein Sext mit dem Gerät erstellen möchte, durch ein Experimentieren mit Beleuchtung und anderen Faktoren irgendwann erfolgreich sein wird. Die Kamera bzw. die App gibt durch die kontinuierliche Kontrolle sofort Rückmeldung über den Erfolg der Erkennung und kann somit als Orakel missbraucht werden.

7.1.2 Trainingsdaten

Die Wichtigkeit von Trainingsdaten wurde in dieser Studie schon mehrmals erwähnt. Eine zuverlässige Erkennung von Sexts ist technisch nur dann umsetzbar, wenn ausreichend Trainingsdaten vorhanden sind, die die folgenden Kriterien erfüllen:

- Beispiele für alle zu erkennenden Fälle sollten enthalten sein. Das betrifft Geschlecht, Hautfarbe, Alter, Grad der Nacktheit und eventuell weitere Faktoren.

- Die Trainingsdaten müssen mit Metadaten versehen sein, die das auf den Bildern zu sehende beschreiben.

- Die Beschreibung muss einheitlich sein. Ähnliche Begriffe sollten aufeinander abgebildet werden. Werden beispielsweise »Brust« und »Busen« zufällig verwendet, kann die Erkennungsrate sinken.

- Die Beschreibung muss fehlerfrei sein. Wird »Nacktheit« unterschiedlich bewertet, kann die Fehlerrate steigen.

In Abbildung 7.3 sind vier Beispiel aus Pixabay samt den dort vorhandenen Metadaten zu sehen. In 7.3a ist kein Hinweis auf Nacktheit enthalten, was zumindest einen Grenzfall darstellen sollte. Das Gleiche gilt für 7.3b, hier allerdings noch stärker. Hier fehlt jeglicher Hinweis auf »Busen« oder »Erotik«. In 7.3c wurden dafür »Brust« und »Akt« vergeben. Die letzte Abbildung 7.3d zeigt deutlich Nacktheit. Das Label »Nackt« wurde vergeben, »Akt« oder »Brust« hingegen nicht.

Ein Netz, welches mit diesen Bildern und Metadaten trainiert würde, hätte nun entweder das Problem, dass nur deutliche Nacktheit wie in 7.3d erlernt werden könnte oder es zu einem Verschwimmen des Konzeptes »Nacktheit« kommen würde. Würde neben »Nackt« auch »Akt« als Erkennungskriterium für Nacktheit hinzugezogen, so würde das Netz durch 7.3c fälschlich auch erotische Unterwäsche als Nacktheit klassifizieren.

Eine bereits existierende Sammlung von annotierten Bildern zum Thema »Sexting« ist den Verfassern nicht bekannt. Daher müsste analog zu der Evaluierung in dieser Studie eine eigene Sammlung von Sexts erstellt und annotiert werden. Dieser Aufwand ist notwendig, um eine zuverlässige Erkennung von Sexts zu erreichen. Die Anzahl von notwendigen Bildern ist abhängig von der Komplexität, die das Motiv »Sext« darstellen kann. In unserem Experiment haben wir mit wenigen tausend Bildern eine befriedigende Erkennungsleistung erreicht, die allerdings für die Praxis nicht ausreichen würde. Außerdem sind hier nur legale Bilder, die absichtlich im Internet veröffentlicht wurden, verwendet worden.

Experten zum Thema Jugendschutz müssten hier zuerst typische Fälle erarbeiten, die dann wieder das Spektrum der zu erkennenden Motive aufzeigen. Darauf basierend muss dann eine Abschätzung der zu erstellenden Trainingsdaten erfolgen. Dabei ist zu beachten, dass die juristischen Aspekte bezüglich des Bildmaterials eine entscheidende Rolle spielen. Ob eventuell vorhandene Bilder bei Ermittlern zum Training herangezogen werden dürfen, ist keine technische, sondern eine juristische Frage.

(a) Frau, Schönheit, Posieren, Unterwäsche, Sexy

(b) Mädchen, Blick, Denken, Porträt, Frau, Gesicht

(c) Frau, Akt, Brust, Dame, Blick, Erotik, Porträt, Gesicht

(d) Nackt Frau Erotik Körper Weiblichkeit Sexy

Abbildung 7.3:
Beispiele aus Pixabay, alle CC0.

7.2 Erkennung von Cybergrooming in Foren

Um die Ansätze aus 6.2 in der Praxis nutzen zu können, ist eine Integration der Erkennungsmechanismen in Foren, in denen Erwachsene Kinder potenziell ansprechen, die erfolgversprechendste Herangehensweise.

Hier kann ein trainiertes Netz die Texte der Teilnehmer automatisiert zusammenfassen und so eine ausreichend große Grundlage für Entscheidungen schaffen. Die Erkennung des Alters durch Profiling ist zwar fehleranfällig, diese Fehler bewegen sich aber üblicherweise im Rahmen von 5 Jahren. Dementsprechend ist das Unterscheiden von Kindern und älteren Erwachsenen möglich. Junge Erwachsene und ältere Kinder könnten allerdings durch diese Fehler verwechselt werden.

Aufgrund des Fehlerrisikos erscheint es ratsam, die Erkennung des Alters durch Profiling eher als einen Hinweisgeber für menschliche Moderatoren zu nutzen. Wird ihnen vom System eine auffällige Kommunikation gemeldet, beziehungsweise eine hohe Diskrepanz zwischen angegebenem und abgeleitetem Alter erkannt, so sollte der Moderator hier eine manuelle Sichtung der Kommunikation durchführen und gegebenenfalls weitere Schritte einleiten. Dies kann eine Warnung an das potenzielle Opfer, eine Verwarnung des möglichen Täters oder das Einschalten der Polizei sein.

Technisch dürfte die Integration in ein Forensystem keine Herausforderung sein. Das trainierte Netz würde kontinuierlich Kommunikationsverläufe einzelner Nutzer aggregieren, bewerten und mit den gemachten Altersangaben vergleichen. Zu prüfen ist allerdings, ob ein durchgehendes Profiling der Nutzer durch entsprechende Methoden rechtlich akzeptabel ist oder dem Datenschutz widerspricht.

Das Beschaffen von Trainingsdaten für eine Alterserkennung sollte bei einer Kooperation mit Forenbetreibern unproblematisch sein. Als Trainingsdaten können aus technischer Sicht alle Daten von Nutzern verwendet werden, die sich als nicht auffällig erwiesen haben.

Ebenfalls in 6.2 wird auch das Wiedererkennen von Autoren anhand ihres Schreibstils diskutiert. Diese Methode kann dabei helfen, Personen zu identifizieren, die bereits aus einschlägigen Gründen aus einem Forum verbannt wurden und sich nun unter einem neuen Pseudonym erneut anmelden wollen. Ebenso denkbar ist der Einsatz der Methode, um Personen, die unter mehreren Pseudonymen parallel in einem Forum aktiv sind, aufzudecken. Auch hier gilt allerdings wieder der Einsatz unter Vorbehalt des Datenschutzes.

Auch unabhängig vom Autor kann versucht werden, durch maschinelles Lernen die typische Vorgehensweise von Cybergrooming zu erlernen und erkennen. Allerdings ist hier die Herausforderung, geeignete Trainingsdaten zu nutzen, besonders hoch. Hier spielt der Schutz von Opfern und Tätern eine große Rolle. Aufzeichnungen von überführten Cybergrooming-Handlungen wären aber notwendig, um ein entsprechendes Training durchzuführen.

8 Handlungsempfehlungen

Dieses Kapitel fasst die Erkenntnisse zusammen, um Handlungsempfehlungen auszusprechen und eine Abschätzung des Aufwands ihrer Umsetzung zu liefern.

8.1 Sexting-Unterdrückung auf Smartphones

Um zu verhindern, dass ein Smartphone eines Minderjährigen verwendet wird, um ein Sext zu erstellen, können verschiedene Ansätze gewählt werden.

1. Kombination der Erkennung von Nacktheit und Alter: Hier greifen die Erkennung von Nacktheit durch ein trainiertes Neuronales Netz und die Altersbestimmung ineinander. Somit könnte automatisiert erkannt werden, wenn ein beliebiger Minderjähriger eine Nacktaufnahme von sich erstellt.

2. Kombination der Erkennung von Nacktheit und der Person: Wird die Nacktheitserkennung mit einer biometrischen Methode zum Erkennen einer Person kombiniert, wie sie beispielsweise bei Zugriffskontrollen eingesetzt wird, lässt sich ein Smartphone so einrichten, dass nur der Besitzer keine Sexts von sich erstellen kann.

3. Reine Nacktheitserkennung: Die Funktion beschränkt sich auf eine Erkennung von Nacktheit. Ein Smartphone kann dann keine Aufnahmen von Personen erstellen, die unbekleidet sind.

Offensichtlich haben diese Ansätze unterschiedliche Ausprägungen hinsichtlich ihrer Wirkung und ihrer Komplexität. Mit der Komplexität steigt auch die Wahrscheinlichkeit von Fehlern. Die einfachste Methode ist die reine Nacktheitserkennung. Es lässt sich argumentieren, dass ein Smartphone in der Hand von Minderjährigen, insbesondere Kindern, nicht dazu verwendet werden sollte, nackte Personen zu fotografieren. Probleme könnte höchstens eine zu strikte Erkennung sein, die beispielsweise auch verhindert, Kunstwerke aufzunehmen.

Insbesondere die Kombination von Nacktheitserkennung und einer Erkennung der Person ist hier eine offenere Variante. Eltern können bei der Bereitstellung des Smartphones an ihre Kinder dieses so trainieren, dass es den neuen Besitzer erkennt und Nacktaufnahmen von diesem verweigert. Der Nachteil hierbei ist natürlich, dass andere Personen nackt fotografiert werden können. Gelingt ein Umgehen der biometrischen Erkennung durch Verfremdung, so sind Nacktaufnahmen des Besitzers möglich.

Eine zuverlässige Verknüpfung der Erkennung von Nacktheit und Alter wäre ebenfalls ein guter Kompromiss zwischen Restriktion und Offenheit. Somit würde verhindert werden, dass Minderjährige sich gegenseitig nackt aufnehmen. Allerdings ist eine zuverlässige Erkennung des Alters technisch bisher schwer möglich und unterliegt starken Schwankungen. Dass Make-up künstlich verjüngt, ist hier eher unproblematisch. Fehler in die andere Richtung, also das Schätzen eines höheren Alters minderjähriger Personen, können allerdings zu einem Scheitern des Schutzes führen.

8.1.1 Integration

Eine Funktion wie die Sexting-Erkennung wird von den Minderjährigen vermutlich eher als Einschränkung und weniger als Schutz wahrgenommen. Würden Apps zur Auswahl stehen, die eine starke Kontrolle der Nutzung aufweisen und beispielsweise die Sexting-Erkennung beinhalten und andere Apps, die das nicht tun, würde wahrscheinlich die Wahl der Jugendlichen eher auf die unbeschränkten Apps fallen. Im Falle von Kommunikationsdiensten wird durch die Wahl der App zumeist auch festgelegt über welche Kanäle der Anwender kommunizieren kann. Ein einzelner Minderjähriger könnte dann nicht mit einer App mit Jugendschutz-Funktion mit Gleichaltrigen kommunizieren, die diese Funktion und damit die App ablehnen.

Der erfolgversprechendste Ansatz ist daher die Prüfung auf Nacktheit direkt in das Betriebssystem des Smartphones zu integrieren. Damit würde jedes Foto, welches mit dem Gerät gemacht wird, die Prüfung durchlaufen. Eine App kann dann nicht mehr entscheiden, ob sie den Jugendschutz umsetzt oder nicht.

In Abbildung 8.1 wird das Konzept aufgezeigt. Ein Nutzer öffnet (1) eine App, mit der er ein Foto aufnehmen möchte. Diese App startet (2) nun den Vorgang im Smartphone. Sie ruft das Betriebssystem auf (3), welches wiederum die Kamera startet (4). Diese erstellt eine digitale Aufnahme des Motivs und übergibt sie an das Betriebssystem. Dieses prüft nun das aufgenommene Bild zuerst mit der Nacktheitserkennung (5). Nun wird entweder das Foto an die App weitergegeben (6) oder eine Meldung an sie gesendet, dass das Foto nicht erstellt werden kann (7). Alternativ könnte das Foto auch durch einen Hinweis oder ein schwarzes Bild ersetzt werden, allerdings könnte im zweiten Fall der Nutzer auch von einer Fehlfunktion ausgehen.

Eine solche Lösung kann nur gemeinsam mit den Entwicklern der Smartphones und der entsprechenden Betriebssysteme erstellt werden, hat aber den Vorteil, dass keine Abstimmung mit einer Vielzahl von App-Anbietern nötig ist. Ein Foto schon beim Entstehen zu filtern und so schon das Speichern im Gerät selbst zu verhindern ist die sicherste Methode, zuverlässig das Entstehen ungewollter Sexts zu verhindern.

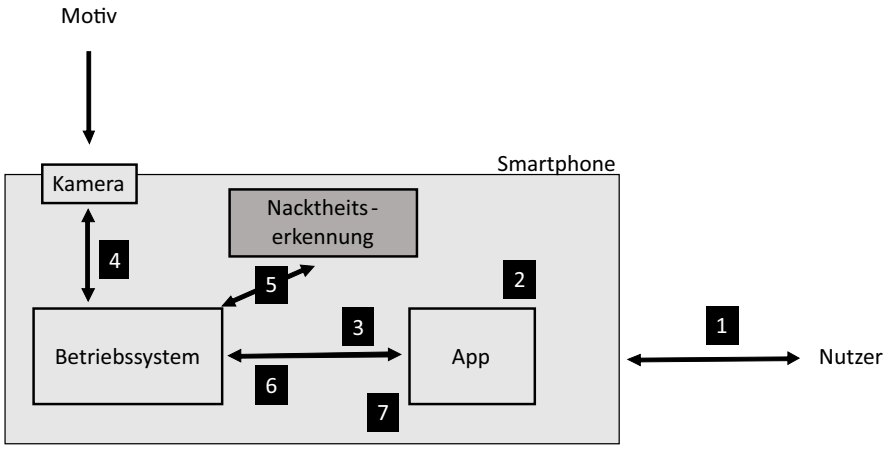


Abbildung 8.1:
Ein Nutzer erstellt ein Foto über eine App. Diese ruft dazu das Betriebssystem auf, welches automatisch das Motiv auf Nacktheit prüft.

8.1.2 Sammlung von Trainingsdaten

Nach derzeitigem Stand der Technik ist eine Erkennung von Sexts Minderjähriger am zuverlässigsten durch entsprechend trainierte neuronale Netze erreichbar. Wie bereits in Kapitel 7.1.2 diskutiert, dürfte das Erstellen einer ausreichend großen Sammlung entsprechender Daten, also Bilder und die zugehörige Annotation, eine der größten Herausforderungen bei der Umsetzung einer Lösung sein.

Wir raten dementsprechend, das Thema »Trainingsdaten« zumindest parallel zu den technischen und integrativen Fragestellungen zu behandeln, wenn eine Lösung für die Praxis gewünscht ist. Insbesondere ist eine frühzeitige juristische Prüfung der Sachlage wichtig, um Klarheit darüber zu erlangen, ob eine Sammlung von Sexts Minderjähriger überhaupt erstellt werden darf und, wenn dies bejaht werden kann, wer und unter welchen Bedingungen dies darf.

Da Jugendschutz und Internet ein internationales Problem ist, an dem viele Nationen forschen und arbeiten, kann hinsichtlich der Trainingsdaten auch eine internationale Kooperation mit Forschern und Behörden ratsam sein. Gegebenenfalls existieren entsprechende Trainingsdaten bereits in anderen Ländern.

8.1.3 Aufwand Kernverfahren

Hinsichtlich des Kerns der Sexting-Erkennung können wir eine grobe Abschätzung der notwendigen Personenaufwände abgeben. Diese setzt voraus, dass Trainingsdaten vorhanden sind und die Lösung in ein bekanntes Betriebssystem eines Smartphones integriert werden kann.

Der Aufwand bewegt sich dabei von einem Monat für eine einfache Lösung, die beispielsweise als Prototyp für Gespräche mit Herstellern von Smartphones oder Entwicklern von Betriebssystemen verwendet werden kann, bis zu neun Monaten, wenn die Lösung hinsichtlich Ressourcenverbrauch und Parametrisierung optimiert und evaluiert werden soll. Explizit ausgenommen sind hierbei die notwendigen Änderungen am Betriebssystem, welche durch den Hersteller erfolgen müssen, sowie das Erstellen der Trainingsdaten.

Soll eine ausführliche Optimierung und Evaluierung erfolgen, so ist wieder zu prüfen, wie diese geschehen kann. Dies könnte es notwendig machen, nicht-technisches Fachpersonal, welches im Umgang mit Minderjährigen geschult ist, mit einzubinden. Eventuell ist eine Optimierung durch die Technik in einer ersten Phase nur auf allgemeine Sexts sinnvoll und umsetzbar, während eine zweite Phase bei Erfolg der ersten durch entsprechende Experten für Jugendschutz angegangen werden muss.

8.2 Erkennung von Cybergrooming

Unsere Experimente in Kapitel 6.2 haben gezeigt, dass die Erkennung von Cybergrooming hinsichtlich realistischer Daten mithilfe von Autorenprofiling (AP) und Autorschaftsverifikation (AV)

funktioniert. Wichtig jedoch ist hierbei zu beachten, dass in unseren Experimenten entsprechende Trainingsdaten (also Daten, deren Kategorie im Vorfeld bekannt ist) vorlagen. Somit war es uns möglich, die Verfahren entsprechend zu justieren, um sie anschließend auf ungesehenen Daten testen zu können. Um solche Lösungen in Ermittlungsszenarien einsetzen zu können, bedarf es daher einer großen Menge an Daten, um die Verfahren vernünftig trainieren zu können. Idealerweise sollten diese von Experten klassifiziert werden, da sonst bei einer schlechten Datenqualität das verwendete ML-Verfahren fehlerhafte Klassifikationen ebenfalls erlernen könnte und damit unbrauchbar wird. Es existieren zwar einige wenige Ansätze für ML-Verfahren, die ohne jegliches Training auskommen, jedoch haben diese den markanten Nachteil, dass sie nicht optimiert werden können, da keine einstellbaren Parameter vorliegen.

Gilt es Cybergrooming direkt in Foren und Chats zu erkennen, so ist auch hier wieder wichtig zu verstehen, dass eine Datengrundlage benötigt wird, um die eingesetzten ML-Verfahren kontinuierlich optimieren zu können.

Bei der Erkennung von Cybergrooming können drei unterschiedliche Strategien verfolgt werden, die alle auf Verfahren aus Kapitel 6.2 basieren:

- Erkennung von Abweichungen bei Altersangaben: Hier wird mittels Profiling abgeleitet, welches Alter bei einem gegebenen Schreibstil angenommen werden kann. Dieses wird mit vorhandenen Angaben verglichen.

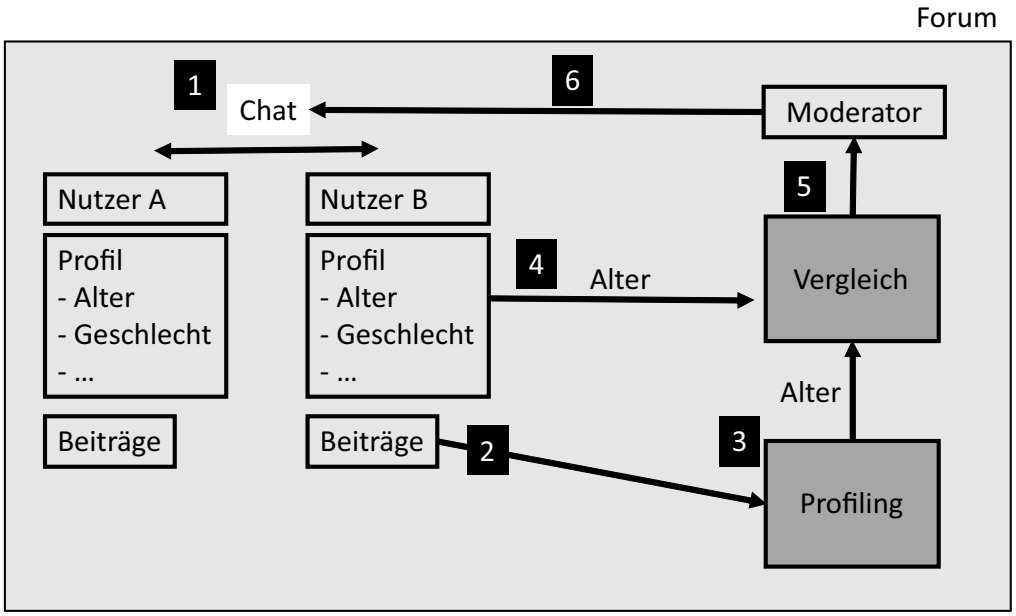


Abbildung 8.2:
Mittels Profiling kann eine Abweichung von angegebenem und abgeleitetem Alter erkannt werden und eine Warnung an den Moderator erfolgen.

- Wiedererkennen von Personen: Durch Autorschaftsattribution wird geprüft, ob ein Nutzer einen Schreibstil aufweist, der einem bereits bekannten Autor zugeordnet werden kann. Damit können beispielsweise bereits von Foren ausgeschlossene Nutzer wiedererkannt werden, wenn sie sich unter einem neuen Namen anmelden.

- Erkennen von Kommunikationsverläufen: Denkbar ist es auch, mittels maschinellem Lernen typische Vorgehensweisen beim Cybergrooming automatisiert zu erkennen. Dies könnte dann unabhängig von den Autoren geschehen. Es setzt allerdings umfangreiche Trainingsdaten voraus.

8.2.1 Integration

Auch zur Erkennung von Cybergrooming soll an dieser Stelle eine kurze Darstellung einer möglichen Integration, hier in ein Onlineforum, dargestellt werden. Diese ist schematisch in Abbildung 8.2 wiedergegeben. Zwei Nutzer kommunizieren in einem Chat miteinander (1). Zu jedem Benutzer ist bei der Anmeldung ein Profil erstellt worden, das Angaben wie Alter und Geschlecht enthält. Nun erzeugt der Nutzer Beiträge (2), die nach einiger Zeit ausreichend umfangreich sind, um daraufhin ein Profiling (3) durchzuführen. Dadurch wird ein Alter abgeleitet, welches dem Sprachstil des Nutzers entspricht. Dieses wird mit dem Alter aus dem Profil (4) verglichen. Stimmt es nicht überein und ist die Abweichung signifikant, so erhält der Moderator eine Warnung (5). Er kann nun den Chat überwachen und gegebenenfalls schützend eingreifen.

8.2.2 Sammeln von Trainingsdaten

Gerade in Bezug auf Sprache ist es wichtig, neue Trainingsdaten zu sammeln, da Sprache sich (insbesondere in Foren und Chats) permanent weiterentwickelt und die jeweiligen ML-Verfahren sonst nicht in der Lage wären sich anzupassen. Allerdings betrifft dies in erster Linie AP-Verfahren, da AV-Verfahren nicht auf die Thematik der Texte angewiesen sind (wie in Kapitel 6.2 gezeigt), um erfolgreiche Ergebnisse zu liefern. Hervorzuheben ist auch die Tatsache, dass die Anwendung von NNs auf Texten (im Gegensatz zu Bildern) nicht zwingend auf viele Daten angewiesen ist, um erfolgversprechende Ergebnisse zu liefern (dies wurde ebenfalls in Kapitel 6.2 gezeigt).

9 Exkurs: Kinderpornografie

Ein neben Cybergrooming durch die Digitalisierung der letzten Jahrzehnte ebenfalls rasant wachsendes Problem im Bereich der Pädokriminalität ist die Verbreitung von Kinderpornografie über das Internet. Da Technologien zur Unterstützung der Ermittlungsarbeit im Bereich Kinderpornografie eine starke Überschneidung zu denen der Sexting-Erkennung aufweisen und aufgrund der thematischen Nähe, bildet dieses Kapitel einen Exkurs zum Thema Kinderpornografie. Es liegt darüber hinaus nahe, dass beide Phänomene Hand in Hand miteinander gehen und einige Cybergrooming-Fälle in direktem Zusammenhang zur Kinderpornografie stehen.

9.1 Handel und Verbreitung

Das Tor-Netzwerk[1] (ugs. auch als »Darknet« bezeichnet) bietet durch das Prinzip des Onion-Routings[2] sowohl Anbietern als auch Konsumenten die Möglichkeit, weitestgehend unerkannt Kinderpornografie zugänglich zu machen bzw. zu beziehen. So können sie sich durch die Verwendung des entsprechenden Tor-Browsers polizeilichem Zugriff in vielen Fällen entziehen, wenn auch nicht vollständig, da auch Strafverfolgungsbehörden im Darknet aktiv sind und ermitteln. Aufgrund der Anonymität ist hier die Identifikation straffälliger Subjekte jedoch deutlich schwieriger, da die Ursprungsadressen von Datenpaketen nicht identifiziert werden können.

Wie Forscher der RWTH Aachen und der Goethe-Universität in Frankfurt am Main herausfanden, lassen sich kinderpornografische Inhalte oder Verweise auf Darknet-Webseiten mit solchen in der Blockchain[3] der Online-Währung Bitcoin nachweisen [84].

9.2 Probleme bei der Ermittlungsarbeit

In Gesprächen mit verschiedenen Landeskriminalämtern wurde auf Vorgehensweisen sowie Herausforderungen bei der Sicherstellung kinderpornografischen Bild- und Videomaterials eingegangen. Dabei zeigte sich, dass das Hauptproblem, vor dem die Ermittler behördenübergreifend stehen, die große Menge beschlagnahmter Daten ist. Werden in einem Verdachtsfall auf den Besitz von oder Handel mit Kinderpornografie Datenträger beschlagnahmt, müssen diese von Beamten nach relevantem Bild- bzw. Videomaterial durchsucht werden. Dieser Prozess wird als Sichtung bezeichnet.

Die zunehmende Größe von Datenträgern ist diesbezüglich für die Ermittler zu einem großen Problem geworden. So sind externe Festplatten mit 8 TB Speicher für unter 200 Euro erhältlich. Bei einer durchschnittlichen Größe von 1,5 MB lassen sich über 5 Millionen Bilder auf einer solchen

1 Akronym für The Onion Routing

2 Anonymisierungsverfahren für Internetverkehr

3 Eine Blockchain ist eine Liste von Datenblöcken, die in einer festen Reihenfolge miteinander verkettet sind. Sie wird bspw. von Kryptowährungen wie Bitcoin als dezentrales Buchführungssystem genutzt.

Festplatte speichern. Selbst das iPhone X von Apple ist mit einem Speicher von 256 GB verfügbar, was Platz für etwa 170.000 Bilder gleicher Größe bietet. Die Sichtung solch großer Datenbestände beansprucht die Ermittler sowohl zeitlich als auch psychisch intensiv. Da die Konzentrationsfähigkeit mit zunehmender Dauer der monotonen Arbeit stark abnimmt, ist auch das unabsichtliche Übersehen relevanter Bilder oder Videos ein Problem.

In einem Artikel des Bundeskriminalamts wird überdies darauf hingewiesen, dass verpflichtende Mindestspeicherfristen für Internet-Verkehrsdaten von 6 Monaten (Vorratsdatenspeicherung) in vielen Fällen notwendig ist, um Personen, die online Kinderpornografie veröffentlichen oder beziehen, ausfindig zu machen [85]. Wie bereits in Abschnitt 3.1 erwähnt, wurde die Wiedereinführung der Vorratsdatenspeicherung bereits beschlossen, jedoch wären laut Bundeskriminalamt in vielen Fällen deutlich länger Speicherfristen erforderlich, da die Behörde selbst zwar innerhalb von sieben Tagen tätig werde, verbotenes Material jedoch häufig eine gewisse Zeit unentdeckt bleibe und eine Vorratsdatenspeicherung von wenigen Wochen somit nicht ausreichend sei. Im Jahr 2017 seien die Ermittler mangels Verfügbarkeit von Verkehrsdaten in 8400 mutmaßlichen Fällen von Kinderpornografie der Täter nicht habhaft geworden [86].

9.3 Aktuell eingesetzte technische Hilfsmittel

Nach dem aktuellen Kenntnisstand der Autoren sind derzeit drei Arten von technischen Hilfsmitteln im Einsatz, die bei der Sichtung großer Bildmengen unterstützend eingesetzt werden.

9.3.1 Bildbetrachtungsprogramme

Dies sind zum einen spezielle Bildbetrachtungsprogramme, die eine bestimmte Menge von Bildern gleichzeitig anzeigen und so den Durchsatz gesichteter Bilder pro Zeiteinheit erhöhen. Hierbei ist jedoch anzumerken, dass gerade durch das gleichzeitige Betrachten von Bildern die Wahrscheinlichkeit, ein relevantes Bild zu übersehen, steigt.

9.3.2 Kryptografische Hash-Verfahren

Kryptografische Hash-Verfahren werden eingesetzt, um bereits bekannte kinderpornografische Inhalte einfach wiederzuerkennen. Der Hash-Wert einer Datei wird mittels einer Hash-Funktion berechnet und bildet eine eindeutige Prüfsumme, anhand derer die Datei (bzw. ein Duplikat, d. h. eine Datei mit identischem Inhalt) identifiziert werden kann. Er kann als eine Art Fingerabdruck einer Datei betrachtet werden und kann dazu genutzt werden, um Dateien mit bestimmtem Inhalt wiederzufinden. Um bekannte Bilder mit kinderpornografischem Inhalt schnell zu identifizieren, verwaltet das Bundeskriminalamt die sogenannte PERKEO[4]-Datenbank, welche Hash-Werte bekannter kinderpornografischer Bilder enthält, und bietet den Behörden auf Landesebene über eine Schnittstelle Zugriff darauf. So können diese in einem schnellen Scan-Verfahren beschlagnahmte Datenträger auf das Vorhandensein bekannter kinderpornografischer Inhalte untersuchen. Bilder,

4 Abk. für Programm zur Erkennung Relevanter Kinderpornografischer Eindeutiger Objekte

deren Hash-Wert nicht in der PERKEO-Datenbank vorhanden ist, werden so jedoch nicht gefunden.

Ein generelles Problem beim Einsatz kryptografischer Hash-Werte zum automatischen Auffinden von Kinderpornografie ist die Tatsache, dass selbst kleinste Änderungen an einer Datei (hier genügt das erneute Speichern eine JPEG-Datei) ihren Hash-Wert vollkommen verändern. Dies hat zur Folge, dass ein Scan-Verfahren mit der PERKEO-Datenbank sehr einfach umgangen werden kann, indem sämtliche auf dem Datenträger befindliche einschlägige Bilder explizit neu abgespeichert werden. Das von einem Beamten des Landeskriminalamt Niedersachsen entwickelte Programm URANOS kombiniert beide Funktionalitäten miteinander ein geeignetes Bildbetrachtungsprogramm sowie den Zugriff auf die PERKEO-Datenbank. Das Programm und wurde den Bundesländern kostenlos zur Verfügung gestellt [87].

9.3.3 Robuste Hash-Verfahren

Im Gegensatz zu kryptografischen Hash-Verfahren sind robuste Hash-Verfahren ihrem Namen nach robust gegenüber Manipulationen der Ausgangsdatei. Das bedeutet, dass selbst bei einigen Änderungen an einem Bild die entsprechende Funktion dennoch den gleichen Hash-Wert errechnet und somit das veränderte Bild nach wie vor über den Hash-Wert identifiziert werden kann. Software-Lösungen auf Basis robuster Hash-Verfahren zur Erkennung bekannter Kinderpornografie sind das vom Fraunhofer-Institut für Sichere Informationstechnologie und Partnern entwickelte ForBild[5] sowie das von Microsoft entwickelte PhotoDNA[6].

9.4 Verbesserung von Sichtungsprozessen durch Neuronale Netze

Die bisher eingesetzten technischen Hilfsmittel für die Sichtung großer Datenbestände erleichtern den Ermittlern zwar ihre Arbeit zu einem gewissen Grad, aber sie können das grundlegende Problem der nicht handhabbaren Datenmenge nicht ausreichend lösen. So müssen Datenträger selbst nach einem Hash-Vergleich vollständig manuell untersucht werden, um das Vorhandensein unbekannter oder veränderter relevanter Bilder zu überprüfen. Ein auf maschinellem Lernen basierendes Verfahren könnte die Sichtung dagegen – besonders im Hinblick auf die benötigte Zeit sowie die Konzentrationsfähigkeit des Ermittlers – weitaus effizienter unterstützen. Ebenso, wie auf Nacktbildern trainierte KNNs zur Sexting-Erkennung eingesetzt werden können, wäre es möglich, KNNs auch auf die Erkennung von Kinderpornografie zu trainieren.

Sichtungsprozesse könnten dadurch wie folgt unterstützt werden: Die Bilder innerhalb eines Sichtungsprozesses werden anhand der Klassifizierungswerte des KNN sortiert. Bilder von potenziell relevanter Natur stehen ganz oben auf der Liste; diejenigen von geringerer Relevanz sind am unteren Ende. Dies hilft schnell zu entscheiden, ob illegale Inhalte auf dem betreffenden Datenträ-

5 http://sit4.me/forbild

6 https://www.microsoft.com/en-us/photodna

ger gespeichert sind. Der Ermittler kann seine begrenzte Konzentration auf jene Bilder verwenden, die mit einer hohen Wahrscheinlichkeit relevant sind. Dazu wird die Annahme getroffen, dass eine größere Sammlung gemischter Bilder mit unterschiedlichen Inhalten vorliegt, die über verschiedene Verzeichnisse verteilt sind und keine aussagekräftigen Namen aufweisen. Diese Sammlung enthält willkürlich gemischt eine Teilmenge A mit Bildern mit kinderpornografischem Inhalt und eine andere Teilmenge B mit den restlichen Bildern. Die Teilmenge B ist somit Rauschen, welches das Erkennen der Bilder aus Teilmenge A erschwert. Wird dem Ermittler nun zuerst die relevante Teilmenge A gezeigt, so wird damit die Ablenkung vermieden, die durch ein gemischtes Betrachten von A und B erzeugt würde. Wenn das System sogar automatisch erstellte Metadaten zur Verfügung stellen kann, die helfen, die Relevanz der einzelnen Bilder zu quantifizieren, zum Beispiel basierend auf angenommenem Alter oder sexueller Aktivität, kann das Sortieren noch hilfreicher sein, da besonders relevante Bilder innerhalb von A oben auf der Liste stehen würden. Eine Untersuchung einer riesigen Bildersammlung mit nur einem geringen Anteil von Pornografie würde dadurch erheblich unterstützt. Eine vereinfachte Darstellung des Konzepts wird in Abbildung 9.1 dargestellt. Das Verhältnis der Mengen A und B beträgt hier nur 1:10, dementsprechend übersichtlich fällt die Illustration aus. Kritisch kann vor allem die falsche Zuordnung eines illegalen Inhalts in die Gruppe der legalen Inhalte in Abbildung 9.1b sein.

Die in Abschnitt C.4 beschriebenen Angriffsszenarien auf ein ML-basiertes System gelten auch für das in diesem Kapitel vorgestellte Verfahren einer Bildsortierung durch ein KNN. Um zu verhindern, dass kinderpornografisches Material durch den Algorithmus sehr niedrig bewertet wird, d. h. am unteren Ende der Liste einsortiert wird, muss das trainierte KNN zur Erkennung von Kinderpornografie ausgiebig getestet werden. Es muss insbesondere sichergestellt werden, dass durch einfache Bildmanipulationen (z. B. das Hinzufügen von Rauschen) keine unvorteilhafte Sortierung der Bilder zustande kommt.

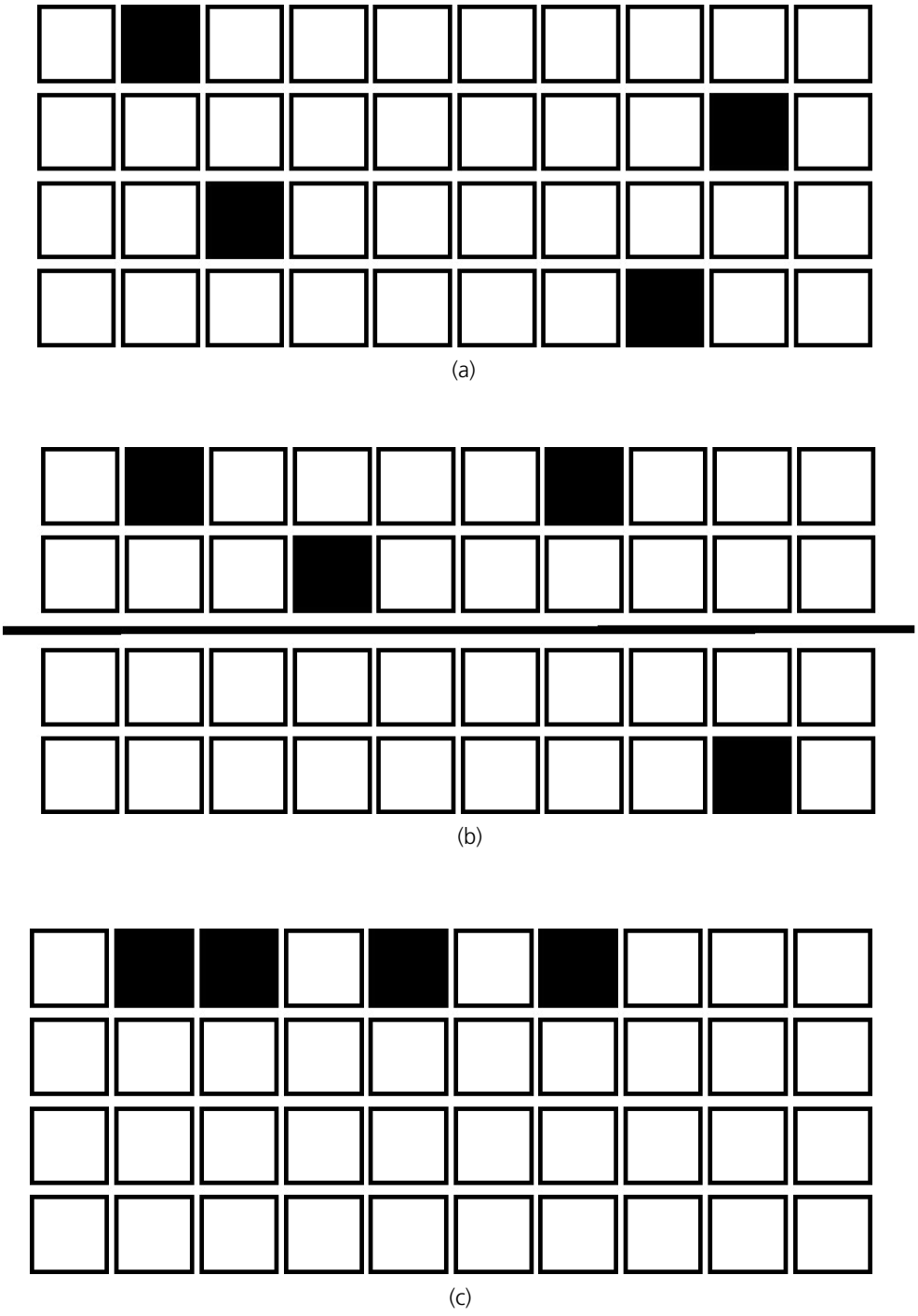


Abbildung 9.1:
Illustration des Konzepts: (a) Zufällige Verteilung legaler (weiß) und kinderpornografischer (schwarz) Inhalte. 10% der Inhalte sind hier Kinderpornografie; (b) Variante 1: Bei einer automatisierten Klassifizierung werden legale und kinderpornografische Inhalte in zwei Gruppen eingeteilt. Durch falsch-positive und falsch-negative Zuordnungen geraten legale Inhalte in die Kategorie »Kinderpornografie« und umgekehrt; (c) Variante 2: Anstelle einer expliziten Klassifikation wird eine Sortierung der Bilder anhand ihrer Wahrscheinlichkeitswerte für enthaltene Kinderpornografie vorgenommen. Die zu erwartende Wahrscheinlichkeit kinderpornografische Inhalte zu finden, ist somit am Anfang der sortierten Liste am höchsten, während sie mit zunehmendem Fortschreiten in der Liste abnimmt. Es werden dabei keine Gruppen gebildet und die Liste enthält sämtliche auf dem Datenträger vorhandene Bilder.

10 Zusammenfassung

Durch das stetige Wachstum der digitalen Welt und die rasche Zunahme der Erdbevölkerung, kann man davon ausgehen, dass auch die Anzahl der sexuellen Übergriffe über digitale Medien wachsen wird. Da auch immer mehr Kinder und Jugendliche das Internet und dessen Möglichkeiten nutzen, ist es zwingend erforderlich, Maßnahmen zu ergreifen, um diese vor (pädophilen) Straftätern zu schützen. In dieser Studie wurden zwei technische Maßnahmen betrachtet, die dabei helfen können, Minderjährige vor sexuellen Übergriffen zu schützen.

Dazu wurden zuerst die Themenfelder Sexting und Cybergrooming ausführlich betrachtet. Danach wurde nach einer kurzen Darlegung der technischen Grundlagen der Stand der Technik erörtert, der geeignete Technologien zur Bekämpfung der Phänomene aufzeigt und die Grundlage für die darauf folgende Eignungsprüfung ausgewählter Methoden schafft.

Wir zeigten in 6.1 wie Deep Learning dabei helfen kann, Nacktaufnahmen von Jugendlichen automatisiert zu klassifizieren und welche Vorteile diese Herangehensweise gegenüber herkömmlichen pixelbasierten Lösungen aufweist. In 6.2 legten wir dar, wie Methoden der Computerlinguistik in Kombination mit maschinellem Lernen zur Alterserkennung und zum Identifizieren von Personen mit unterschiedlichen Pseudonymen verwendet werden können.

Neben der rein technischen Betrachtung erörterten wir in den Abschnitten 7.1 und 7.2 auch, wie der Einsatz in der Praxis möglich wäre. Unsere Handlungsempfehlungen zeigten dann, wie die in der Studie untersuchten Technologien in Smartphones (in Abschnitt 8.1) und in Foren (in Abschnitt 8.2) eingebracht werden können. Hier wurde klar, dass ein Einsatz nur im Zusammenspiel mit weiteren Parteien effektiv möglich ist.

Wir sind zu dem Schluss gekommen, dass bei einer Zusammenarbeit mit den entsprechenden Herstellern und Betreibern aus technischer Sicht höchstens ein ausreichender Fundus an Trainingsdaten eine Herausforderung darstellt. Notwendig ist allerdings eine rechtliche Prüfung, die in dieser Studie nicht durchgeführt wurde. Damit reiht sich die Frage, wie Missbrauch von Minderjährigen im Internet bekämpft werden kann, in zahlreiche andere Fragestellungen im Kontext von maschinellem Lernen und Big Data ein.

Ideal wäre eine interdisziplinäre Betrachtung, die neben den technischen und (datenschutz-)rechtlichen Aspekten auch beispielsweise wirtschaftliche, soziologische oder psychologische Aspekte einbezieht. Auch eine kriminologische Perspektive würde für eine ganzheitliche Betrachtung in Frage kommen.

Immer wieder ist im Laufe der Studie das Thema »Kinderpornografie« neben Sexting als Gegenstand automatisierter Erkennung angesprochen worden. In beiden Fällen sind Altersabschätzung und Nacktheitserkennung von besonderer Bedeutung. Dementsprechend haben wir am Ende der Studie einen Exkurs zum Thema eingefügt, welcher primär die Unterstützung bei der Sichtung von potenziell kinderpornografischem Material adressiert.

11 Literatur

[1] Döring, Nicola: Erotischer Fotoaustausch unter Jugendlichen: Verbreitung, Funktionen und Folgen des Sexting. In: Zeitschrift für Sexualforschung, Band 25, Seiten 4–25, Stuttgart, 2012. Georg Thieme Verlag.

[2] Döring, Nicola: Sexting. Aktueller Forschungsstand und Schlussfolgerungen für die Praxis, Seiten 15–43. Bundesarbeitsgemeinschaft Kinder- und Jugendschutz e.V., Berlin, 2015, ISBN 978-3-00-049233-4.

[3] Feierabend, Sabine, Theresa Plankenhorn und Thomas Rathgeb: Jugend, Information, (Multi-) Media, JIM 2015. Medienpädagogischer Forschungsverbund Südwest (mpfs), Stuttgart, 2015.

[4] Döring, Nicola: Sexting. Fakten und Fiktionen über den Austausch erotischer Handyfotos unter Jugendlichen. In: merz. medien + erziehung, Band 56, Seiten 47–52, München, 2012. kopaed Verlag.

[5] Döring, Nicola: Consensual sexting among adolescents: Risk prevention through abstinence education or safer sexting? In: Cyberpsychology: Journal of Psychosocial Research on Cyberspace, Band 8, Jostova, CZE, 2014. Masaryk University, Faculty of Social Studies.

[6] Klettke, Bianca, David Hallford und David Mellor: Sexting prevalence and correlates: A systematic literature review. In: Clinical Psychology Review, Band 34, Seiten 44–53, Amsterdam, NL, 2014. Elsevier.

[7] WeltN24 GmbH: Sexting: Wenn das Verschicken von Nacktbildern tragisch endet. Welt, Online-Artikel, 01.11.2013.

[8] Lorenz, Sabine: Anonymous will Amanda Todd rächen. Frankfurter Rundschau, Online-Artikel, 19.10.2012.

[9] Saferinternet.at: Sexting in der Lebenswelt von Jugendlichen. Studie, Saferinternet.at, 2015.

[10] Koordination klicksafe.de: Sexting - was ist das? klicksafe.de, 06.06.2014. http://www.klicksafe.de/nc/themen/problematische-inhalte/sexting/sexting-was-ist-das, besucht: 16.11.2016.

[11] Kalenda, Florian: FBI macht Stimmung gegen anonyme Messaging-App Kik. ZDNet, 08.02.2016. http://www.zdnet.de/88259329/fbi-macht-stimmung-gegen-anonyme-messaging-app-kik/, besucht: 16.11.2016.

[12] HPMG News: Sexting: Instagram verbannt Auberginen-Emoji. The Huffington Post, 30.04.2015.

[13] Herbig, Daniel: Facebook Messenger: Verschlüsselte Chats verfügbar. Heise Online, Online-Artikel, 27.09.2016. https://www.heise.de/newsticker/meldung/Facebook-Messenger-Verschluesselte-Chats-verfuegbar-3332724.html, besucht: 16.11.2016.

[14] Leavitt, Trent: SNAPCHAT UNVEILED: AN EXAMINATION OF SNAPCHAT ON ANDROID DEVICES. www.decipherforensics.com, 23.01.2014. http://www.decipherforensics.com/snapchat/, besucht: 12.01.2017.

[15] Mathiesen, Asbjørn: Cybermobbing und Cybergrooming. In: Jahrbuch des Kriminalwissenschaftlichen Instituts der Leibniz Universität Hannover. Kriminalwissenschaftliches Institut der Leibniz Universität Hannover, 2014.

[16] Katzer, Catarina: Gefahr aus dem Netz - Der Internet-Chatroom als neuer Tatort für Bullying und sexuelle Viktimisierung von Kindern und Jugendlichen. Dissertation, Universität zu Köln, 2007.

[17] Kampa, Patrick: Cybergrooming: Sexueller Missbrauch von Kindern via WhatsApp. Rechtsindex, 14.01.2016. http://www.rechtsindex.de/strafrecht/5446-Cybergrooming-sexueller-missbrauch-von-kindern-via-whatsapp, besucht: 19.12.2016.

[18] Burgert, Julian: Kritik an ungenauen Vorschriften: Sachverständige plädieren für Nachbesserungen bei der geplanten Reform des Sexualstrafrechts. Deutscher Bundestag, 20.10.2014.

[19] Bundeskriminalamt: Fallentwicklung und Aufklärung bei Tatmittel Internet. Polizeiliche Kriminalstatistik, 2013–2017.

[20] Stadler, Lena, Steffen Bieneck und Christian Pfeiffer: Repräsentativbefragung Sexueller Missbrauch 2011. Kriminologisches Forschungsinstitut Niedersachsen e.V., Hannover, 2012.

[21] Balzer, Vladimir und Axel Rahmlow: Warum "Cybergrooming" so erfolgreich ist. Deutschlandradio Kultur, 28.09.2016.

[22] Osterheider, Michael und Janina Neutze: Mikado - Missbrauch von Kindern: Aetiologie, Dunkelfeld, Opfer. 17.09.2015.

[23] Rüdiger, Thomas-Gabriel: Cybergrooming in virtuellen Welten - Chancen für Sexualtäter? In: Zeitschrift der Gewerkschaft der Polizei, Band 2, Seiten 29–35, Hilden, 2012. Gewerkschaft der Polizei.

[24] Arnsperger, Malte: Erst Chat, dann Vergewaltigung. Stern Online, 10.12.2008.

[25] Mackensen, Gisela: Eingeschleimt, angemacht, verurteilt. Stern Online, 22.12.2008.

[26] Reinsch, Melanie: Missbrauch via Spiele-Chat. Frankfurter Rundschau, 21.10..2016.

[27] ARD: Wie Sie Ihr Kind vor Cybergrooming schützen können. DasErste.de, 28.09.2016. https://www.daserste.de/unterhaltung/film/themenabend-cyber-grooming/interviews/interview-nina-luebbesmeyer100.html, besucht: 20.12.2016.

[28] mbH, Stuttgarter Zeitung Verlagsgesellschaft: 45-Jähriger muss mehrere Jahre in Haft. Stuttgarter-Zeitung.de, 28.07.2016.

[29] dpa: Missbrauch: 45-Jähriger lockte Mädchen im Netz an. Neue Osnabrücker Zeitung, 02.06.2016.

[30] Holland, Martin: Cyber Grooming: Prozess gegen Düsseldorfer beginnt. heise online, 02.12.2016.

[31] Bach, Solveig: Cybergrooming über "Minecraft": Täter im Fall Paul droht lange Haft. heise online, 02.12.2016.

[32] Feierabend, Sabine, Theresa Plankenhorn und Thomas Rathgeb: KIM-Studie 2014: Kinder + Medien, Computer + Internet. Medienpädagogischer Forschungsverbund Südwest (mpfs), Stuttgart, 2016.

[33] Beer, Kristina: Kinder fühlen sich bei Cybergrooming schnell schuldig. heise online, 26.09.2016.

[34] jugendschutz.net: Sexuelle Belästigung - Einfallstor Messenger. http://www.jugendschutz.net/cybermobbing-sexuelle-belaestigung/, besucht: 13.12.2016.

[35] Sulake Corporation Oy: Welche Funktionen verwendet Habbo, um die Sicherheit meines Kindes zu garantieren, wenn es Habbo Hotel besucht? Habbo.de Help Tool. https://help.habbo.de/hc/de/articles/221254607-Welche-Funktionen-verwendet-Habbo-um-die-Sicherheit-meines-Kindes-zu-garantieren-wenn-es-Habbo-Hotel-besucht-, besucht: 22.12.2016.

[36] Carr, John: Viewpoint: What went wrong at Habbo Hotel? BBC News, 14.06.2012.

[37] Livingston, Anne: What Parents need to know about Clash of Clans. https://kidsprivacy.net/2014/07/09/parents-clash-of-clans/, besucht: 22.12.2016.

[38] Williams, Amanda: Internet paedophile used mobile phone game Clash of Clans to groom boy, nine. Mail Online, 05.01.2016.

[39] WindowsArea.de: Minecraft zählt nun 74 Millionen aktive Spieler, 144 Millionen Verkäufe. windowsarea.de, 21.01.2018.

[40] Hausding, Matthias: Der Sex-Täter spielt mit. Märkische Online Zeitung, 11.04.2016.

[41] Ap-Apid, Rigan: An Algorithm for Nudity Detection. In: Proceedings of the 5th Philippine Computing Science Congress (PCSC), Seiten 201–205, Cebu City, PH, 2005.

[42] Jones, Michael J. und James M. Rehg: Statistical Color Models with Application to Skin Detection. In: Technical Report Series, Cambridge, MA, USA, 1998. Cambridge Research Laboratory.

[43] Santos, Clayton, Eulanda M. dos Santos und Eduardo Souto: Nudity Detection Based on Image Zoning. In: 11th International Conference on Information Sciences, Signal Processing and their Applications (ISSPA), Band 11, Seiten 1098–1103, Piscataway, New Jersey, USA, 2012. IEEE.

[44] Lee, Jiann–Shu, Yung–Ming Kuo, Pau–Choo Chung und E–Liang Chen: Naked image detection based on adaptive and extensible skin color model. In: Pattern Recognition, Band 40, Seiten 2261–2270, New York, NY, USA, 2007. Elsevier.

[45] Platzer, Christian, Martin Stuetz und Martina Lindorfer: Skin Sheriff: A Machine Learning Solution for Detecting Explicit Images. In: Proceedings of the 2nd International Workshop on Security and Forensics in Communication Systems, SFCS '14, Seiten 45–56, New York, NY, USA, 2014. ACM.

[46] Roheda, Siddharth: A Multi-Scale Approach to Skin Pixel Detection. In: Electronic Imaging, Band 2017, Seiten 18–23, 2017.

[47] Krizhevsky, Alex, Ilya Sutskever und Geoffrey E. Hinton: ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012.

[48] He, Kaiming, Xiangyu Zhang, Ren Shaoqing und Sun Jian: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: ICCV, Seiten 1026–1034, Washington, DC, USA, 2015. IEEE.

[49] Saito, Masaki und Yusuke Matsui: Illustration2Vec: A Semantic Vector Representation of Illustrations. In: SIGGRAPH Asia 2015 Technical Briefs, Seiten 5:1–5:4, New York, NY, USA, 2015. ACM.

[50] Almeida, Verónica, Malay Kishore Dutta, Carlos M. Travieso, Anushikha Singh und Jesús B. Alonso: Automatic Age Detection Based on Facial Images. In: Communication Control and Intelligent Systems, 2016.

[51] Rothe, Rasmus, Radu Timofte und Luc Van Gool: Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. In: International Journal of Computer Vision, 2016.

[52] Ranjan, Rajeev, Swami Sankar, Carlos D. Castillo und Rama Chellappa: An All-In-One Convolutional Neural Network for Face Analysis. In: IEEE FG, Band 12, 2017.

[53] Halvani, Oren, Martin Steinebach und Svenja Neitzel: Lässt sich der Schreibstil verfälschen um die eigene Anonymität in Textdokumenten zu schützen? In: Katzenbeisser, Stefan, Volkmar Lotz und Edgar R. Weippl (Herausgeber): Sicherheit 2014: Sicherheit, Schutz und Zuverlässigkeit, Beiträge der 7. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI), 19.-21. März 2014, Wien, Österreich, Band 228 der Reihe LNI, Seiten 229–241. GI, 2014, ISBN 978-3-88579-622-0. http://subs.emis.de/LNI/Proceedings/Proceedings228/article38.html.

[54] Deza, M.M. und E. Deza: Encyclopedia of Distances. . Springer Berlin Heidelberg, 2009, ISBN 9783642002342. https://books.google.de/books?id=LXEezzccwcoC.

[55] Jiang, Zhile, Shuai Yu, Qiang Qu, Min Yang, Junyu Luo und Juncheng Liu: Multi-task Learning for Author Profiling with Hierarchical Features. In: Companion Proceedings of the The Web Conference 2018, WWW '18, Seiten 55–56, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee, ISBN 978-1-4503-5640-4. https://doi.org/10.1145/3184558.3186926.

[56] Schler, Jonathan, Moshe Koppel, Shlomo Argamon und James W Pennebaker: Effects of age and gender on blogging. In: AAAI spring symposium: Computational approaches to analyzing weblogs, Band 6, Seiten 199–205, 2006.

[57] Koppel, Moshe und Yaron Winter: Determining if Two Documents are Written by the Same Author. JASIST, 65(1):178–187, 2014. http://dx.doi.org/10.1002/asi.22954.

[58] Neuman, Yair, Dan Assaf, Yochai Cohen und James L. Knoll: Profiling School Shooters: Automatic Text-Based Analysis. Frontiers in Psychiatry, 6:86, 2015, ISSN 1664-0640. https://www.frontiersin.org/article/10.3389/fpsyt.2015.00086.

[59] Payer, M., L. Huang, N. Z. Gong, K. Borgolte und M. Frank: What You Submit Is Who You Are: A Multimodal Approach for Deanonymizing Scientific Publications. IEEE Transactions on Information Forensics and Security, 10(1):200–212, Jan 2015, ISSN 1556-6013.

[60] Nguyen, Dong, Rilana Gravel, Dolf Trieschnigg und Theo Meder: "How Old Do You Think I Am? A Study of Language and Age in Twitter. In: ICWSM, 2013.

[61] Flesch, Rudolph: A new readability yardstick. Journal of applied psychology, 32(3):221, 1948.

[62] Amstad, T: Wie verständlich sind unsere Zeitungen?[How readable are our newspapers?]. Dissertation, Doctoral thesis, Universität Zürich, Switzerland, 1978.

[63] McClure, Glenda M: Readability formulas: Useful or useless? IEEE Transactions on Professional Communication, (1):12–15, 1987.

[64] Hedman, Amy S: Using the SMOG formula to revise a health-related document. American Journal of Health Education, 39(1):61–64, 2008.

[65] Coleman, Meri und Ta Lin Liau: A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60(2):283, 1975.

[66] Dieckmann, W.: Sprache in der Politik. Sprachwissenschaftliche Studienbücher. Winter, 1969. https://books.google.de/books?id= DvVaAAAAIAAJ.

[67] Neal, Tempestt J., Kalaivani Sundararajan und Damon L. Woodard: Exploiting Linguistic Style as a Cognitive Biometric for Continuous Verification. In: 2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018, Seiten 270–276, 2018. https://doi.org/10.1109/ICB2018.2018.00048.

[68] Barbon, Jr, Sylvio, Rodrigo Augusto Igawa und Bruno Bogaz Zarpelão: Authorship Verification Applied to Detection of Compromised Accounts on Online Social Networks. Multimedia Tools Appl., 76(3):3213–3233, Februar 2017, ISSN 1380-7501. https://doi.org/10.1007/s11042-016-3899-8.

[69] Rexha, Andi, Mark Kröll, Hermann Ziak und Roman Kern: Extending Scientific Literature Search by Including the Author's Writing Style. In: Proceedings of the Fifth Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 39th European Conference on Information Retrieval (ECIR 2017), Aberdeen, UK, April 9th, 2017, Seiten 93–100, 2017. http://ceur-ws.org/Vol-1823/paper9.pdf.

[70] Potthast, M., J. Kiesel, K. Reinartz, J. Bevendorff und B. Stein: A Stylometric Inquiry into Hyperpartisan and Fake News. ArXiv e-prints, Februar 2017.

[71] Hirst, Graeme und Vanessa Wei Feng: Changes in Style in Authors with Alzheimer's Disease. 93:357–370, Mai 2012.

[72] Koppel, Moshe und Yaron Winter: Determining if Two Documents are Written by the Same Author. JASIST, 65(1):178–187, 2013. http://dx.doi.org/10.1002/asi.22954.

[73] Hernández, Josué Gerardo Gutiérrez, José Casillas, Paola Ledesma, Gibran Fuentes Pineda und Iván Vladimir Meza Ruíz: Homotopy Based Classification for Author Verification Task: Notebook for PAN at CLEF 2015. In: Working Notes of CLEF 2015 - Conference and Labs

of the Evaluation forum, Toulouse, France, September 8-11, 2015, 2015. http://ceur-ws.org/Vol-1391/74-CR.pdf.

[74] Khonji, Mahmoud und Youssef Iraqi: A Slightly-modified GI-based Authorverifier with Lots of Features (ASGALF). In: Cappellato, Linda et al. [88], Seiten 977–983, http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-KonijEt2014.pdf.

[75] Kocher, Mirco und Jacques Savoy: A Simple and Efficient Algorithm for Authorship Verification. Journal of the Association for Information Science and Technology, 68(1):259–269, 2017, ISSN 2330-1643. http://dx.doi.org/10.1002/asi.23648.

[76] Seidman, Shachar: Authorship Verification Using the Impostors Method Notebook for PAN at CLEF 2013. In: Forner, Pamela et al. [89]. http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-Seidman2013.pdf.

[77] Potha, Nektaria und Efstathios Stamatatos: An Improved Impostors Method for Authorship Verification. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings, Seiten 138–144, 2017. https://doi.org/10.1007/978-3-319-65813-1_14.

[78] Juola, Patrick und Efstathios Stamatatos: Overview of the Author Identification Task at PAN 2013. In: Forner, Pamela et al. [89]. http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-JuolaEt2013.pdf.

[79] Stamatatos, Efstathios, Walter Daelemans, Ben Verhoeven, Benno Stein, Martin Potthast, Patrick Juola, Miguel A. Sánchez-Pérez und Alberto Barrón-Cedeño: Overview of the Author Identification Task at PAN 2014. In: Cappellato, Linda et al. [88], Seiten 877–897. http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-StamatosEt2014.pdf.

[80] Koppel, Moshe und Jonathan Schler: Authorship Verification as a One-Class Classification Problem. In: Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004, 2004. http://doi.acm.org/10.1145/1015330. 1015448.

[81] Stein, Benno, Nedim Lipka und Sven Meyer zu Eissen: Meta Analysis within Authorship Verification. In: 19th International Workshop on Database and Expert Systems Applications (DEXA 2008), 1-5 September 2008, Turin, Italy, Seiten 34–39, 2008. https://doi.org/10.1109/DEXA.2008.20.

[82] Halvani, Oren, Christian Winter und Lukas Graner: On the Usefulness of Compression Models for Authorship Verification. In: Proceedings of the 12th International Conference on Availability, Reliability and Security, ARES '17, Seiten 54:1–54:10, New York, NY, USA, 2017. ACM, ISBN 978-1-4503-5257-4. http://doi.acm.org/10.1145/ 3098954.3104050.

[83] Stamatatos, Efstathios: Authorship Attribution Using Text Distortion. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Seiten 1138–1149. Association for Computational Linguistics, 2017. http://aclweb.org/anthology/E17-1107.

[84] Matzutt, Roman, Jens Hiller, Martin Henze, Jan Henrik Ziegeldorf, Dirk Müllmann, Oliver Hohlfeld und Klaus Wehrle: A Quantitative Analysis of the Impact of Arbitrary Blockchain Content on Bitcoin. In: Financial Cryptography and Data Security, Band 27, 2018.

[85] Münch, Holger: Praktische Nutzung der Vorratsdatenspeicherung. In: Zeitschrift für Rechtspolitik, Band 5. C.H. Beck, 2015.

[86] Frankfurter Allgemeine: Tausende Kinderporno-Fälle mangels Speicherung nicht aufgeklärt. faz.net, 06.02.2018.

[87] Schleswig-Holsteinischer Landtag: Kleine Anfrage des Abgeordneten Dr. Patrick Breyer (PIRATEN) und Antwort der Landesregierung - Minister für Inneres und Bundesangelegenheiten. Drucksache 18/4831, 15.11.2016. http://www.landtag.ltsh.de/infothek/wahl18/drucks/4800/drucksache-18-4831.pdf.

[88] Cappellato, Linda, Nicola Ferro, Martin Halvey und Wessel Kraaij (Herausgeber): Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15–18, 2014, Band 1180 der Reihe CEUR Workshop Proceedings. CEUR-WS.org, 2014. http://ceur-ws.org/Vol-1180.

[89] Forner, Pamela, Roberto Navigli, Dan Tufis und Nicola Ferro (Herausgeber): Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23–26, 2013, Band 1179 der Reihe CEUR Workshop Proceedings. CEURWS.org, 2014. http://ceur-ws.org/Vol-1179.

[90] Halvani, Oren, Lukas Graner und Inna Vogel: Authorship Verification in the Absence of Explicit Features and Thresholds. In: Pasi, Gabriella, Benjamin Piwowarski, Leif Azzopardi und Allan Hanbury (Herausgeber): Advances in Information Retrieval, Seiten 454–465, Cham, 2018. Springer International Publishing, ISBN 978-3-319-76941-7.

[91] Zeiler, Matthew D. und Rob Fergus: Visualizing and Understanding Convolutional Networks. In: Computer Vision, Band 8689, Seiten 818–833, New York, USA, 2013. Springer.

[92] Goodfellow, Ian, Jonathon Shlens und Christian Szegedy: Explaining and Harnessing Adversarial Examples. In: ICLR, 2015.

12 Glossar

Accuracy	Gütemaß für Klassifikationsalgorithmen. Misst den Anteil der korrekt klassifizierten Beispiele an der Gesamtmenge. 37
Bias	Wert, der in einer Berechnung zu Verzerrung führt. 24
Chatroom	Thematisch abgetrennter Bereich eines Chat-Forums. In einem Chat-Forum über PKWs könnten z.B. verschiedene Chatrooms zu diversen PKW-Herstellern existieren. 15
CNN	Abk. für »Convolutional Neural Network«. Spezielle Art eines KNN, das insbesondere für die Verarbeitung statischer Daten wie Bilder geeignet ist. 25, 38, 39, 42, 44
Computer Vision	Teilgebiet der Informatik, welches sich mit der maschinellen Erkennung bzw. Interpretation von Bild- und Videomaterial beschäftigt. 24, 38, 42
FNR	Abk. für »False Negative Rate« (Falsch-Negativ-Rate). Anteil der durch einen binären Klassifizierer fälschlicherweise als negativ klassifizierten positiven Beispiele an der Gesamtmenge aller positiven Beispiele. 23, 30
FPR	Abk. für »False Positive Rate« (Falsch-Positiv-Rate). Anteil der durch einen binären Klassifizierer fälschlicherweise als positiv klassifizierten negativen Beispiele an der Gesamtmenge aller negativen Beispiele. 23, 30, 33, 36, 37, 46
Hautpixel	Pixel (Bildpunkt), das von einem Algorithmus als hautfarben klassifiziert wurde. 30
KNN	Abk. für »Künstliches Neuronales Netz«. Ein den Rechenprinzipien des Gehirns nachempfundenes Konstrukt aus mathematischen Funktionen, welches anhand von Trainingsdaten lernen kann, bestimmte Muster zu erkennen. 24, 25, 54, 55, 63
MAE	Abk. für »Mean Absolute Error« (mittlerer absoluter Fehler). Gütemaß für Regressionsalgorithmen. Gibt die durchschnittliche betragsmäßige Abweichung vom Zielwert an. 42
Neuron	Im Kontext von Deep Learning: Mathematische Funktion, welche gewichtete Eingangswerte entgegennimmt. Wird durch diese ein bestimmter Schwellwert überschritten, erzeugt eine Aktivierungsfunktion einen Ausgangswert, welcher wiederum an ein folgendes Neuron als Eingangswert weitergegeben werden kann. 24

Nickname	Pseudonym, unter welchem sich eine Person in einem Chat-Forum, Online-Spiel oder Messenger zu erkennen gibt. 11, 15, 19
NSFW	Abk. für »Not Safe For Work«. Bezeichnung für Bilder oder andere digitale Medien, welche für das Arbeitsumfeld unangemessene Inhalte (insbesondere Nacktheit oder sexuell anstößige Darstellungen) enthalten. 28, 29, 38
Pädokriminalität	Beschreibt verallgemeinernd die sexuelle Ausbeutung von Kindern (z.B. durch sexuellen Missbrauch, Kinderprostitution oder Kinderpornografie). 10, 51
Regression	Verfahren, um mittels Statistik oder maschinellem Lernen eine Zielfunktion in Abhängigkeit gegebener Merkmale zu ermitteln. 42
TNR	Abk. für »True Negative Rate« (Richtig-Negativ-Rate). Anteil der durch einen binären Klassifizierer korrekt als negativ klassifizierten Beispiele an der Gesamtmenge aller negativen Beispiele. 23, 35
TPR	Abk. für »True Positive Rate« (Richtig-Positiv-Rate). Anteil der durch einen binären Klassifizierer korrekt als positiv klassifizierten Beispiele an der Gesamtmenge aller positiven Beispiele. 23, 33, 35, 37, 46
Trainingsbeispiel	Einzelnes Element einer Trainingsmenge. 22

A Gesetzestexte

§ 176 StGB: Sexueller Missbrauch von Kindern [Stand: 17.01.2015]

(1) Wer sexuelle Handlungen an einer Person unter vierzehn Jahren (Kind) vornimmt oder an sich von dem Kind vornehmen lässt, wird mit einer Freiheitsstrafe von sechs Monaten bis zu zehn Jahren bestraft.

(2) Ebenso wird bestraft, wer ein Kind dazu bestimmt, daß es sexuelle Handlungen an einem Dritten vornimmt oder von einem Dritten an sich vornehmen lässt.

(3) In besonders schweren Fällen ist auf Freiheitsstrafe nicht unter einem Jahr zu erkennen.

(4) Mit Freiheitsstrafe von drei Monaten bis zu fünf Jahren wird bestraft, wer

1. sexuelle Handlungen vor einem Kind vornimmt,

2. ein Kind dazu bestimmt, dass es sexuelle Handlungen vornimmt, soweit die Tat nicht nach Absatz 1 oder Absatz 2 mit Strafe bedroht ist,

3. auf ein Kind mittels Schriften (§ 11 Absatz 3) oder mittels Informations- oder Kommunikationstechnologie einwirkt, um

   a) das Kind zu sexuellen Handlungen zu bringen, die es an oder vor dem Täter oder einer dritten Person vornehmen oder von dem Täter oder einer dritten Person an sich vornehmen lassen soll, oder

   b) um eine Tat nach § 184b Absatz 1 Nummer 3 oder nach § 184b Absatz 3 zu begehen, oder

4. auf ein Kind durch Vorzeigen pornografischer Abbildungen oder Darstellungen, durch Abspielen von Tonträgern pornografischen Inhalts, durch Zugänglichmachen pornografischer Inhalte mittels Informations- und Kommunikationstechnologie oder durch entsprechende Reden einwirkt.

(5) Mit Freiheitsstrafe von drei Monaten bis zu fünf Jahren wird bestraft, wer ein Kind für eine Tat nach den Absätzen 1 bis 4 anbietet oder nachzuweisen verspricht oder wer sich mit einem anderen zu einer solchen Tat verabredet.

(6) Der Versuch ist strafbar; dies gilt nicht für Taten nach Absatz 4 Nr. 3 und 4 und Absatz 5.

§ 184c StGB: Verbreitung, Erwerb und Besitz jugendpornografischer Schriften [Stand: 27.01.2015]

(1) Mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe wird bestraft, wer

1. eine jugendpornografische Schrift verbreitet oder der Öffentlichkeit zugänglich macht; jugendpornografisch ist eine pornografische Schrift (§ 11 Absatz 3), wenn sie zum Gegenstand hat:

   a) sexuelle Handlungen von, an oder vor einer vierzehn, aber noch nicht achtzehn Jahre alten Person oder

   b) die Wiedergabe einer ganz oder teilweise unbekleideten vierzehn, aber noch nicht achtzehn Jahre alten Person in unnatürlich geschlechtsbetonter Körperhaltung,

2. es unternimmt, einer anderen Person den Besitz an einer jugendpornografischen Schrift, die ein tatsächliches oder wirklichkeitsnahes Geschehen wiedergibt, zu verschaffen,

3. eine jugendpornografische Schrift, die ein tatsächliches Geschehen wiedergibt, herstellt oder

4. eine jugendpornografische Schrift herstellt, bezieht, liefert, vorrätig hält, anbietet, bewirbt oder es unternimmt, diese Schrift ein- oder auszuführen, um sie oder aus ihr gewonnene Stücke im Sinne der Nummer 1 oder 2 oder des § 184d Absatz 1 Satz 1 zu verwenden oder einer anderen Person eine solche Verwendung zu ermöglichen, soweit die Tat nicht nach Nummer 3 mit Strafe bedroht ist.

(2) Handelt der Täter in den Fällen des Absatzes 1 gewerbsmäßig oder als Mitglied einer Bande, die sich zur fortgesetzten Begehung solcher Taten verbunden hat, und gibt die Schrift in den Fällen des Absatzes 1 Nummer 1, 2 und 4 ein tatsächliches oder wirklichkeitsnahes Geschehen wieder, so ist auf Freiheitsstrafe von drei Monaten bis zu fünf Jahren zu erkennen.

(3) Wer es unternimmt, sich den Besitz an einer jugendpornografischen Schrift, die ein tatsächliches Geschehen wiedergibt, zu verschaffen, oder wer eine solche Schrift besitzt, wird mit Freiheitsstrafe bis zu zwei Jahren oder mit Geldstrafe bestraft.

(4) Absatz 1 Nummer 3, auch in Verbindung mit Absatz 5, und Absatz 3 sind nicht anzuwenden auf Handlungen von Personen in Bezug auf solche jugendpornografischen Schriften, die sie ausschließlich zum persönlichen Gebrauch mit Einwilligung der dargestellten Personen hergestellt haben.

(5) Der Versuch ist strafbar; dies gilt nicht für Taten nach Absatz 1 Nummer 2 und 4 sowie Absatz 3.

(6) § 184b Absatz 5 und 6 gilt entsprechend.

B Weiterführende Informationsquellen

Dieser Anhang listet weiterführende Informationsquellen zu den Themen Sexting, Cybergrooming sowie Kinderpornografie auf.

Online-Beratungsangebote für Kinder, Jugendliche und Eltern

gegen-missbrauch e.V.	http://www.gegen-missbrauch.de
Innocence in Danger e.V.	http://www.innocenceindanger.de
Projekt SCHAU HIN!	https://www.schau-hin.info
Saferinternet.at	https://www.saferinternet.at
Save Me Online	http://www.nina-info.de/save-me-online

C Maschinelles Lernen

Maschinelles Lernen (ML) ist ein Kerngebiet der künstlichen Intelligenz (KI), welches dem Bereich der angewandten Informatik zuzuordnen ist. Vereinfacht gesprochen beschäftigt sich ML mit der Generierung von Wissen aus Daten und lässt sich in die folgenden vier Teilgebiete einteilen: **überwachtes, semiüberwachtes, unüberwachtes** sowie **bestärkendes Lernen.** Die Kategorie »überwachtes Lernen« spielt, hinsichtlich der diskutierten Lösungsstrategien in dieser Studie, eine wichtige Rolle. Sie stellt das populärste Teilgebiet des MLs dar, zu dem seit Jahrzehnten intensiv geforscht wird, und ist Gegenstand einer Vielzahl von Lösungen, die in der Industrie und Wirtschaft branchenunabhängig eingesetzt werden. Zu den wichtigsten Vertretern des überwachten Lernens zählen die **Klassifikation, Regression** und **Anomalieerkennung**. **Deep Learning** hingegen repräsentiert einen Spezialfall des MLs, welcher sich allen vier Teilgebieten zuordnen lässt. Im Folgenden werden die genannten Disziplinen vorgestellt, die eine Anwendung in dieser Studie finden.

C.1 Klassifikation

Klassifikation[1] ist der bedeutendste und populärste Teilbereich des MLs. Hintergrund der (computergestützten) Klassifikation ist es, die elementare Fähigkeit eines Menschen nachzuahmen, unbekannte Objekte oder Subjekte hinsichtlich einer festgelegten Menge von Kategorien (ab hier **Klassen** genannt) zuzuordnen. Die Nachahmung erfolgt i. d. R. automatisch, sodass dadurch die Klassifikation großer Mengen an Daten bewältigt werden kann – eine Aufgabe, die ein Mensch, im Hinblick auf Aufwand und Kosten, nicht alleine bewältigen kann.

Das Prinzip der automatisierten Klassifikation beruht darauf, aus Erfahrungen zu lernen. Umgesetzt wird dies, indem aus bereits **bekannten**[2] Daten ein Modell erlernt wird, das die Beziehung zwischen den Daten und ihre dazugehörigen Klassen beschreibt. Anschließend wird das konstruierte Modell eingesetzt, um ungesehene Daten, für die keine Klassen vorliegen, zu klassifizieren. Ein ML-Algorithmus, der dies bewerkstelligt, wird **Klassifikator** genannt. Die Daten, die ein Klassifikator benötigt, um daraus ein Modell zu erlernen, werden **Trainingsdaten** genannt. Das von einem Klassifikator generierte Modell besteht im Wesentlichen aus einem Regelwerk, welches exakt beschreibt, wie die unbekannten Daten auf Basis der darin befindlichen Merkmale zu klassifizieren sind. Je nach eingesetztem Klassifikationsverfahren variieren die erzeugten Modelle bezüglich ihrer Form. So kann ein Modell etwa einen Schwellwert, eine Wahrscheinlichkeitsverteilung, eine Menge von Entscheidungsbäumen oder eine Hyperebene in einem Merkmalsraum darstellen. Der gemeinsame Nenner von ML-Modellen ist es, **Entscheidungsgrenzen** zu definieren, die festlegen, ab wann ein unbekanntes Datum einer vordefinierten Klasse zugeordnet wird.

1 In den Medien werden oft beide Begriffe **ML** und **Klassifikation** fälschlicherweise synonym verwendet. Klassifikation ist ein Teilgebiet von ML.

2 Im Fachjargon wird hier von **gelabelten** Daten gesprochen, also solche Daten, die bereits eine Klassenzuordnung besitzen.

C.1.1 Klassifikationsprobleme

In der Praxis existiert eine Vielzahl von Szenarien, die sich mithilfe von Klassifikationsverfahren lösen lassen. Allerdings muss zuerst eingegrenzt werden, um welches Klassifikationsproblem es sich konkret handelt. Die Identifikation des Klassifikationsproblems richtet sich in erster Linie danach, wie viele Klassen in einem gegebenen Szenario vorliegen. In ML werden primär drei Klassifikationsprobleme unterschieden, die nachfolgend beispielhaft vorgestellt werden.

Unäre Klassifikationsprobleme: Unäre Klassifikationsprobleme haben zahlreiche Anwendungsszenarien, beispielsweise in der Biometrie. Ein Szenario aus dieser Domäne wäre etwa einen legitimen Benutzer B eines Geräts zu erkennen, beispielsweise durch einen Iris-Erkenner, der zum Entsperren eines Smartphones dient. Hier müssen Trainingsdaten von B gesammelt werden (im Beispiel also Bilder der Augen von B), aus denen Merkmale abgeleitet werden, die B beschreiben. Nachdem geeignete Merkmale bestimmt wurden, muss ein Modell konstruiert werden, um B zu einem beliebigen Zeitpunkt[3] zu erkennen und somit das Smartphone zu entsperren. Die Besonderheit unärer Klassifikationsprobleme ist, dass hier ausschließlich Trainingsdaten einer Klasse c_1 existieren (im Beispiel die Augenbilder von B). Die Aufgabe besteht darin, unbekannte Daten, die c_1 angehören, als solche zu erkennen und dagegen alle anderen Daten, die c_1 nicht angehören, zurückzuweisen (siehe Abbildung C.1). Um unäre Klassifikationsprobleme zu identifizieren, kann folgende Faustregel betrachtet werden: Gilt es eine Klasse zu **erkennen** oder **abzuweisen**, so handelt es sich höchstwahrscheinlich um ein unäres Klassifikationsproblem. Zu den bekanntesten Verfahren, mit deren Hilfe unäre Klassifikationsprobleme gelöst werden können, zählen z.B. »One-Class Support Vector Machine«, »Local Outlier Factor« oder auch »AutoEncoder«, die eine spezielle Form von neuronalen Netzen darstellen. Mehr über Neuronale Netze wird im Kapitel C.2 erläutert.

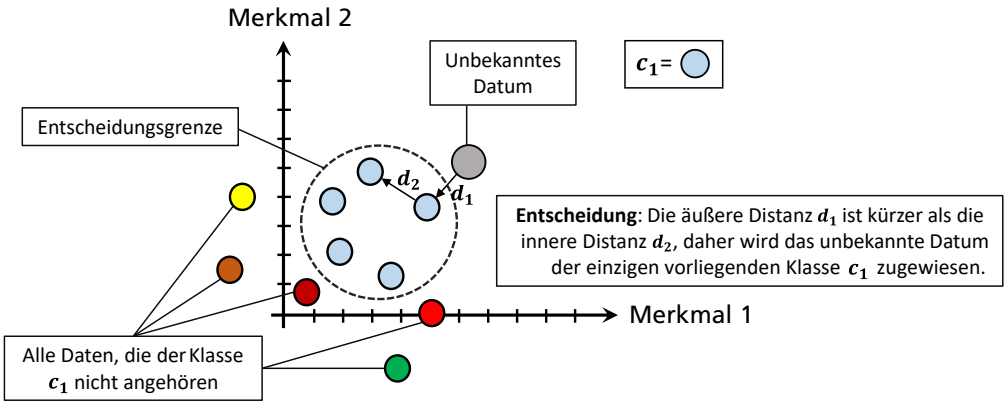


Abbildung C.1: Ein beispielhaftes unäres Klassifikationsproblem, adaptiert aus [90].

Binäre Klassifikationsprobleme: Binäre Klassifikationsprobleme liegen dann vor, wenn zu einem unbekannten Datum eine Auswahl zwischen zwei Klassen c_1 und c_2 besteht. Ein binäres Klassifikationsszenario wäre etwa zu unterscheiden, ob ein Foto F zu der Klasse c_1 = »Erotik« oder c_2 = »Pornografie« gehört. Die Entscheidungsgrenze, ob das Datum zu c_1 oder c_2 zugewiesen wird, ist

3 Soll das Smartphone zur nächtlicher Stunde entsperrt werden, so sollten die Trainingsdaten auch solche Fotos enthalten, die unter entsprechenden Lichtbedingungen aufgenommen wurden.

abhängig davon, ob Merkmale der einen oder anderen Klasse überwiegen (siehe Abbildung C.2). In dem Beispiel von F wäre c_1 die korrekte Wahl, sofern F erotische jedoch nicht pornografische Darstellungen enthalten würde.

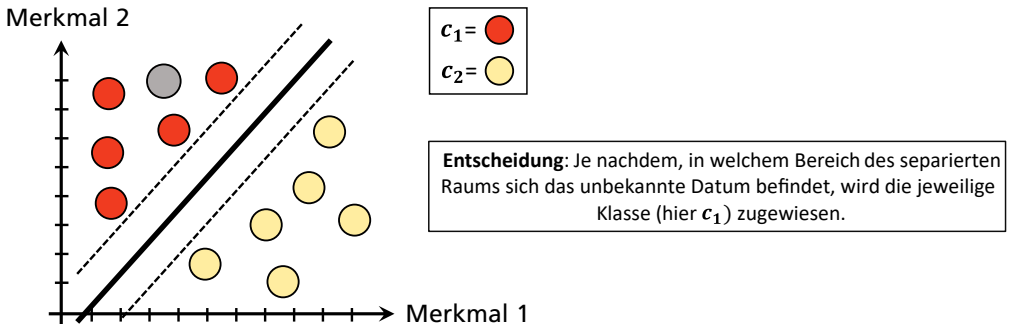


Abbildung C.2: Ein beispielhaftes binäres Klassifikationsproblem.

Manche unäre Klassifikationsprobleme können als binäre Probleme aufgefasst und modelliert werden. So wird in der Praxis z.B. Spam-Erkennung i.d.R. als binäres Klassifikationsproblem betrachtet. Binäre Klassifikationsprobleme können mit einer Vielzahl von ML-Verfahren gelöst werden, z.B. »Support Vector Machines«, »Logistic Regression« oder auch »Perzeptron«. Letzterer repräsentiert im Wesentlichen ein einzelnes Neuron, das Daten linear separieren kann (siehe Abbildung C.4).

N-äre Klassifikationsprobleme: N-äre Klassifikationsprobleme haben in der Praxis eine breite Anwendung und liegen dann vor, wenn n Klassen c_1, c_2, ..., c_n zur Auswahl stehen. Ein Beispiel für ein n-äres Klassifikationsproblem (im Kontext von Cybergrooming) wäre etwa die Altersbestimmung einer Person ausgehend von natürlicher Sprache. So muss hier als Datengrundlage eine Menge von Texten bzw. Textfragmente, wie z. B. Chat-Verläufe vorliegen, deren Klassen Altersstufen repräsentieren. Die Altersstufen können dabei entweder als fortlaufende Jahre (z.B. 27, 28, 30, ...) oder als Jahresintervalle (z.B. [20 - 30], [31 - 40], ...) angegeben werden. Ein anderes Szenario, welches ebenfalls in Sexting Anwendung findet, ist die Klassifikation von pornografischen Bildern hinsichtlich der dargestellten Körperteile. In beiden Szenarien bestimmt der Klassifikator diejenige Klasse, die am ehesten für ein unbekanntes Datum in Frage kommen könnte (siehe Abbildung C.3). Dies geschieht oftmals anhand einer Aggregation (z.B. Mittelwert, Median, Minimum oder Maximum) bezüglich bestimmter Werte wie Gewichte, Wahrscheinlichkeiten, Fehlerraten oder Ähnlichkeitswerte. Gängige Ansätze, die sich für das Lösen von n-ären Klassifikationsproblemen eignen, sind beispielsweise traditionelle ML-Verfahren wie »Random Forests«, »Support Vector Machines«, »Naive Bayes« oder »k Nearest Neighbours«, aber auch modernere Verfahren wie etwa mehrschichtige »Feed-Forward-Netze«, die in der Praxis vielversprechende Ergebnisse liefern können.

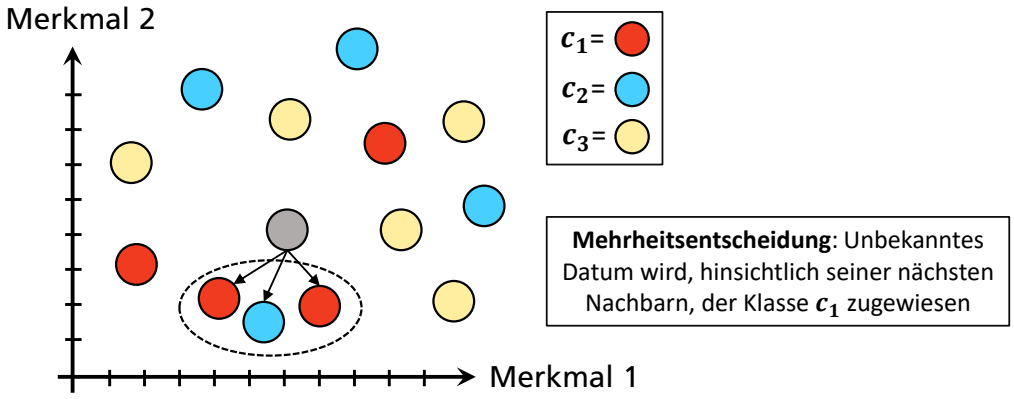


Abbildung C.3:
Ein beispielhaftes n-äres Klassifikationsproblem.

C.1.2 Regression

Regression ist ein weiterer wichtiger Bereich innerhalb von ML und ist sehr eng mit der Klassifikation verwandt. So lassen sich viele Klassifikationsverfahren leicht modifizieren, um eine Regression zu ermöglichen. Umgekehrt lassen sich wiederum Regressionsverfahren mit einer kleinen Anpassung in Klassifikationsverfahren umwandeln. Es bleibt also zu klären, was den Unterschied beider Disziplinen ausmacht. Tatsächlich liegt der Unterschied im Detail: Klassifikationsalgorithmen liefern stets als Resultat, hinsichtlich unbekannter Daten, eine Einteilung in diskrete Klassen (z.B. Nationalität ⟶ {deutsch, englisch, deutsch, ...}). Die Ausgabe von Regressionsverfahren sind dagegen kontinuierlich (z.B. Alter ⟶ {28.5, 29.2, 30.1, ...}). Bildlich gesprochen kann Regression auch wie folgt verstanden werden: Es wird mithilfe eines Modells versucht, Daten bestmöglich zu beschreiben (etwa indem z. B. eine Gerade oder Kurve durch eine Punktwolke gelegt wird), während in der Klassifikation dagegen versucht wird, die Punktwolke, die die Daten beschreibt, in eine oder mehrere Klassen einzuteilen.

Im Rahmen dieser Studie wird ein Regressionsverfahren basierend auf einem neuronalen Netzwerk vorgestellt, mit dessen Hilfe das ungefähre Alter eines Cyber-Groomers vorhergesagt werden kann.

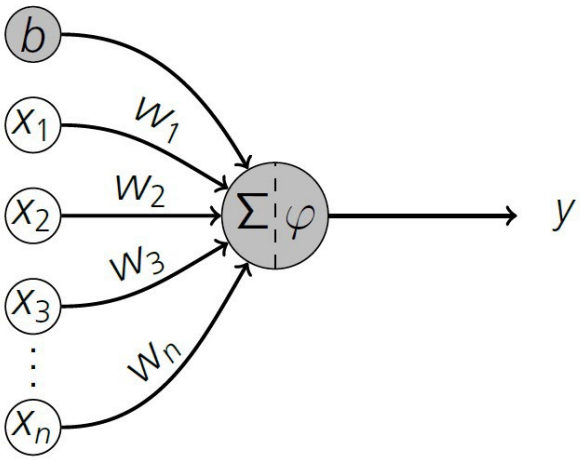


Abbildung C.4:
Schematische Darstellung eines künstlichen Neurons. Die Eingangswerte x_1 *bis* x_n *werden jeweils mit den Gewichten* w_1 *bis* w_n *multipliziert. Die Produkte* $x_i \cdot w_i$ *sowie der Bias* b *werden aufsummiert und die Aktivierungsfunktion* φ *bildet die Summe der Eingangswerte schließlich auf einen Ausgangswert* y *ab.*

C.2 Deep Learning (tiefe neuronale Netze)

Im Bereich des Maschinellen Sehens (Computer Vision) haben sich in jüngster Vergangenheit künstliche neuronale Netze (kurz **NN** genannt) als äußerst erfolgreich erwiesen. Diese bestehen, ähnlich zu den Neuronenverbindungen im Gehirn, aus einer Vielzahl miteinander verknüpfter Recheneinheiten, welche bei bestimmten Eingangssignalen aktiviert werden und das Signal an die nächste Schicht weiterleiten. Einzelne Recheneinheiten werden als künstliche Neurone bezeichnet. Der Aufbau eines Neurons ist schematisch in Abbildung C.4 dargestellt. Jedes künstliche Neuron erhält eine Anzahl von Eingangssignalen x_i, die mit entsprechenden Gewichten w_i multipliziert werden. Je nachdem, wo sich das Neuron innerhalb des neuronalen Netzes befindet, können die Eingangssignale dabei entweder von Vorgängerneuronen oder von der Eingabeschicht (engl. »input layer«) stammen. Der Eingangswert b (engl. »bias«) dient als zusätzlicher Korrekturwert. Überschreitet die Summe der Produkte $x_i \cdot w_i$ und b einen bestimmten Schwellwert, so gibt das Neuron über eine Aktivierungsfunktion φ ein Ausgangssignal y aus. Dieses dient, je nachdem wo sich das Neuron innerhalb des NNs befindet, entweder als Eingangssignal seiner Nachfolgerneuronen oder als Ausgabe des NNs. Ist Letzteres der Fall, so wird das Ausgangssignal y, welches eine reellwertige Zahl darstellt, i. d. R. diskretisiert, um eine Klassenentscheidung abzubilden, oder so belassen, wie es ist, um eine Regression zu ermöglichen. Sowohl die Gewichte w_i, als auch der Bias b sind trainierbare (also veränderliche) Parameter eines Neurons, die während eines Lernprozesses mit Trainingsdaten angepasst werden. Diese Anpassung der Parameter führt letztlich dazu, dass das NN die Erkennung bestimmter Muster »erlernt«.

Innerhalb eines NNs sind die Neuronen in aufeinanderfolgenden Schichten angeordnet. So ist in jedem Fall eine Eingabeschicht zur Entgegennahme von Daten (z. B. Bilder) in das Netz vorhanden sowie eine Ausgabeschicht, die die Eingabedaten einer oder mehreren Klassen zuordnet. Zwischen Eingabe- und Ausgabeschicht befinden sich sogenannte versteckte Schichten (engl. »hidden layer«), die für die Mustererkennung verantwortlich sind. Abbildung C.5 veranschaulicht schematisch den Aufbau eines einfachen NNs mit einer versteckten Schicht.

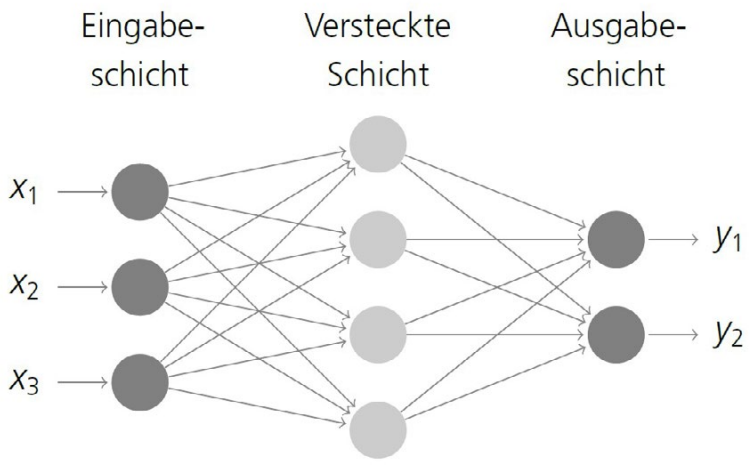


Abbildung C.5:
Aufbau eines einfachen NNs mit nur einer versteckten Schicht. Die einzelnen Kreise symbolisieren Neuronen, während die Pfeile die Verknüpfungen der Neuronen untereinander angeben. Die Eingabeschicht nimmt in diesem Beispiel drei Merkmale (im Fachjargon **Features** *genannt) x_1, x_2 und x_3 entgegen und gibt über die Ausgabeschicht zwei Werte zurück. Diese können durch einen sogenannten 1-aus-n-Code oder auch die* **One-Hot-Kodierung** *($y_1 = 1$, $y_2 = 0$) und ($y_1 = 0$, $y_2 = 1$) zwei verschiedene Klassen (z. B. »Hund« und »Katze«) repräsentieren. Die Bias-Eingänge der einzelnen Neuronen sind hier zwecks Überschaubarkeit nicht dargestellt.*

Enthält ein NN eine größere[4] Anzahl versteckter Schichten, so wird diese oft als tiefes NN (engl. »deep neural network«) bezeichnet. Dieser Begriff wird heute oftmals synonym zum Begriff **Deep Learning** verwendet. Eine Vielzahl moderner NNs gehört dieser Kategorie an, so auch die sogenannten faltenden NNs »Convolutional Neural Networks« (CNNs), welche seit wenigen Jahren als Industriestandard im Bereich des Maschinellen Sehens gelten. Hierbei handelt es sich um eine besondere Form von NNs, die zur Verarbeitung statischer Daten wie z. B. Bildern[5] geeignet sind. So werden zusammenhängende Pixelbereiche eines Bildes auf das Vorhandensein gelernter Merkmale hin untersucht, was zu guten Ergebnissen bei der Erkennung von Bildinhalten führt. Das Lernen sowie die Erkennung von Objekten in Bildern erfolgt bei einem CNN hierarchisch. Das bedeutet, je tiefer eine Schicht innerhalb des Netzes liegt, desto komplexer sind die Bildmerkmale, die sie erkennen kann. So werden z. B. in den ersten Schichten verschiedene Arten von Kanten, Farben u.ä. erkannt. Tiefer gelegene Schichten verknüpfen benachbarte Kanten zu komplexeren Strukturen und sind dadurch beispielsweise in der Lage, geometrische Figuren zu erkennen. In den tiefsten Schichten eines NNs werden schließlich ganze Objekte erfasst und die Ausgangsschicht nimmt die letztendliche Klassifikation des Bildes vor. Abbildung C.6 verdeutlicht, wie unterschiedliche CNN-Schichten Merkmale mit zunehmender Komplexität erkennen können [91].

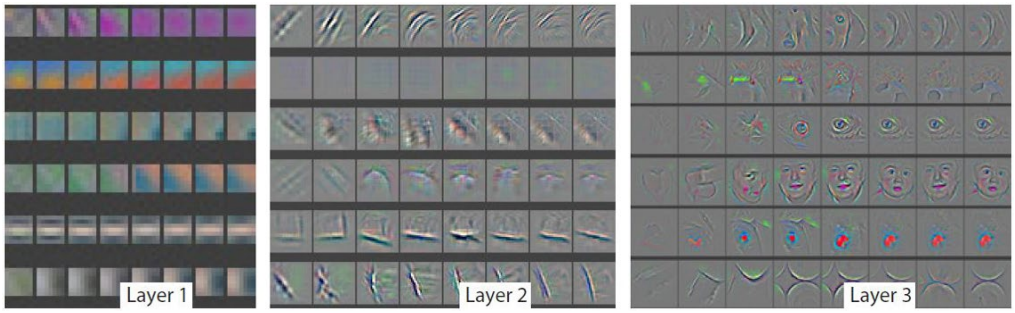


Abbildung C.6: Bildmerkmale, die durch ein CNN in aufeinanderfolgenden Schichten gelernt wurden. Während Layer 1 lediglich auf bestimmte Kanten und Farben anspricht, sind die von Layer 2 erkannten Merkmale bereits deutlich komplexere geometrische Gebilde. In Layer 3 sind bereits Gesichter oder Gesichtsbereiche zu erkennen. Quelle: [91].

C.3 Gütemaße in Machine Learning

Um die Güte von ML-Modellen bewerten zu können, existieren zahlreiche Evaluierungsmöglichkeiten, wobei sich die Wahl aus verschiedenen Faktoren zusammensetzt. So muss zunächst festgehalten werden, ob ein Klassifikations-, Regressions- oder ein anderweitiges Verfahren evaluiert werden soll. Weiterhin muss ein adäquates Verfahren bestimmt werden, um die Prognosefähigkeit des zu trainierenden Modells bestimmen zu können. Außerdem muss in Abhängigkeit der Klassenverteilung bezüglich der Datengrundlage ein geeignetes (bzw. robustes) Evaluierungsmaß ausgewählt werden.

Im Folgenden beziehen wir uns zunächst auf die Evaluierung von Klassifikations- und anschließend auf Regressionsverfahren.

4 In der Paxis existieren NNs mit mehreren Tausenden versteckten Schichten.

5 Jenseits von Bildern haben sich CNNs auch für Texte als nützlich erwiesen, z. B. im Bereich der Stimmungsanalyse (engl. »sentiment analysis«).

C.3.1 Evaluierungsmaße für Klassifikationsverfahren

Grundlage für die Evaluierung der meisten unären, binären und n-ären Klassifikationsverfahren stellen vier mögliche Fälle dar:

- **True Positives (TP):** Tatsächlich positive Beispiele, die als positiv klassifiziert wurden.
- **True Negatives (TN):** Tatsächlich negative Beispiele, die als negativ klassifiziert wurden.
- **False Positives (FP):** Tatsächlich negative Beispiele, die als positiv klassifiziert wurden.
- **False Negatives (FN):** Tatsächlich positive Beispiele, die als negativ klassifiziert wurden.

Die vier Fälle lassen sich in einer sogenannten Konfusionsmatrix (siehe Abbildung C.7) anordnen und werden daher auch als »Ausprägungen« dieser Matrix bezeichnet. Basierend auf diese Ausprägungen lassen sich eine Reihe von Kennzahlen berechnen, wobei wir uns allerdings nur auf jene (siehe Tabelle C.1) fokussieren, die im Rahmen dieser Studie Anwendung gefunden haben.

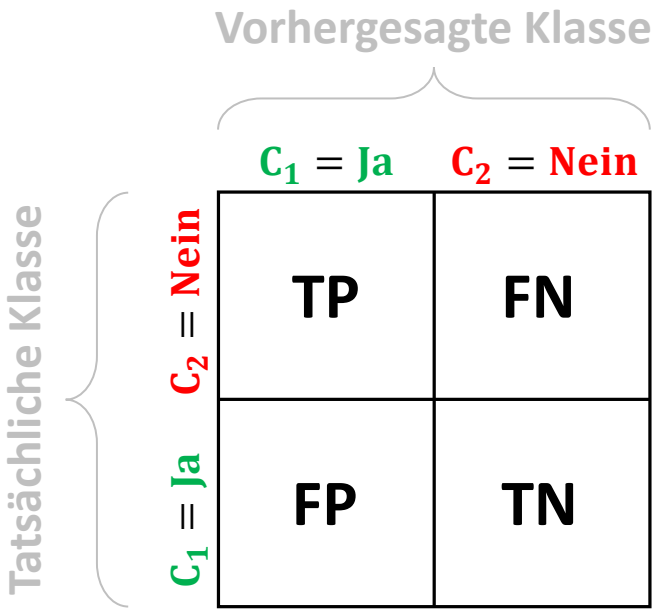


Abbildung C.7:
Eine Konfusionsmatrix, die zur Beurteilung der Güte von unären oder binären Klassifikationsverfahren verwendet werden kann. C_1 und C_2 bezeichnen hierbei die zwei Klassen.

Kennzahl	Formel
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R)	$\frac{TP}{TP + FN}$
F_1	$\frac{2PR}{P+R}$
True Positive Rate (TPR)	$\frac{TP}{TP + FP}$
False Positive Rate (FPR)	$\frac{FP}{FP + TN}$
Area Under the Curve (AUC)	Fläche unterhalb der ROC-Kurve, die aus einer Menge von (FPR, TPR)-Paaren gebildet ist.

Tabelle C.1: Evaluierungsmaße für ML-Modelle, die im Rahmen dieser Studie verwendet werden.

C.3.2 Evaluierungsmaße für Regressionsverfahren

Um die Güte von Verfahren zu bewerten, die auf Regression basieren, sind andere Evaluierungsmaße nötig. Hintergrund ist, dass bei der Regression keine diskreten Ausgaben wie z. B. »wahr« oder »falsch« vorliegen, sondern stattdessen kontinuierliche Werte, sodass Maße benötigt werden, die damit umgehen können. Zu den zwei bekanntesten Performanzmaßen für regressionsbasierte Verfahren, die auch im Rahmen dieser Studie verwendet werden, zählen **MSE** (engl. Mean Squared Error) und **RMSE** (engl. Root Mean Squared Error). Erstere ist definiert als:

$$\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2$$

Hier bezeichnet n die Zahl der existierenden Trainingsdaten, f_i den vorhergesagten und y_i den tatsächlichen Wert des i-ten Datenpunkts. Der RMSE hingegen repräsentiert die Wurzel aus der MSE-Formel.

C.3.3 Überanpassung / Unteranpassung

Neben der Performanz eines ML-Verfahrens stellt dessen »Generalisierungsfähigkeit« mit ungesehenen Daten eine weitere zentrale Anforderung dar. Eine optimale Generalisierungsfähigkeit besteht, wenn ein ML-Verfahren hinsichtlich der Daten weder unter- noch überangepasst ist.

Zur Verdeutlichung des Problems betrachten wir ein binäres Klassifikationsproblem, bei dem ein ML-Verfahren Hunde- von Katzenbildern unterscheiden soll. Um dies zu realisieren, muss ein Klassifikationsmodell basierend auf entsprechenden Trainingsdaten konstruiert werden. Eine **Überanpassung** (engl. »overfitting«) würde nun dann vorliegen, wenn sich das ML-Verfahren zu viele

Details (die nicht allesamt relevant sind) in den Trainingsbildern merken würde. Dazu zählen z. B. Anzahl der Schnurrhaare, Verschmutzungen auf dem Fell, Narben oder aber auch der Hintergrund in den Bildern (die heimische Wohnung bei Katzen oder ein Park/Wald bei Hunden). Ein Klassifikator, der sich auf viele solcher Details fokussiert, um zu seiner Entscheidung zu kommen, wird es schwer haben, ungesehene Daten zu klassifizieren, in denen jene Details nicht vorkommen. In diesem Fall wird ein solcher Klassifikator als überangepasst bezeichnet. Eine **Unteranpassung** (engl. »underfitting«) hingegen liegt vor, wenn sich der Klassifikator viel zu wenige (oder irrelevante) Merkmale merkt, mit denen Hunde von Katzen unterschieden werden können. Dazu zählen z. B. die Fellfarbe oder die Anzahl der Krallen, Ohren oder Beine.

Liegt eine Über- oder Unteranpassung bei einem Klassifikator vor, so hat dies i. d. R. einen direkten (negativen) Einfluss auf dessen Performanz bzgl. ungesehener Daten. Ein praxistauglicher Klassifikator benötigt daher eine vernünftige Auswahl an Merkmalen und muss zudem tolerant genug sein, wenn sich diese nur unvollständig jenseits der Trainingsdaten wiederfinden. Während sich eine Unteranpassung relativ einfach beheben lässt (einfach mehr bzw. bessere Merkmale betrachten), verlangt die Vermeidung oder Minderung von Überanpassung andere Strategien, welche nachfolgend kurz erläutert werden.

Regularisierung: Es existieren unterschiedliche Regularisierungsmechanismen, die dabei helfen können, den Grad der Überanpassung zu senken. Eine populäre Methode, die häufig beim Trainieren von neuronalen Netzen Anwendung findet, ist z. B. *Dropout*. Hierbei werden Aktivierungen von zufällig ausgewählten Neuronen innerhalb der versteckten Schichten eines NNs auf Null gesetzt und somit »ausgeschaltet«. Dadurch wird das NN gezwungen, sich nicht alles zu merken, sondern sich nur auf bestimmte Konstellationen von (idealerweise stark diskriminierenden) Merkmalen zu fokussieren. Eine andere Art von Regularisierung in NNs ist *early stopping*. Hier wird der Lernprozess frühzeitig gestoppt, wenn keine (größere) Verbesserung des Klassifikationsergebnisses festgestellt werden kann.

Datenaufteilung: Eine andere Möglichkeit, die Generalisierungsfähigkeit von ML-Verfahren zu kontrollieren, ist die Aufteilung der Datenmenge in verschiedene Partitionen, wobei jede einen anderen Zweck erfüllt. Eine gängige Praxis ist es, die Daten in eine Trainings-, Validierungs- und Testmenge aufzuteilen. Erstere dient dabei dazu, das ML-Verfahren hinsichtlich **Modellparameter** zu trainieren (z. B. die Gewichte in einem NN). Eine Validierungsmenge hingegen ist nützlich, um sogenannte **Hyperparameter** (also solche Parameter, die der Benutzer von außen einstellt, wie z. B. die Anzahl der versteckten Schichten in einem NN) zu optimieren. Die Testmenge dient letztendlich ausschließlich dazu, das ML-Modell (bestehend aus Modell- und Hyperparametern) auf ungesehenen Daten zu evaluieren. Wenn nun ein ML-Verfahren auf den Trainingsdaten sehr gute Ergebnisse erzielt, dafür aber auf der Testmenge schlecht abschneidet, so liegt mit großer Wahrscheinlichkeit eine Überanpassung vor. Erzielt das ML-Verfahren jedoch bereits auf der Trainingsmenge schlechte Ergebnisse, so liegt wiederum eine Unteranpassung vor.

Für den Fall, dass der Umfang der Datenmenge klein ist, kann eine weitere Technik verwendet werden, die als k-fache[6] Kreuzvalidierung (engl. *k-fold cross validation*) bezeichnet wird. Dabei wird die Datenmenge in k gleich große Teilmengen $t_1, t_2, ..., t_k$ partitioniert, wobei eine Teilmenge t_i als Testmenge und die verbliebenen $t_{i+1}, t_{i+2}, ...$ als Trainingsmenge fungiert. Im nächsten Schritt repräsentiert t_{i+1} die Testmenge und die verbliebenen die Trainingsmenge. Dieser Vorgang wird k-mal wiederholt und die Ergebnisse hinsichtlich der k Aufteilungen gemittelt. Anhand der Klassifikationsfehler, die in jeder Aufteilung entstehen, lässt sich erkennen, ob der Klassifikator zu einer Überanpassung neigt.

C.4 Sicherheitsrisiken

Wie alle IT-Systeme sind auch ML-Algorithmen angreifbar und weisen gewisse Schwachstellen auf. Um Klassifikationsmodelle anzugreifen oder zu manipulieren, gibt es verschiedene Herangehensweisen, die in diesem Abschnitt diskutiert werden.

Manipulation von Modellen: Ist nach der Trainingsphase der Lernprozess eines ML-Algorithmus, z.B. eines NNs, abgeschlossen, könnte ein Angreifer, der Zugang zu dem gespeicherten Modell erlangt, dieses durch Veränderung der gelernten Parameter manipulieren. Das Modell würde in der Folge keine zuverlässige Klassifikation mehr durchführen und wäre somit unbrauchbar. Ein solcher Angriff ließe sich jedoch sehr schnell bemerken. Dass ein Angreifer das gelernte Modell eines NN jedoch gezielt verändert, um bestimmte Fehlklassifikationen zu erreichen, ist nach derzeitigem Stand der Wissenschaft sehr unwahrscheinlich, da komplexe NN-Modelle über mehrere Millionen Parameter verfügen, was sie schwer interpretierbar macht. Das Verständnis der einzelnen Modellparameter ist jedoch notwendig, um ein Modell gezielt zu manipulieren. Einfachere ML-Algorithmen wie z.B. solche, die auf Entscheidungsbäumen basieren, können dagegen durchaus analysiert und gezielt verändert werden.

Manipulation von Trainings- und Testdaten: Ist ein Angreifer in der Lage, unbemerkt die Trainings- und Testdaten für das zu lernende Klassifikationsmodell zu manipulieren, kann im schlimmsten Fall das gesamte Modell unbrauchbar gemacht werden. Beispielsweise durch das unbemerkte Hinzufügen von bestimmten Rauschmustern, die durch bloßes Betrachten nicht zu erkennen sind, kann das Modell so angelernt werden, dass spätere unverfälschte Daten im Produktiveinsatz des Modells nicht korrekt erkannt werden. Ein solcher Angriff lässt sich jedoch leicht durch zusätzliche Tests erkennen, woraufhin ein neues Modell mit unverfälschten Daten generiert werden kann. Ersetzt der Angreifer gar die gesamten Daten durch eigene oder manipuliert er die Labels der einzelnen Beispiele, kann er bewirken, dass ein Modell gelernt wird, welches seinen Zielen entspricht.

Manipulation von Produktivdaten: Anstatt das Modell selbst bzw. die ihm zugrunde liegenden Daten zu manipulieren, wie es in den beiden erstgenannten Szenarien der Fall ist, kann ein Angreifer sich auch lediglich auf die Produktivdaten beschränken, also die Daten, die das gelernte

6 Gebräuchlich sind Werte k = 5 oder k = 10

Modell in einem konkreten Anwendungsszenario klassifizieren soll. NNs sind aufgrund ihrer hohen Komplexität nur schwer interpretierbar, weshalb sie häufig als Blackboxen behandelt werden, die zwar gute Klassifikationsergebnisse liefern, wobei jedoch oft nicht hinterfragt wird, wie diese zustande kommen. Es existieren mittlerweile eigene Forschungszweige, die sich mit der Frage der Interpretierbarkeit neuronaler Netze sowie deren Täuschung durch leicht veränderte Eingabedaten auseinandersetzen. Die in der entsprechenden Fachliteratur als »advesarial examples« (zu Deutsch gegnerische Beispiele) bezeichneten manipulierten Daten unterscheiden sich von den Originaldaten meist nur durch von Menschen kaum oder nicht wahrnehmbare Veränderungen [92]. Abbildung C.8 zeigt, wie durch das Hinzufügen eines Rauschmusters zu einem Bild die Klassifikation durch ein NN bewusst korrumpiert werden kann.

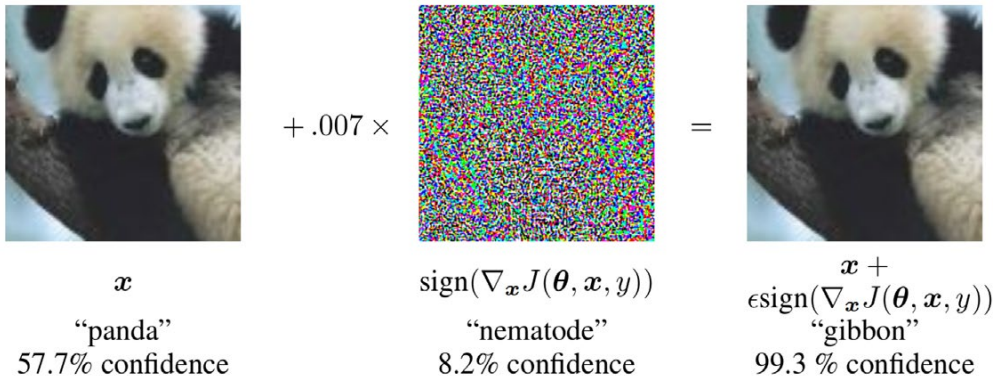


Abbildung C.8:
Beispiel für die Korrumpierung einer Bildklassifikation durch Hinzufügen eines Rauschmusters zu einem Eingabebild. Anstelle der durch das Neuronale Netz zu erwartenden Klassifikation als Panda, wird die Klasse Gibbon zugeordnet [92].