HOCHSCHULE
MITTWEIDA
University of Applied Sciences

BACHELORARBEIT

Frau
Maria Starke

Altersbestimmung zur Charakterisierung von Blog-Autoren im Kontext der digitalen forensischen Kommunikationsanalyse

Mittweida, Mai 2024

Fakultät **Angewandte Computer- und Biowissenschaften**

BACHELORARBEIT

Altersbestimmung zur Charakterisierung von Blog-Autoren im Kontext der digitalen forensischen Kommunikationsanalyse

Autorin:
Maria Starke

Studiengang:
Allgemeine und Digitale Forensik

Seminargruppe:
FO20w5-B

Erstprüfer:
Prof. Dr. rer. nat. Michael Spranger

Zweitprüferin:
Jenny Felser, M.Sc.

Einreichung:
Mittweida, 03.05.2024

Verteidigung/Bewertung:
Mittweida, 2024

Faculty of **Applied Computer Sciences and Biosciences**

BACHELOR THESIS

Age determination for the characterization of blog authors in the context of digital forensic communication analysis

Author:
Maria Starke

Course of Study:
General and Digital Forensic Science

Seminar Group:
FO20w5-B

First Examiner:
Prof. Dr. rer. nat. Michael Spranger

Second Examiner:
Jenny Felser, M.Sc.

Submission:
Mittweida, 03.05.2024

Defense/Evaluation:
Mittweida, 2024

Bibliografische Beschreibung

Starke, Maria:
Altersbestimmung zur Charakterisierung von Blog-Autoren im Kontext der digitalen forensischen Kommunikationsanalyse. – 2024. – 44 S.
Mittweida, Hochschule Mittweida – University of Applied Sciences, Fakultät Angewandte Computer- und Biowissenschaften, Bachelorarbeit, 2024.

Referat

Die stetig ansteigende Internetnutzung hat zur Folge, dass auch ein großer Teil der Kommunikation digital passiert. Durch die Möglichkeit der Anonymität ist es gerade für Strafverfolgungsbehörden interessant relevante Texte ihrem Urheber zuordnen zu können oder mithilfe von ihnen Rückschlüsse auf den Autor zu ziehen. So beschäftigt sich diese Arbeit mit der Altersbestimmung von Menschen anhand ihrer verfassten Texte mithilfe von Ansätzen des maschinellen Lernens. Nach dem Training und Test der Modelle auf Blogbeiträgen, wurden sie weiterhin auf strafrechtlich relevanten Daten getestet, mit dem Ziel Cybergrooming-Fälle durch Altersbestimmung der Chatpartner festzustellen. Dafür wurden n-Gramme, Second Order Attributes und statistische Features getestet. Die Ergebnisse zeigen, dass die Alterserkennung eine Herausforderung darstellt, ihr Einsatz aber zur Erhöhung der Sicherheit im digitalen Bereich beitragen kann und sie dadurch ein wichtiges Forschungsgebiet bildet.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AA . . . . . Authorship Attribution

AP . . . . . Author Profiling

AV . . . . . Authorship Verification

BoW . . . . . Bag-of-Words

POS . . . . . Part-of-Speech

RF . . . . . Random Forest

SOA . . . . . Second Order Attributes

SVM . . . . . Support Vector Machine

1 Einleitung

Das Internet ist zu einem festen Bestandteil im Leben vieler Menschen geworden und für die meisten vermutlich nicht mehr wegzudenken. Das zeigt auch die zunehmende Nutzung über die letzten zwei Jahrzehnte hinweg [1]. Die Möglichkeit der einfachen Kommunikation und weltweiten Vernetzung sind dabei zentrale Vorteile. Einer Umfrage des IfD Allensbach im Jahr 2023 zufolge ist die beliebteste Aktivität der Deutschen im Internet das Verschicken von Textnachrichten, Bildern und Videos über Messengerdienste und zwar unter allen der befragten Generationen [2]. Die Möglichkeit zur Kommunikation auf digitalem Weg kann aber auch für kriminelle Handlungen genutzt werden, was durch die Möglichkeit auf Anonymität noch begünstigt wird.

Im Sinne der Strafverfolgung und Forensik bietet das Internet also eine enorme Menge an Daten, die zur Erkennung oder Aufklärung von Straftaten interessant sein können. Da eine manuelle Überprüfung und Auswertung der Daten mit viel Zeit- und Personalaufwand verbunden ist, wird versucht solche Aufgaben durch computergestützte Systeme effektiver durchzuführen und so Strafverfolgungsbehörden zu unterstützen. Es gilt dabei einerseits die relevanten Daten ausfindig zu machen und andererseits die dahinterstehenden Personen zu identifizieren. Da eine direkte Identifizierung aber bspw. durch Persönlichkeitsverschleierung oder mangelnde Vergleichsdokumente nicht immer möglich ist, kann versucht werden den Personen Merkmale zuzuordnen, die bei der Suche nach der richtigen Identität hilfreich sind, da sie zumindest den Kreis möglicher Personen eingrenzen. Eine Möglichkeit für diese Art der Charakterisierung ist die Ableitung verschiedener Persönlichkeitsaspekte eines Autors von seinen Texten und wird als Author Profiling (AP) bezeichnet [3].

Im Rahmen dieser Arbeit geht es speziell um die Altersbestimmung von Autoren anhand ihrer geschriebenen Texte mithilfe maschinellen Lernens. Es wurden Modelle mit verschiedenen Features wie n-Grammen, Second Order Attributes und statistischen Textmerkmalen trainiert und getestet, um Autoren von Blogbeiträgen einer von drei Altersgruppen zuzuordnen. Da die Unterscheidung zwischen minderjährigen Personen und Erwachsenen einen besonderen Stellenwert in der Strafverfolgung hat, wurden die entwickelten Modelle weiterhin auf eine Klassifikation nach Kind und Erwachsener trainiert und final in einer Cross-Genre-Evaluation auf strafrechtlich relevanten Daten getestet, mit dem Ziel der Detektion von Grooming-Fällen anhand der Alterserkennung von Chatpartnern.

Nach einer Einführung in das Thema Textklassifikation und Altersbestimmung in der Forensik in Kapitel 2, wird in Kapitel 3 ein Überblick über den Stand der Forschung zum Thema des AP in Bezug auf das Alter gegeben. Es folgt eine Darstellung der in dieser Arbeit verwendeten Daten und der Zusammenstellung der Datensätze in Kapitel 4 sowie eine Erläuterung der angewandten Methodiken in Kapitel 5. In Kapitel 6 werden die Ergebnisse vorgestellt, diskutiert und in einen Zusammenhang gebracht. Den Abschluss bildet eine Zusammenfassung mit einem Ausblick in Kapitel 7.

2 Grundlagen

Dieses Kapitel vermittelt die Grundlagen der Textklassifikation, welche Basis dieser Arbeit ist, und diskutiert die Bedeutung der Alterserkennung im Kontext der Kriminalität.

2.1 Textklassifikation

Textklassifikation ist eine gängige Aufgabe des Natural Language Processings, bei der einem Eingabetext ein vordefiniertes Label zugeordnet wird [4]. Abhängig von der Anzahl an verfügbaren Labels handelt es sich um ein binäres (zwei Klassen) oder ein Mehrklassenproblem [5]. Zudem wird danach unterschieden, ob dem Eingabetext genau ein Label zugeordnet wird (z.B. Wurde der Eingabetext von einer KI erstellt oder nicht?) oder aber mehrere Label einem Text zugeordnet werden können (z.B. Mit welchen Themen befasst sich der Eingabetext?). So wird die Zuordnung eines einzigen Labels als „Single-Label-Klassifikation“ bezeichnet, die Zuordnung von mehreren Labels als „Multi-Label-Klassifikation“ [5].

Typische Aufgaben, die als Textklassifizierungsproblem behandelt werden können, sind z.B.:

- Sentimentanalyse: Stimmung eines Textes erkennen [z.B. 6]
- Spam-Erkennung: Kategorisierung von Nachrichten in Spam und Nicht-Spam [z.B. 7]
- Spracherkennung: Texte einer Sprache zuordnen [z.B. 8]

Der grundlegende Ablauf einer Textklassifikation folgt dem nachstehenden Schema:

Abbildung 2.1: Allgemeiner Ablauf einer Textklassifikation angelehnt an das Schema von [9].

Der übergebene Text muss zuerst in eine maschinenlesbare Darstellung überführt werden, was als Feature-Extraktion bezeichnet wird (siehe Abschnitt 2.1.1). Oftmals erfolgt vorher eine Vorverarbeitung der Textdaten zur Bereinigung und Vereinheitlichung. Meist schließt sich ein Schritt zur Feature-Selektion an (siehe Abschnitt 2.1.2), um einerseits die Dimensionalität der Textrepräsentation zu verringern und andererseits irrelevante Features auszusortieren [10, S. 266].

Daraufhin folgt die Klassifikation, bei der das Modell von den Beispielen in den Trainingsdaten lernt. Hierbei können verschiedene Klassifikationsverfahren zum Einsatz kommen. Im Bereich der Textklassifikation finden und fanden bereits viele verschiedene Techniken und Algorithmen Anwendung, sowohl traditionelle als auch Deep Learning Methoden (siehe Abschnitt 2.1.3). Der finale Schritt ist die Evaluation des Modells (siehe Abschnitt 2.1.4) auf

Grundlage der Vorhersagen für ungesehene Testdaten. Dafür gibt es verschiedene Maße, welche zum Vergleich von Modellen und ihrer Leistung berechnet werden können.

2.1.1 Textrepräsentation und Features

Da es sich bei Texten in den meisten Fällen um unstrukturierte Daten handelt, ist es notwendig die Dokumente in eine systematische von Maschinen lesbare Repräsentation zu bringen. Dazu erfolgt oftmals zuerst eine Vorverarbeitung der Texte, in der sie bereinigt / vereinheitlicht werden, teilweise abhängig von der zu bearbeitenden Textklassifizierungsaufgabe und der verwendeten Klassifikationstechnik. Einige Maßnahmen der Vorverarbeitung sind bspw. Tokenisierung des Textes in kleinere Einheiten (z.B. Worte oder Sätze), Umgang mit Stoppworten (z.B. Entfernung), Umgang mit Großbuchstaben (z.B. alle Groß- in Kleinbuchstaben umwandeln), Lemmatisierung (Rückgabe der Wörter in ihrer Grundform), Stemming (Rückgabe der Worte als Wortstamm) oder auch der Umgang mit Abkürzungen (z.B. alle Abkürzungen ausschreiben) [9].

Anschließend folgt die Extraktion von Features, durch welche die Texte repräsentiert werden und die diese charakterisieren. Im einfachsten Falle wird eine Bag-of-Words (BoW)-Repräsentation gewählt, bei der jedes im Vokabular des Korpus vorkommende Wort als ein Feature betrachtet wird und die Anzahl an Vorkommen jeden Wortes pro Dokument den Wert des Features setzt. Darauf kann dann aufgebaut werden mit Termgewichtung, bspw. mittels der TF-IDF-Heuristik [11], welche die Termfrequenz mit der inversen Dokumentenfrequenz kombiniert, wodurch spezifische Wörter höher gewichtet werden als allgemeine Begriffe wie Stoppwörter.

Neben Wort-Unigrammen sind Wort-Bi- und -Trigramme weitere häufig verwendete n-Gramme [9]. Zusätzlich zu Wort-n-Grammen ist auch die Verwendung von Zeichen-n-Grammen üblich. Da verschiedene n-Gramm-Feature-Sets unterschiedliche Informationen enthalten, ist die Verwendung mehrerer solcher Feature-Sets in Kombination ebenfalls eine gängige Methode [12, S. 305].

Zusätzlich zu den lexikalischen Features werden auch andere Merkmale hinzugezogen, die den Stil oder statistische Kennzahlen der Texte wiedergeben. Dazu zählen z.B. Wort-, Satz- und Dokumentlängen, Anzahl bestimmter Zeichen und Zeichenkombinationen, Verwendung von bestimmten Wortarten oder auch das Sentiment eines Textes.

2.1.2 Feature-Selektion

Vor allem bei einer hohen Anzahl an extrahierten Features bietet sich der Schritt der Feature-Selektion an. Da oftmals nur ein Teil der Features hilfreich für die Klassifizierung ist, wird versucht diese Teilmenge ausfindig zu machen [12, S. 304] und irrelevante sowie redundante Features auszusortieren [10, S. 266]. Der Prozess der Feature-Selektion hat aber auch noch

das Ziel dem sogenannten Fluch der Dimensionalität [13] entgegenzuwirken [10, S. 266]. Denn aus einer hohen Feature-Anzahl folgt eine hohe Dimensionalität, die zu Problemen wie geringer Datendichte und Schwierigkeiten bei der Abstandsberechnung führt [10, S. 244].

Bei der Feature-Selektion wird grundsätzlich nach drei Arten unterschieden: Wrapper-Methoden, Filter-Methoden und Embedded-Methoden [14].
Wrapper-Methoden nutzen einen bestimmten Lernalgorithmus, mit dem eine Teilmenge der Features gesucht und diese anschließend evaluiert wird. Diese zwei Schritte werden bis zur Erfüllung bestimmter Abbruchkriterien oder dem Erreichen der angestrebten Lernleistung wiederholt [14].
Filter-Methoden hingegen nutzen keinen Lernalgorithmus, sondern ein Feature-Bewertungskriterium, anhand dessen die Features entsprechend ihrer Bewertung eingestuft werden. Die Features mit der niedrigsten Bewertung werden dann aussortiert. Die Feature-Selektion über Filter-Methoden ist zwar effizienter, es kann jedoch sein, dass die ausgewählten Features nicht optimal sind [14].
Embedded-Methoden vereinen die Vorzüge der beiden anderen Methoden, die Effizienz der Filter- und die Genauigkeit der Wrapper-Methoden, indem sie die Features, die für die Genauigkeit des Modells verantwortlich sind, während der Erstellung des Trainingsmodells kalkulieren [15].

2.1.3 Klassifikationstechniken

Für Textklassifizierungsaufgaben sind bereits eine Vielzahl an Klassifikationstechniken zum Einsatz gekommen. Nachfolgend werden einige gängige Methoden benannt [9]. Zu den eher traditionellen Klassifikationstechniken für die Textklassifizierung zählen Logistische Regression, Naive Bayes und k-nearest Neighbor, welche trotzdem noch immer Anwendung finden. Auch die Support Vector Machine (SVM) ist weit verbreitet in der Textklassifizierung. Entscheidungsbaum-basierte Algorithmen wie Decision Tree und Random Forest (RF) zählen ebenso zu oft genutzten Algorithmen.

Zudem werden Deep Learning-Methoden wie Recurrent Neural Networks und Convolutional Neural Networks verwendet sowie die Transformer-Architektur. Ein sehr populäres Modell, welches auf dieser Architektur aufbaut, ist das BERT-Modell [16] (weitere sind z.B. RoBERTa, DistilBERT). Im Gegensatz zu den traditionellen Klassifizierungsalgorithmen ist für die Neuronalen Netze keine manuelle Feature-Extraktion mehr notwendig, da diese Modelle die Merkmale selbst erkennen. Einerseits bietet dies eine Einsparung bezüglich des Arbeitsaufwandes, andererseits ist nicht nachvollziehbar, welche Features für die Vorhersagen in Betracht gezogen wurden, was für bestimmte Bereiche wie Gesundheit ein Ausschlusskriterium darstellt [9].

Da in dieser Arbeit die beiden Methoden SVM und RF Anwendung fanden, werden diese nachfolgend kurz näher erläutert.
SVMs [17] bilden zur Klassifikation eine Entscheidungsgrenze (genannt Hyperebene) im Feature-Raum, welche die Datenpunkte der unterschiedlichen Klassen voneinander trennt. Diese Hyperebene wird so gewählt, dass der Abstand zwischen ihr und den nächsten Datenpunkten,

die als Support-Vektoren bezeichnet werden, maximiert wird [10].
Der Ensemble-Algorithmus RF (erstmals von Ho [18] als Random Decision Forest vorgestellt) nutzt eine Sammlung von Entscheidungsbäumen, die unabhängig voneinander auf einem zufälligen Teil des Trainingsdatensatzes trainiert werden und auf einer zufälligen Auswahl von Features bei jedem Knoten basieren [10]. Die vorhergesagte Klasse eines Datenpunktes wird dann nach dem Voting-Prinzip ausgewählt [9].

Weitere Informationen zu den genannten Klassifikationstechniken können den in diesem Abschnitt genannten Quellen entnommen werden.

2.1.4 Evaluation

Zur Evaluation einer Textklassifikation gibt es unterschiedliche Maße, viele davon beruhen auf der Konfusionsmatrix. Eine solche für eine binäre Klassifikation, bei der es üblicherweise eine positive und eine negative Klasse gibt, ist in Abbildung 2.2 dargestellt. Diese Konfusionsmatrix umfasst die korrekt klassifizierten True Positives (TP) und True Negatives (TN), sowie die falsch klassifizierten False Positives (FP) und False Negatives (FN). In der Abbildung beziehen sich die Zeilen auf die durch den Algorithmus vorhergesagten Klassen und die Spalten auf die tatsächliche Klasse. Diese Anordnung kann auch umgekehrt sein. Nachfolgend werden häufig genutzte Evaluationsmaße kurz vorgestellt. Außerdem wird zuletzt noch auf das Problem der Cross-Domain-Evaluation eingegangen.

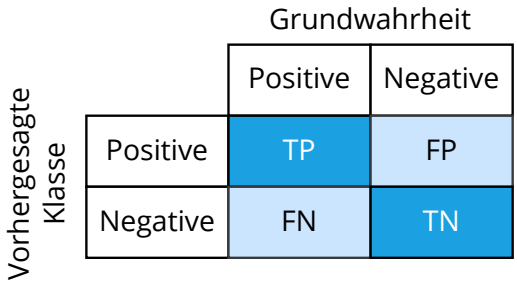

Abbildung 2.2: Konfusionsmatrix für eine binäre Klassifikation

Accuracy

Die Accuracy ist ein sehr gängiges Evaluationsmaß zum Vergleich von Textklassifikationsmodellen, das den Anteil an korrekten Vorhersagen über allen Vorhersagen angibt (siehe Gleichung (2.1)) und neben binären auch für Mehrklassen-Probleme verwendet werden kann. Das Accuracy-Maß hat jedoch den Nachteil, dass es bei unausgeglichenen Datensätzen keine repräsentative Aussage zur Güte der Klassifikation trifft.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.1}$$

Precision und Recall

Precision und Recall geben Auskunft über die Güte der Klassifikation einer bestimmten Klasse (hier: Positive). So gibt Precision den Anteil der als positiv klassifizierten Datenpunkte an, die tatsächlich positiv sind (siehe Gleichung (2.2)). Recall hingegen beschreibt den Anteil der positiven Datenpunkte, die das Modell korrekt klassifiziert hat (siehe Gleichung (2.3)).

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

F1-Maß

Das F1-Maß kombiniert Precision und Recall und wird nach Gleichung (2.4) berechnet.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.4}$$

Wie die Accuracy kann auch das F1-Maß für Mehrklassen-Probleme genutzt werden, indem das Maß für jede Klasse einzeln berechnet und von diesen Werten der Mittelwert gebildet wird, welcher dem macro-F1-Wert entspricht. Dieser gibt auch bei nicht ausbalancierten Daten jeder Klasse das gleiche Gewicht. Zur Berechnung des micro-F1-Maßes ohne Gewichtung der Klassen wird zuerst pro Klasse eine Konfusionsmatrix erstellt. Aus den vier Bestandteilen dieser Konfusionsmatrizen wird elementweise, also jeweils für TP, TN, FP und FN der Mittelwert gebildet, womit die micro-Precision und der micro-Recall errechnet werden. Diese werden dann zur Berechnung des micro-F1-Wertes genutzt.

Cross-Domain-Evaluation

Im Bereich der Textklassifikation beschreibt die Cross-Domain-Evaluation ein Szenario, bei dem sich die Testdaten in gewisser Weise von den Trainingsdaten unterscheiden [19]. So können die Testdaten in einer anderen Sprache verfasst sein (Cross-Language), einer anderen Textart entstammen (Cross-Genre) oder thematisch andere Inhalte behandeln (Cross-Topic). Modelle, die für Cross-Domain-Klassifikationen ausgelegt sind, bieten den Vorteil einer erweiterten Anwendbarkeit und die Möglichkeit auf mehr verfügbare Trainingsdaten [19]. Geht es bspw. in einem Ermittlungsverfahren darum die Urheberschaft einer belastenden E-Mail zuzuordnen, kann ein Modell auch mit anderen Textdaten wie Blogbeiträgen oder Restaurantbewertungen der potenziellen Autoren trainiert werden. Die Unterschiede in den Trainings- und Testdaten wirken sich außerdem auf die Performance des Modells aus. So wurden z.B. von Stamatatos [20] Accuracy-Abnahmen verzeichnet bei der Cross-Topic Authorship Attribution (AA) im Gegensatz zur Singel-Topic AA. Solche Beobachtungen machten auch Kaati et al. [21] in Bezug auf Cross-Genre AP.

2.2 Nutzen der Altersbestimmung im Kontext der Kriminalität

Gerade im Bereich der digitalen Kommunikation ist die Möglichkeit gegeben anonym zu bleiben oder sich als jemand anders auszugeben. Das wird sich insbesondere dann zunutze gemacht, wenn es sich um moralisch verwerfliche oder kriminelle Aktivitäten (wie Cybermobbing, Erpressung, Verbreitung von Fake-News etc.) handelt. Für die forensische Arbeit kann es also interessant sein, bestimmte digitale Texte (wie E-Mails, Chatnachrichten oder Blogbeiträge) der Person zuzuordnen, die sie verfasst hat. Für solch eine Arbeit können maschinelle Textklassifikationsmethodiken, wie im vorigen Abschnitt 2.1 erläutert, genutzt werden.

Spezielle Textklassifikationsaufgaben, die in diesem Zusammenhang für die Forensik interessant sein können, da sie darauf abzielen den Autor eines Textes auf die ein oder andere Weise einzuordnen, sind Authorship Attribution (AA), Authorship Verification (AV) und Author Profiling (AP).
Die ersten beiden Aufgaben beschäftigen sich damit herauszufinden, ob ein Text von einer bestimmten Person verfasst wurde und sind Variationen der Authorship Identification [22]. Dabei geht es bei AA darum, herauszufinden ob ein unbekannter Text einem bekannten Autor, von dem bereits Schriftstücke vorliegen, zugeordnet werden kann. AV hat zum Ziel zu erkennen, ob zwei gegebene Texte von derselben Person geschrieben wurden oder nicht.
Beim AP hingegen wird nicht zwischen individuellen Autoren unterschieden sondern zwischen Autorenklassen. So wird versucht Merkmale wie Alter, Geschlecht oder Muttersprache anhand des Textes zu ermitteln [23]. Das AP mit dem Ziel den Autor eines Textes einer Altersgruppe zuzuordnen ist Inhalt dieser Arbeit.

In der digitalen Forensik kann die Altersbestimmung anhand von verfassten Texten für verschiedene Bereiche eingesetzt werden. Die Möglichkeit falsche oder keine Angaben zu seiner Person zu machen, welche bei der Nutzung des Internets häufig besteht, erleichtert Regelverstöße und die Durchführung von Straftaten.
Die manuelle Überwachung von sozialen Netzwerken, um solche Vorfälle zu verhindern bzw. zu erkennen, ist bei der Masse an Inhalten mit einem sehr großen Zeit- und Personalaufwand verbunden. Computergestützte Systeme können hier zum Einsatz kommen, um die Arbeitslast zu verringern und zumindest eine Vorselektion der Daten vorzunehmen, wie bspw. das von van de Loo et al. [24] vorgeschlagene System. Potenzielle Probleme können dann an einen Moderator oder Strafverfolgungsbehörden weitergeben werden, welche die Situation dann genauer betrachten, um einzuschätzen, ob es sich tatsächlich um eine problematische oder illegale Aktivität handelt.

Ein Beispiel dafür sind Systeme, die die Kommunikation zwischen Erwachsenen mit illegaler Intention und Kindern frühzeitig erkennen. Ein Problem, welches in diese Kategorie fällt, ist Cybergrooming. Es meint den Prozess der Vorbereitung von Kindern auf spätere illegale sexuelle Begegnungen durch Erwachsene über das Internet [25]. Es ist nicht unüblich, dass solche Personen sich dabei als Kinder ausgeben, und die Opfer denken, sie würden mit einem Gleichaltrigen kommunizieren. Auf Grundlage dieses Verhaltens wurden bereits Systeme zur Erkennung von Grooming und Pädophilie gestaltet.

Zum Beispiel haben van de Loo et al. [24] einen Mechanismus geschaffen, welcher die Profilangaben von Nutzern sozialer Netzwerke im Hinblick auf Alter und Geschlecht mit den Informationen, die aus ihren geschriebenen Texten abgeleitet werden, vergleicht. Bei Unstimmigkeiten wird die Konversation von einem Klassifikator analysiert, der sexuelle Inhalte und Grooming-Verhalten detektiert.
Ein weiteres System zur frühzeitigen Aufdeckung von Grooming-Fällen durch die Verwendung von AP wurde von Ashcroft et al. [25] vorgestellt, die sich auf die Unterscheidung von Kindern und Erwachsenen, die sich als Kinder ausgeben, konzentrierten. Die Intention der Autoren bestand darin, Kinder während ihrer Onlinekommunikation über das reale Alter ihrer Chatpartner aufzuklären.
Auch Siva et al. [26] stellten ein System zur Grooming-Detektion vor, welches ein vorangestelltes Modul zur Alterserkennung aufweist, um anschließend die Kommunikation von Kindern kontinuierlich zu überwachen und bei Verdacht auf Grooming die Moderatoren zu alarmieren. Auf diese Weise sollte die Anzahl der zu überwachenden Konversationen reduziert und zudem das Eindringen in die Privatsphäre von zwei kommunizierenden Erwachsenen verhindert werden.

Ein weiterer Anwendungsbereich aus präventiver Sicht ist die Überprüfung der Einhaltung eines gesetzlichen Mindestalters für das Besuchen von Webseiten oder Plattformen mit dem Ziel Kinder und Jugendliche vor Sicherheitsrisiken zu schützen. Zum Beispiel ist für viele Social-Media-Plattformen mittlerweile ein Mindestalter in den Nutzungsbedingungen festgelegt (in vielen Fällen 13 Jahre), welches jedoch nicht alle Kinder davon abhält diese zu nutzen. So haben in einer Studie zur Internetnutzung von Kindern in den EU-Ländern durchschnittlich 28% der 9- bis 11-Jährigen angegeben, dass sie soziale Netzwerke mindestens täglich besuchen [27].
Automatisch zu erkennen, ob ein User nicht das Mindestalter einer Plattform erreicht hat, kann helfen diese Regelungen besser durchzusetzen, indem die detektierten Vorstöße zu einem Ausschluss von der Plattform führen.
Solche Sicherheitsmechanismen sind besonders dann interessant, um vor allem Kinder vor unangemessenen und potenziell verstörenden Inhalten, wie Videos und Bilder von gewalttätigen oder sexuellen Szenen, zu bewahren.

Die Altersbestimmung kann auch hilfreich bei der Aufklärung begangener Straftaten sein, indem sie zur Profilerstellung eines unbekannten Täters eingesetzt wird. Die Erkennung der Altersgruppe oder besser noch des genauen Alters anhand von Dokumenten, die in Zusammenhang mit einem Verbrechen gesichert wurden, kann den Täterkreis eingrenzen und so bei den Ermittlungen unterstützen. Zu solchen Dokumenten zählen zum einen solche mit offensichtlich rechtlicher Relevanz wie anonyme Droh- oder Erpresserbriefe, und zum anderen solche, die auf subtilere Weise die Verwicklung einer Person in Situationen mit strafrechtlichen Auswirkungen aufdecken können, wie Blog- und Tagebucheinträge oder geschäftliche und persönliche E-Mails [28].

3 Literaturdiskussion

Dieses Kapitel gibt einen Überblick über den Stand der Forschung zum Thema AP und geht dabei auf die verwendeten Features (Abschnitt 3.1) und Klassifikationsalgorithmen (Abschnitt 3.2) ein, stellt häufig verwendete Datensätze vor (Abschnitt 3.3) und präsentiert die erreichten Resultate (Abschnitt 3.4).

Im letzten Jahrzehnt nahm die Aufmerksamkeit für den Task des AP zu, vor allem durch seine Auftritte bei den Shared Tasks der PAN in den Jahren 2013 bis 2018 [23, 29–33]. Dabei standen meist das Geschlecht (jedes Jahr) und Alter (2013-2016) im Fokus, aber auch die Sprachvarietät (2017) und Persönlichkeitsmerkmale (2015) wurden betrachtet. Auch außerhalb der PAN erforschte man diese und weitere Autorenmerkmale, wie ethnische Herkunft [34–36], Bildungsniveau [37] und Muttersprache [38].

Ansätze, die nun speziell das Alter von Autoren erfassen wollen, betrachten dies oftmals in Kombination mit anderen Merkmalen, am häufigsten zusammen mit dem Geschlecht [z.B. 23, 29, 31, 39–41]. Bei solch kombinierten Vorhaben wurde entweder im Bereich der Features und/oder des Klassifikationsprozesses nach Merkmal unterschieden [z.B. 42–47] oder es wurde das gleiche System zur Klassifizierung von beiden Merkmalen verwendet [z.B. 48–50]. Teilweise wurden die Merkmale auch durch die Bildung von neuen Subklassen, die das Alter und das Geschlecht abdecken, als Einheit betrachtet [z.B. 51–53]. Letztere Vorgehensweise führt zu dem Schluss, dass es Zusammenhänge zwischen Texten von Autoren bestimmten Alters und bestimmten Geschlechts gibt. Eben solche stellten z.B. Argamon et al. [54] fest: Sie fanden Überschneidungen bei den jeweils 1.000 Worten mit dem höchsten Information Gain für die Kategorien Alter und Geschlecht. Diese Wörter wurden - mit einer Ausnahme - eher von weiblichen und jüngeren bzw. von männlichen und älteren Autoren verwendet.

Da der Kern dieser Arbeit jedoch die Bestimmung des Alters ist, wird bei der nachfolgenden Darstellung von verwendeten Features und Klassifikationsalgorithmen nur auf diejenigen für die Altersklassifizierung eingegangen, auch wenn die erwähnten Ansätze zusätzlich andere Merkmale von Autoren erforschten.

3.1 Verwendete Features

Dieser Abschnitt befasst sich mit den Features, die in verschiedenen Ansätzen für den AP-Task herangezogen wurden. Sie werden im Folgenden getrennt nach lexikalischen und weiteren Features betrachtet. Erstere fanden merklich am häufigsten Verwendung und letztere umfassen Features unterschiedlicher Art, welche in der Verwendungshäufigkeit variieren. Zudem wird im letzten Abschnitt auf die verschiedenen genutzten Feature-Selektionsmethoden eingegangen.

3.1.1 Lexikalische Features

Ein Großteil der Ansätze zum AP in Bezug auf das Alter nutzt n-Gramme [z.B. 43, 46, 55, 56]. Dabei unterscheiden sich teilweise die n-Gramm-Typen (diese umfassen n-Gramme von Worten/Termen, Zeichen und Part-of-Speech (POS) Tags) sowie die für n eingesetzten Werte. So verwendeten Poulston et al. [49] und Gencheva et al. [46] nur Wort-Uni- und -Bigramme, und Bougiatiotis und Krithara [43] nur Wort-Trigramme. Ausschließlich Zeichen-Trigramme nutzten Garciarena Ucelay et al. [57] und von Cruz et al. [44] wurden nur POS-Tag-n-Gramme verwendet, indem nach der Anwendung eines POS-Taggers die POS-Tags anstelle der Wörter für die Uni-, Bi- und Trigramm-Bildung genutzt wurden.
Wort- und Zeichen-n-Gramme in Kombination nutzten Modaresi et al. [47] und Rahman und Akter [58], mit Wort-Uni- und Bigrammen sowie Zeichen-n-Grammen mit $n = 4$ bzw. $n = \{3, 4, 5\}$ respektive. Eine Kombination von Wort- und POS-Tag-n-Grammen nutzten bspw. Flekova und Gurevych [51] mit entsprechend $n = \{1, 2\}$ bzw. $n = \{1, 2, 3, 4\}$.
In einem Versuch von Peersman et al. [56], bei dem Wort- und Zeichen-n-Gramme jeweils im Bereich $n = \{1, 2, 3\}$ einzeln als sechs verschiedene Feature-Sets getestet wurden, erreichten alle Wort-n-Gramm-Sets höhere Accuracies als die Zeichen-n-Gramme, wobei die Wort-Unigramme am besten abschnitten. Bei dem Versuch von Ameer et al. [59] wurde die Nützlichkeit von Wort-, Zeichen- und POS-Tag-n-Grammen verglichen, wobei auf mehrere Datensätze die höchste Accuracy mit einem Feature-Set bestehend aus Wort-1- bis -4-Grammen erreicht wurde.

Zur Repräsentation der vorgestellten lexikalischen Features wurde bspw. das BoW-Modell verwendet [43, 46, 53], welches die Dokumente als Termfrequenz-Vektoren darstellt. Bei den gewählten Termen konnte es sich um die einzelnen Wörter (Unigramme) im Dokument handeln [53], es wurden aber auch andere Wort-n-Gramme verwendet, z.B. Wort-Trigramme [43] oder Wort-Uni- und -Bigramme [46].
Häufiger als die BoW-Repräsentation wurde jedoch die TF-IDF-Darstellung genutzt [z.B. 42, 45, 49, 60], welche die Termfrequenz (TF) mit der Inversen Dokumentenfrequenz (IDF) verbindet und wo die Vektoren aus gewichteten Werten bestehen. Auch hierbei waren die gewählten Terme variabel je nach Ansatz, Cruz et al. [44] verwendeten bspw. neben Wort- auch POS-Tag-n-Gramme.
Ein von Rahman und Akter [58] durchgeführter Vergleich zwischen BoW, TF-IDF und Wortlänge zur Repräsentation sowie der Kombinationen derer zeigte die besten Klassifikationsresultate bei der alleinigen Darstellung durch die TF-IDF.
Auch eine erstmals bei der PAN 2013 von López-Monroy et al. [52] vorgestellte niedrig-dimensionale und dichte Repräsentation der Dokumente, genannt Second Order Attributes (SOA), wurde seither mehrfach angewendet [61–63]. Die Methode basiert auf Beziehungen zwischen Termen, Dokumenten und Klassen, wobei Terme jeglichen als Features genutzten Texteinheiten entsprechen können [52]. Für den Task des AP bei der PAN 2014 vertieften López-Monroy et al. [64] diese Methode noch, indem sie die Beziehungen zwischen Dokumenten und automatisch gefundenen feineren Subklassen innerhalb der eigentlichen Klassen mithilfe der SOA darstellten.

Seltener wurden Wort-Embeddings zur Repräsentation der Dokumente verwendet [65]. Bayot und Gonçalves [66] nahmen einen Vergleich zwischen der Verwendung von Wort-Embeddings (erstellt mit word2vec [67]) und TF-IDF-Repräsentation vor. Dieser ließ jedoch keine genera-

lisierbaren Schlussfolgerungen bezüglich einer überlegenen Nützlichkeit einer der beiden Repräsentationen zu.
López-Santillán et al. [68] kombinierten Wort-Embeddings (erzeugt mit word2vec [67], fastText [69], BERT [16]) mit einer Gewichtung der Worte mithilfe klassischer Termfrequenz-Statistiken wie TF oder TF-IDF und erzeugten so neuartige Dokument-Embeddings.

3.1.2 Weitere Features

Obwohl in manchen Fällen ausschließlich n-Gramme als Features genutzt wurden [z.B. 43, 50, 56, 58], zog man oftmals weitere der nachfolgenden Features hinzu oder nutzte nur diese und verzichtete ganz auf n-Gramme [z.B. 48, 70, 71].

Häufig genutzte Features sind durch statistisches / zählendes Vorgehen aus den Texten extrahiert und je nach Ansatz in unterschiedlichen Kombinationen verwendet worden [z.B. 46, 51, 72, 73]. Dazu gehören u.a. Dokumenten-, Satz- und Wortlängen, Anzahl an Wörtern und Sätzen sowie Häufigkeiten bzw. Anteile von bestimmten Worten, POS- oder Named Entity-Tags, Satzzeichen, Links, Emoticons, Großbuchstaben etc. Auch Features basierend auf Termen von vordefinierten Wortlisten, die beispielsweise Emotionsworte, familienbezogene Worte oder Abkürzungen enthalten [42, 51, 55], wurden ebenso genutzt wie falsch geschriebene [46, 47, 55, 62, 72] und Slangwörter [40, 62], die nicht in Wörterbüchern enthalten sind. Gencheva et al. [46] verwendeten zudem besondere Erwähnungen, die auf das Alter hinweisen (z.B. „I am“ gefolgt von einer Zahl) und Pimas et al. [71] arbeiteten mit der Konkretheit von Worten.

Maße, die das Wortreichtum (Vocabulary Richness) eines Dokuments ermitteln, fanden ebenso Anwendung als Features. Das dabei am häufigsten verwendete Maß ist die Anzahl der Worte, die nur ein einziges mal im Dokument erscheinen (Hapax Legomena), bspw. ermittelt von Busger op Vollenbroek et al. [62], Lim et al. [72] und Ashraf et al. [73], wobei letztere noch fünf weitere Maße berechneten: Brunet W [74], Honoré R [75], Sichel S [76], Simpson D [77] und Yule K [78].

Auch die Verständlichkeit (Readability) von Texten kam bspw. in mehreren Ansätzen der PAN 2013 zum Einsatz. So ermittelten Flekova und Gurevych [51] sieben verschiedene Maße: Flesch-Kincaid Grade Level [79], Automatic Readability Index [80], LIX Index [81], Coleman-Liau Index [82], Flesch Reading Ease [83], SMOG [84], Gunning-Fog Index [85]. Bei Gillam [48] ergab sich die Lesbarkeit durch die Verteilung von Satz- und Wortlängen, ähnlich dem Automatic Readability Index und Meina et al. [55] wendeten neben dem Flesch-Kincaid Grade Level und dem Flesch Reading Ease auch die Dale-Chall-Formel [86] an.

Die Themen der Dokumente wurden ebenfalls genutzt, um Features zu extrahieren. Zwei häufig verwendete Methoden, um Themen zu ermitteln, sind dabei die Latent Dirichlet Allocation (LDA) [87], die bspw. eingesetzt wurde von [42, 49, 70, 88] sowie die Latent Semantic Analysis (LSA) [89], die in [55, 61] herangezogen wurde. Poulston et al. [49] ermittelten mit der LDA zehn Themen und verwendeten deren Existenz in den Dokumenten als Feature. Santosh et al. [88] extrahierten ebenfalls mittels LDA einerseits Themen aus dem gesamten Trainings-Korpus und andererseits individuelle Themen aus den Dokumenten pro Klasse. Die

Themenverteilung in den Dokumenten nutzten Pavan et al. [70] als Feature. Auf anderem Wege, mittels der Non-Negative Matrix Factorization [90], extrahierten Gencheva et al. [46] Themen aus den Dokumenten.

Als zusätzliches Feature wurde in einigen Arbeiten auch die Stimmung verwendet [46, 71, 91]. Dabei ermittelten Gencheva et al. [46] mithilfe eines Lexikons Stimmungswerte für positives und negatives Sentiment sowie acht verschiedene Emotionen, von denen nur drei (Freude, Überraschung, Vertrauen) final verwendet wurden, für alle Wörter und akkumulierten diese Werte pro Dokument. Pimas et al. [71] maßen ebenfalls Lexikon-basiert für alle Wörter einzeln die Polarität mit Werten im Bereich [-1,1] und ermittelten pro Dokument die maximale, minimale und durchschnittliche Polarität sowie ihre Standardabweichung. Von Suman et al. [91] wurde dagegen den Dokumenten ein negatives (-1), neutrales (0) oder positives (1) Sentiment zugeordnet.

Zudem wurden vereinzelt Features aus dem Information Retrieval (IR) genutzt [92, 93], was auf der Annahme beruht, dass sich Autoren gleichen Alters ähnlicher ausdrücken als unterschiedlichen Alters. Dabei ließ man zuerst ein IR-System sämtliche Trainingsdokumente indizieren. Anschließend wurden die zu klassifizierenden Texte als Suchanfragen verwendet zum Abruf der k ähnlichsten Texte. Die Ähnlichkeit basiert dabei auf Features, die mittels der Cosine- und Okapi-Metriken extrahiert wurden.

Weiterhin kamen ähnlichkeitsbasierte Features zum Einsatz [72, 94]. Lim et al. [72] erzeugten TF-IDF-Werte jeweils insgesamt für alle Dokumente jeder Klasse, um diese mit den TF-IDF-Werten von den zu klassifizierenden Texten zu vergleichen und die sich daraus ergebenden Ähnlichkeitswerte als Features zu nutzen. Moreau und Vogel [94] erstellten repräsentative Häufigkeitsverteilungen bestimmter klassenspezifischer n-Gramme für jede Klasse, um die Distanz zwischen diesen Verteilungen und denen in den zu klassifizierenden Dokumenten als Features zu verwenden.

3.1.3 Feature-Selektion

Bei einer hohen Anzahl von Features (oftmals erreicht durch die Nutzung von n-Grammen) wurde in mehreren Ansätzen durch verschiedene Selektionsmethoden, mit dem Ziel die relevantesten Features herauszufiltern, die Zahl verringert. Zwei häufig verwendete Methoden dafür sind die Nutzung derjenigen k Features mit den höchsten χ^2-Werten [42, 44, 50, 56] sowie mit dem größten Information Gain [39, 51, 73].

Mithilfe der Transition Point Technique [95] wählten Markov et al. [63] ihre textuellen Features aus, indem sie die Technik dazu verwendeten das Vokabular in hoch- und niedrigfrequente Wörter aufzuteilen und diejenigen Worte, die näher an dem Wert des Transition Points waren (mittelfrequente Wörter) als Features auszuwählen.

Santosh et al. [88] nutzten zur Selektion das Verhältnis der Verteilung bestimmter textueller Features über die verschiedenen Klassen. Dafür berechneten sie erst jeweils ihre Häufigkeiten in den Dokumenten jeder Klasse und dann für jedes das Verhältnis dieser Häufigkeiten zwischen den Klassen, um diejenigen zu finden, welche am unausgeglichensten über die Klassen verteilt sind.

Eine weitere Methode zur Selektion ist die Auswahl über die Häufigkeit des Auftretens bestimmter Worte. So wählten Mechti et al. [96] die 200 häufigsten Worte des gesamten Korpus, Busger op Vollenbroek et al. [62] dagegen die 2.500 häufigsten Worte pro Klasse.

Eine neue Methode der Term-Selektion, bei der neben der Termverteilung auf die Klassen auch die Art der Phrasen, in denen die Terme vorkommen, berücksichtigt werden, wurde von Ortega-Mendoza et al. [97] vorgestellt.

Auch Radha und Chandra Sekhar [98] stellten eine neue Feature-Selektionsmethode vor, die Distributional Class Specific Correlation-Based Technik (DCC), zum Erkennen von Termen mit hoher Abhängigkeit von der Klasse. Im Vergleich zu drei traditionellen Methoden (χ^2, Mutual Information, Information Gain) lieferte die DCC in ihren Experimenten, welche auf einen aus Hotelbewertungen bestehenden Datensatz (Subkorpus des Datensatzes der PAN 2014, siehe Tabelle 3.1) durchgeführt wurden, die besten Ergebnisse.

3.2 Verwendete Klassifikationsalgorithmen

Der am häufigsten verwendete Klassifikator beim AP bezüglich des Alters ist die Support Vector Machine (SVM), genutzt bspw. von [52, 53, 58, 61], was sich durch ihre simple Anwendung und gute Performance erklären lässt. Weitere verwendete Klassifikatoren sind Random Forest [55, 71, 73], Logistische Regression [47, 51], Decision Tree [48], Maximum Entropy [70], Bayesian Multinomial Logistic Regression [99] und Multi-Class Real Winnow [39]. Weiterhin fanden vereinzelt distanz- bzw. ähnlichkeitsbasierte Ansätze Anwendung, wie die Klassifizierung mittels kNN [100] oder dem Simplified Profile Intersection Maß [57]. In manchen Fällen kamen Deep Learning (DL) Algorithmen, wie BERT [41, 101, 102] und LSTM [91], zum Einsatz.

Klassifikationsprozesse unter Verwendung mehrerer (unterschiedlicher) Klassifikatoren in Kombination wurden ebenfalls durchgeführt. So wendeten Agrawal und Gonçalves [60] die Stacking-Methode an, mit vier Basis-Klassifikatoren (Naive Bayes Multinominal, Simple Logistics, Naive Bayes, SVM), deren Entscheidungen der Input für den finalen Meta-Klassifikator (SVM) sind. Stacking nutzten auch Cruz et al. [44], wendeten die zwei initialen Klassifikatoren (SVMs) aber auf unterschiedliche Feature-Sets an und gaben die Wahrscheinlichkeiten für jede Klasse an den finalen Klassifikator, gebildet mit dem regelbasierten Algorithmus JRip [103], weiter. Auch Santosh et al. [88] klassifizierten jedes ihrer drei Feature-Sets einzeln (Style und Inhalt mittels SVM, Themen mit Maximum Entropy) und führten mit den erhaltenen Werten die finale Klassifikation mithilfe von Decision Tree durch. Ein Multi-Genre Ensemble Modell bestehend aus drei Logistische Regression-Klassifikatoren für drei verschiedene Genre nutzten Zahid et al. [50].

Nun ist sehr auffällig, dass im Bereich des AP DL-Algorithmen vergleichsweise selten verwendet werden. Basile et al. [104] versuchten sogar zu belegen, dass simplere Ansätze effektiver sind und in ihren Experimenten zeigte sich diese Theorie auch als wahr: Die besten Ergebnisse wurden mit einer n-Gramm-basierten SVM erreicht. Jedoch geben sie zu bedenken, dass es andere Möglichkeiten für die Überlegenheit des simpleren Ansatzes gibt. So sind sie der Ansicht, dass mit einer größeren Datenmenge und einem Fokus auf der Feinabstimmung der Hyperparameter bessere Ergebnisse durch ein DL-Modell nicht auszuschließen sind.

3.3 Häufig verwendete Datensätze

Dieser Abschnitt stellt Datensätze vor, welche für Forschungsarbeiten im AP-Bereich bezüglich des Alters, häufig verwendet wurden. In Tabelle 3.1 sind diese zusammengefasst dargestellt mit Details hinsichtlich der Dokumententypen, deren Sprache und die abgedeckten Klassen sowie zugehörige Studien, in denen sie Verwendung fanden. Teilweise wurden nur Subkorpora genutzt, welche in der Tabelle in Klammern hinter den jeweiligen Arbeiten vermerkt sind. All diese Datensätze adressieren neben dem Alter auch mindestens eine weitere Kategorie (zumeist das Geschlecht). Jedoch sind in der Übersicht nur die Teile der Korpora betrachtet, die für das Alter annotiert sind.

Tabelle 3.1: Übersicht häufig verwendeter Datensätze im Bereich des AP in Bezug auf das Alter.

Bezeichnung	Dokumententypen	Sprachen	Klassen	Studien
Blog-Authorship-Korpus	Blogbeiträge	en	3 Klassen: 13-17, 23-27, 33-42 bzw. -47	[39], [54], [99], [40], [97]
PAN13-Korpus	Blogbeiträge	en, es	3 Klassen: *10s* (13-17), *20s* (23-27), *30s* (33-47)	[23], [68]
PAN14-Korpus	Social Media Beiträge Blogbeiträge Tweets Hotelbewertungen	en, es en, es en, es en	5 Klassen: 18-24, 25-34, 35-49, 50-64, 65+	[29], [21] (Blogs, Social Media), [98, 105, 106] (Hotelbewertungen, Training), [59] (en, ohne Tweets), [68], [97] (en, Training)
PAN15-Korpus	Tweets	en, es	4 Klassen: 18-24, 25-34, 35-49, 50+	[30], [21], [68], [58]
PAN16-Korpus	Tweets (Training) Blogbeiträge (Test)	en, es	5 Klassen: 18-24, 25-34, 35-49, 50-64, 65+	[31], [59, 102] (en, Training), [68]

In sämtlichen Datensätzen ist die Altersklassifizierung als Mehrklassen-Problem annotiert mit drei bis fünf Klassen, die teilweise fließend ineinander übergehen (PAN14-, PAN15-, PAN16-Korpus) und in anderen Fällen stärker von einander abgegrenzt sind, indem mehrere Jahre

dazwischen liegen (Blog-Authorship-, PAN13-Korpus). Die einzelnen Klassen umfassen dabei immer mindestens fünf Jahre.
Jeder Datensatz enthält zumindest einen englischsprachigen Subkorpus oder enthält nur englische Dokumente. Die meisten PAN-Datensätze enthalten zusätzlich spanischsprachige Subkorpora.
Die häufigsten Dokumententypen sind Tweets und Blogbeiträge. Die weiteren Dokumente sind ebenfalls informell geschriebene im Internet veröffentlichte Texte. Weiterhin handelt es sich in allen Fällen um unausgeglichene Datensätze bezüglich der Verteilung der Altersklassen. So sind (meist) die jüngeren und/oder mittleren Altersgruppen mehr vertreten als die älteren, was mit dem Ursprung der Daten zusammenhängen kann, da ältere Menschen das Internet seltener nutzen [107].

Der Blog-Authorship-Korpus besteht aus englischsprachigen Blogbeiträgen, die im August 2004 gesammelt wurden. Da viele der Blogs über mehrere Jahre gehen und es so zu Uneindeutigkeiten bezüglich des Alters der Blog-Autoren kommen kann, wurden Zwischengruppen entfernt und Ergebnis waren die drei in der Tabelle 3.1 dargestellten Altersklassen.
Weiterhin häufig verwendet wurden die für die AP Shared Tasks erstellten PAN-Korpora der Jahre 2013 bis 2016. Der PAN13-Korpus ähnelt bezüglich dem Dokumententyp und der Klasseneinteilung sehr stark dem Blog-Authorship-Korpus, jedoch enthält er zusätzlich einen spanischsprachigen Subkorpus und wurde im Vergleich zu den restlichen Datensätzen seltener genutzt.
Die anderen PAN-Korpora unterscheiden sich von den vorherigen beiden dadurch, dass sie (1) mehr Klassen haben und (2) diese Klassen lückenlos ineinander übergehen. Der PAN14-Korpus besteht zudem aus vier Subkorpora, die verschiedene Dokumententypen enthalten, was die Möglichkeit bietet, Ansätze bezüglich ihrer Nützlichkeit für unterschiedliche Genre vergleichen zu können.
Der PAN16-Korpus ist darüber hinaus der einzige von den dargestellten Datensätzen, welcher speziell auf das Problem der Cross-Genre-Evaluation ausgelegt ist, da dieser Trainings- und Testdaten verschiedenen Ursprungs enthält, was unterschiedliche Textstrukturen nach sich zieht. Außerhalb der PAN16 wurde dieser Datensatz jedoch kaum verwendet, zumindest nicht komplett in seiner Auslegung als Cross-Genre-Korpus.

3.4 Überblick über die erreichten Resultate

Dieser Abschnitt stellt Ansätze vor, die im AP hinsichtlich des Alters sehr gute Ergebnisse erzielten. Um eine gewisse Vergleichbarkeit zu gewährleisten, wurden dabei auch die verwendeten Datensätze berücksichtigt.

Eine der erfolgreicheren Herangehensweisen ist die Unterscheidung der Terme in den Dokumenten nach „function words“ und „content words“ und die Nutzung dieser als Features [39, 40, 99]. Dabei sind erstere unabhängig vom Inhalt (z.B. Artikel, Präpositionen) und bilden die Verbindung zwischen den inhaltgebenden „content words“ .
So teilten Argamon et al. [99] die in den Texten vorkommenden „function words“ in eine Taxonomie ein, die sich an POS orientiert und nutzten als Features die normalisierten Anzahlen der „function words“ pro POS-Kategorie. Zusätzlich suchten sie die 1.000 „content words“ (in

Paper bezeichnet als „individual words“), die die Klassen am besten von einander trennen. Zur Klassifikation verwendeten sie den Bayesian Multinomial Regression Algorithmus und erreichten eine Accuracy von 0,777.
Auch Schler et al. [39] nutzten POS, „function words“ und die 1.000 Unigramme („content words“ und Wörter spezieller Wortklassen) mit dem höchsten Information Gain. Zusätzlich zogen sie Blog-spezifische Features (z.B. Blog-Worte (Neologismen wie lol, haha, ur), Links) hinzu. Mit dem Multi-Class Real Winnow (MCRW) Algorithmus erreichten sie eine etwas geringere Accuracy von 0,762.
Eine höhere Accuracy von 0,8038 erreichten Goswami et al. [40], indem sie zusätzlich zu den 35 „content words“ mit dem höchsten Information Gain auch 52 Worte, die nicht im Wörterbuch stehen (wie Slang, Emoticons, Chat-Abkürzungen etc.) und die durchschnittliche Satzlänge als Features hinzuzogen. Die Klassifikation führten sie mit dem Naive Bayes Klassifikator durch. Diese drei Ansätze arbeiteten mit dem gleichen Datensatz, dem Blog-Authorship-Korpus.

Die nächsten drei Ansätze verwendeten alle den Hotelbewertungen-Subkorpus des PAN14-Korpus (nur Trainingsdaten) und konzentrierten sich auf Verbesserungen in der Term-Gewichtung [105, 106] bzw. Term-Selektion [98], wodurch sie jeweils Accuracies über 0,80 erreichten.
Das von Reddy Seelam et al. [105] vorgeschlagene Profile specific Supervised Term Weight (PSTW) Maß erweitert Term- und Dokumentenfrequenz dahingehend, dass es auch die Klasse, zu der das Dokument gehört, einbezieht, indem es die Anzahl der Terme sowie Dokumente innerhalb und außerhalb der Dokumentenklasse beachtet. Das Maß weist denjenigen Termen mehr Gewicht zu, die in einer Klasse häufig auftreten und in den anderen seltener sowie denjenigen, die in mehr Dokumenten einer Klasse vorkommen und in weniger Dokumenten der anderen Klassen. In ihren Experimenten erreichten sie die höchste Accuracy von 0,8295 mithilfe dieser Art der Term-Gewichtung der 6.000 häufigsten Terme und dem Random Forest Klassifikator.
Das von Kavuri und Kavitha [106] vorgeschlagene Term-Gewicht-Maß beachtet zusätzlich zu den Faktoren des PSTW-Maßes auch die Anzahl derjenigen Dokumente jeweils innerhalb und außerhalb der Dokumentenklasse, die den Term nicht enthalten. Die höchste Accuracy von 0,8258 und somit etwas geringer als die des vorigen Ansatzes wurde mit den 10.000 Termen, die nach dem entwickelten Feature-Selektionsalgorithmus als am relevantesten eingestuft wurden, und dem Random Forest Klassifikator erreicht.
Radha und Chandra Sekhar [98] arbeiteten nicht mit einem speziellen Term-Gewicht-Maß, sondern nutzten das BoW-Modell. Jedoch brachte ihre vorgeschlagene Technik der Feature-Selektion DCC (siehe Abschnitt 3.1.3) Verbesserungen sowohl gegenüber der Verwendung der häufigsten Terme (0,6654) als auch den drei populären Feature-Selektionstechniken χ^2 (0,7658), Mutual Information (0,6559) und Information Gain (0,7109) mit einer Accuracy von 0,8108 bei der Nutzung der 8.000 Terme mit dem höchsten DCC-Wert und einer Klassifikation mittels Random Forest.

Ansätze, die die Cross-Genre-Evaluation behandeln, sind seltener. Aus einem vergleichenden Versuch von Kaati et al. [21], bei dem ein Modell einmal auf Testdaten des gleichen Genres wie die Trainingsdaten und einmal auf Testdaten eines anderen Genres angewendet wurde, geht hervor, dass die Cross-Genre-Evaluation anspruchsvoller ist und Genre-unabhängiger Features bedarf.
Die besten Ergebnisse auf diesem Gebiet wurden im Zuge der PAN 2016 auf den PAN16-

Korpus erreicht. Busger op Vollenbroek et al. [62] verzichteten explizit auf Genre-spezifische Features und nutzten neben Wort-, Zeichen- und POS-Tag-n-Grammen zusätzlich die folgenden Features: (1) Anzahl der Sätze, die mit einem Großbuchstaben starten, (2) Anteil an großgeschriebenen Worten, (3) Anteil an Großbuchstaben, (4) Anteil der auf Satzzeichen endenden Sätze, (5) Anteil an Satzzeichen, (6) durchschnittliche Wort- und Satzlänge, (7) Anteil an falsch geschriebenen und Slangworten, (8) Anteil an Worten, die nur einmal verwendet wurden, (9) Verhältnis der Häufigkeit eines Terms in einer Klasse gegenüber den anderen Klassen (davon die 2.500 häufigsten Worte pro Klasse) und (10) Anteil an Emoticons. Weiterhin nutzten sie die SOA-Repräsentation (vorgestellt in Abschnitt 3.1.1), um bestimmte Features in relativen Werten darzustellen. Mit einer SVM zur Klassifikation erreichten sie eine Accuracy von 0,5897 auf den englischen und 0,5179 auf den spanischen Teil des Korpus.

Der Vollständigkeit halber wird in diesem Absatz ein kurzer Blick auf Ergebnisse mittels DL-Algorithmen geworfen. Wie in Abschnitt 3.2 bereits angesprochen, wird beim AP hinsichtlich des Alters auf diese eher verzichtet, was auch die Vergleichbarkeit einschränkt.
Eine Möglichkeit für den direkten Vergleich unterschiedlicher Ansätze, bei der für die Altersklassifizierung auch zwei DL-Methoden Anwendung fanden, bot jedoch der Author Profiling and Deception Detection in Arabic (APDA) Shared Task der PAN@FIRE 2019 [108]. Auch hier ergab sich Grund für die Zurückhaltung gegenüber der Nutzung von DL, denn es zeigte sich eine Überlegenheit traditioneller Machine Learning Algorithmen mit manueller Feature-Extraktion: Das Gewinner-Team führte die Klassifikation mittels Logistischer Regression basierend auf Wort- und Zeichen-n-Grammen durch und erreichte eine Accuracy von 0,625 [109]. Dagegen erzielte der auf BERT basierende Ansatz von Zhang und Abdul-Mageed [101] eine Accuracy von 0,5472 (mittlerer Rang im Vergleich) und die niedrigsten Accuracies von 0,275 bzw. 0,2222 wurden mit dem LSTM-Ansatz von Suman et al. [91] erzielt. DL scheint für das AP insbesondere bezüglich des Alters somit den traditionellen Machine Learning Algorithmen durchaus unterlegen zu sein.

4 Daten

Die in dieser Arbeit verwendeten Datensätze wurden aus Daten von drei verschiedenen Quellen zusammengestellt. In den nachfolgenden Abschnitten werden diese sortiert nach der Art der Daten vorgestellt. Dabei wird der Prozess der Anpassung an die Ansprüche dieses Projekts sowie der Grund ihrer Verwendung näher betrachtet.

4.1 Blog-Daten

In erster Linie wurden für diese Arbeit Daten verwendet, die aus Blogbeiträgen bestehen. Der erste Datensatz, welcher dabei Anwendung fand ist der englischsprachige Teil des Korpus der PAN 2013 für den Task des AP [23]. Dieser besteht aus Blogbeiträgen von 236.600 verschiedenen Autoren in den Trainingsdaten und 25.440 in den Testdaten. Er wurde aufgrund seiner Zugänglichkeit ausgewählt sowie aufgrund dessen, dass er neben Dokumenten von erwachsenen Autoren auch Dokumente Minderjähriger enthält.

Die drei Klassen, die in diesem Korpus abgedeckt werden, sind Autoren im Alter von: 13 bis 17 Jahren bezeichnet als *10s*, 23 bis 27 Jahren bezeichnet als *20s* und 33 bis 47 Jahren bezeichnet als *30s*. Die Klasse der *10s* liegt mit ihrer Anzahl an Datenpunkten deutlich unter denen der anderen beiden Klassen. Um diese Unausgeglichenheit zu verringern, wurde die unterrepräsentierte Klasse um Daten ähnlichen Ursprungs (gleicher Kontext, gleiches Genre) erweitert. Dazu wurden Teile des Blog-Authorship-Korpus, zusammengetragen von [39], verwendet, welcher ebenso wie der PAN13-Korpus aus Blogbeiträgen besteht. Diese stammen von insgesamt 19.320 verschiedenen Autoren. Da das genaue Alter der Autoren Teil des Korpus ist, konnten diejenigen Datenpunkte, welche von Autoren unter 18 Jahren (13 bis 17-Jährige) verfasst wurden, extrahiert und der Klasse *10s* des bereits bestehenden Datensatzes zugefügt werden.

So hat sich die Anzahl an Trainingsdaten um 8.240 Autoren erhöht und lag schließlich bei 244.840, wobei die Anzahl an Autoren mit der Anzahl an Datenpunkten gleichzusetzen ist (siehe dazu auch Abschnitt 5.1). Die Unausgeglichenheit ist trotzdem weiterhin recht groß, jedoch entspricht ein Ungleichgewicht auch einem realistischen Szenario.

Tabelle 4.1: Verteilung der Daten in dem Trainings- und Testkorpus für die 3-Klassen-Klassifikation.

Klassen	**Trainingsdaten**	**Testdaten**
	Training 1 Blogbeiträge	*Test 1* Blogbeiträge
10s	17.200 + 8.240	1.776
20s	85.800	9.216
30s	133.600	14.448
	244.840	25.440

Diese Blog-Daten stellen nun einen Trainings- und Testkorpus bestehend aus drei Klassen bereit, deren Zusammensetzung in Tabelle 4.1 dargestellt ist. Da im Bereich der Strafverfolgung gerade die Unterscheidung zwischen minderjährigen und erwachsenen Personen oft von Bedeutung ist, wurden die Blog-Daten zusätzlich an eine binäre Klassifikation angepasst. Dabei entsprechen die *10s* der Klasse *Minderjährige* und die *20s* und *30s* wurden zur Klasse *Erwachsene* zusammengefasst. Die Verteilung dieser Daten ist in Tabelle 4.2 unter *Training 1b* sowie *Test 1b* zu finden.

Tabelle 4.2: Verteilung der Daten in den Trainings- und Testkorpora für die binäre Klassifikation.

Klassen	**Trainingsdaten**	**Testdaten**	
	Training 1b Blogbeiträge	*Test 1b* Blogbeiträge	*Test 2* Chatnachrichten
Minderjährige	17.200 + 8.240	1.776	2.979
Erwachsene	219.400	23.664	2.979
	244.840	25.440	5.958

4.2 Grooming-Daten

Weiterhin wurde ein zweiter Datensatz zusammengestellt, der als Testdatensatz für die finale Cross-Genre-Evaluation genutzt werden sollte. Hierfür wurde auf einen insbesondere für die Strafverfolgung relevanten Datensatz zurückgegriffen: der Korpus der PAN 2012 für den Task der Sexual Predator Identification [110].

Ursprünglich beinhaltet der Trainingsteil dieses Korpus rund 60.000 Dokumente und der Testteil rund 155.000 Dokumente, wobei ein Dokument einer Chat-Konversation entspricht. Die Daten des Korpus setzen sich aus Chats unterschiedlicher Herkunft zusammen. So handelt es sich bei den Cybergrooming-Daten (die Positiv-Beispiele für den Task) um Chats der Perverted Justice Website[1] und bei den Nicht-Grooming-Daten um Internet Relay Chat (IRC) Logs sowie Omegle-Chats (Online Sexgespräche zwischen Erwachsenen).

Im Rahmen dieser Arbeit soll die Altersbestimmung von Chatpartnern zur Unterstützung der Aufdeckung von Grooming-Fällen eingesetzt werden. Dementsprechend muss für die verwendeten Daten zum Trainieren und Testen der Modelle eine Altersangabe vorhanden sein. Bei dem eben vorgestellten Datensatz der PAN 2012 ist das jedoch nicht der Fall und das Alter eines Chatpartners ist nur eingeschränkt erkennbar. Es kann lediglich abgeleitet werden, dass sich diejenigen Chatteilnehmer, welche als Sexualstraftäter gekennzeichnet sind, im Erwachsenenalter befinden und dass es sich bei ihren Chatpartnern um Minderjährige handelt. Hierzu muss gesagt werden, dass es sich bei den Minderjährigen um Pseudo-Opfer handelt [110]. Tatsächlich waren es Erwachsene, die sich als Kinder ausgaben und versuchten ihren Schreibstil zu imitieren. Da es jedoch keine anderen öffentlichen Datensätze zur Cybergrooming-Erkennung gibt, wurde dieser Korpus dennoch verwendet.

[1] http://www.perverted-justice.com/

Zur Verwendung der Daten im Kontext dieser Arbeit mussten sie nun angepasst werden. Da für alle anderen Chatteilnehmer eine Altersbestimmung nicht möglich ist, wurden nur die Chats, bei denen einer der Chatpartner als Sexualstraftäter markiert ist, herausgefiltert. Im gesamten Korpus gibt es 396 verschiedene Groomer, die an insgesamt 2.979 Konversationen beteiligt sind. Die Modelle dieser Arbeit werden auf die Erkennung des Alters einer einzelnen Person trainiert. Aus diesem Grund wurden die Chatnachrichten nach Chatteilnehmer aufgeteilt (siehe auch Abschnitt 5.1) und es ergaben sich schließlich 5.958 Dokumente, von denen jeweils die Hälfte das Label *Minderjährige* bzw. *Erwachsene* trägt. Die Aufgabe der Erkennung, ob es sich um eine minderjährige oder erwachsene Person handelt, konnte an diesen Daten nun durchgeführt werden. Die Verteilung der Daten ist nochmal in Tabelle 4.2 unter *Test 2* zu finden.

5 Methoden

Dieses Kapitel beschreibt die Vorgehensweise dieses Projekts und umfasst die Vorverarbeitung der Daten, die verwendeten Features und wie und warum sie extrahiert wurden sowie die Klassifikationsmethoden, mit denen die anschließende Klassifizierung durchgeführt wurde, und die für die Evaluation berechneten Maße.

5.1 Vorverarbeitung

Bei den Blog-Daten wurden bei Vorhandensein mehrerer Texte pro Autor diese zuerst konkateniert und als ein einziges Dokument behandelt. Bei den Grooming-Daten war das Vorgehen ähnlich. Da es sich bei diesem Datensatz um Chatverläufe handelt, wurden pro Chat die Nachrichten jeweils für jeden Chatpartner extrahiert und ebenfalls zu einem einzigen Dokument konkateniert.

Die erhaltenen Dokumente wurden dann bereinigt, teilweise abhängig von ihrem Ursprung. So wurden die PAN13-Daten von HTML-Ausdrücken befreit. Zudem wurden in allen Dokumenten sämtliche Internetadressen und Links durch einen URL-Token ersetzt, da die spezifische URL für weniger relevant, das generelle Vorhandensein jedoch als aussagekräftig erachtet wurde. Weiterhin wurden für die Extraktion der lexikalischen Features die Dokumente in Kleinschreibung umgewandelt.

Durch den Bereinigungsprozess und die anschließende Entfernung von Dokumenten, die nur URL-Tokens enthielten, verringerte sich die Anzahl an Trainingsdaten. Von den anfänglichen 244.840 Trainingssamples der Blog-Daten blieben 244.703 übrig. Die Anzahl beider Testdatensätze blieb unverändert.

5.2 Features

Dieser Abschnitt beschreibt die getesteten (verwendeten) Features und geht dabei sowohl auf die Extraktion als auch die Selektion ein. Im Hinblick auf die finale Cross-Genre-Evaluation wurde versucht neben inhaltlichen Features auch inhaltsunabhängige zu finden, welche bspw. in Verbindung mit POS-Tags generiert wurden oder sich auf stilistische und statistische Merkmale der Dokumente beziehen.

Die verwendeten Features sind nachfolgend nach lexikalischen Features, Second Order Attributes (SOA) und statistischen Features geteilt. Sie wurden in unterschiedlichen Kombinationen an den Blog-Testdatensätzen getestet (also sowohl für drei als auch für zwei Klassen), wobei zuerst die lexikalischen Features in Form der n-Gramme mit den SOA verglichen wurden, um die besseren der beiden mit statistischen Features zu kombinieren für Verbesserungen in der Klassifikation. Diejenigen Features, welche für die besten Ergebnisse sorgten, wurden dann für die Grooming-Testdaten genutzt.

5.2.1 Lexikalische Features

Als lexikalische Features wurden die gängigen Wort- und Zeichen-n-Gramme herangezogen. Auf Wortebene wurden Uni-, Bi- und Trigramme extrahiert, auf Zeichenebene 2- bis 5-Gramme. Dabei wurde die Anzahl etwas eingegrenzt, indem nur diejenigen n-Gramme hinzugezogen wurden, die in mindestens 0,2% und maximal 80% der Trainingsdokumente vorkommen. Die niedrige Untergrenze sollte dazu beitragen, dass Terme, die nur von einem kleinen Teil einer einzigen Klasse verwendet wurden und hilfreich für die Klassifikation sind, nicht direkt verloren gehen und dass aber gleichzeitig Terme, die nur von sehr wenigen Personen insgesamt verwendet wurden, entfernt werden. Die Obergrenze sollte wie eine Art Stoppwortfilter dienen, indem Terme, die in fast allen Dokumenten vorkommen und dadurch keinen nützlichen Einfluss auf die Differenzierung der Klassen haben, nicht in Betracht gezogen werden. Die n-Gramme wurden extrahiert, nachdem alle Großbuchstaben in den Dokumenten in Kleinbuchstaben umgewandelt wurden.

Weiterhin wurden auch die Unigramme von Satzzeichen extrahiert, da diese bei den Zeichen-n-Grammen nicht abgedeckt wurden und angenommen wird, dass eine Nutzung bestimmter Satzzeichen dennoch hilfreich für eine Differenzierung der Altersgruppen sein könnte.

Zusätzlich wurden im Hinblick auf inhaltsunabhängige Features POS-Tag-n-Gramme getestet, wofür 1- bis 4-Gramme extrahiert wurden. Für die POS-Tag-Generierung wurde der POS-Tagger von SpaCy[2] verwendet, der zwei Arten von Tags bereitstellt. Zum Ersten gibt es die universellen POS-Tags, die 17 verschiedene Instanzen umfassen[3], zum Zweiten gibt es detaillierte Tags, von denen über 50 verschiedene für die englische Sprache zur Verfügung stehen. Sie beinhalten sowohl genauere Instanzen für die Einordnung von Worten als auch von Satzzeichen und Symbolen.

Alle n-Gramme wurden mittels TF-IDF gewichtet und pro n-Gramm-Typ in sämtlichen Kombinationen als Features getestet. Dabei wurde zur Feature-Selektion der häufig verwendete χ^2-Berechnung gewählt. Neben der Auswahl aller n-Gramme wurden so auch Versuche mit unterschiedlichen Teilmengen der Features, ausgewählt anhand der höchsten χ^2-Werte, durchgeführt, um jeweils die beste Feature-Untermenge, d.h. die aussagekräftigsten Wort-, Zeichen, Satzzeichen- bzw. POS-Tag-n-Gramme, herauszufiltern.

Alle n-Werte der fünf n-Gramm-Typen (Worte, Zeichen, Satzzeichen, universelle POS-Tags und detaillierte POS-Tags) wurden separat extrahiert. Pro n-Gramm-Typ wurde dann jeder n-Wert einzeln getestet sowie alle Kombinationen von aufeinanderfolgenden n-Werten. So wurden z.B. für die Wort-n-Gramme sechs verschiedene Experimente durchgeführt mit $n = \{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{2, 3\}$ und $\{1, 2, 3\}$. Die Wort-Unigramme dienten gleichzeitig als Baseline.

[2] https://spacy.io/
[3] Eine Liste dieser universellen POS-Tags ist hier zu finden: https://universaldependencies.org/u/pos/

5.2.2 Second Order Attributes

Weiterhin wurden als Alternative für die n-Gramm-basierte Feature-Repräsentation SOA ausprobiert, erstmalig vorgestellt von López-Monroy et al. [52], da diese in mehreren Ansätzen zum AP bereits recht erfolgreich eingesetzt wurden [z.B. 61–63]. Hierbei handelt es sich um eine niedrig-dimensionale und dichte Repräsentation der Dokumenten anhand von Termen, die jeder beliebigen Texteinheit entsprechen können. In dieser Arbeit wurden hierfür vier verschiedene textuelle Einheiten gewählt:

1. Worte
2. universelle POS-Tags
3. detaillierte POS-Tags
4. Worte in Verbindung mit Satz- und Sonderzeichen

Das Vorgehen zum Erhalt der SOA nach López-Monroy et al. [52] lässt sich in zwei Stufen gliedern, welche im Folgenden erläutert werden:

1. Term-Repräsentation (Berechnung von Termvektoren)

Der erste Schritt dient dazu, den Zusammenhang zwischen jedem Term und den verschiedenen Klassen abzubilden. Dafür wurde ein Beziehungswert tp_{ij} berechnet, welcher die Verwendung jedes Terms t_i in jeder Klasse p_j aufzeigt. Die Beziehung zwischen einem Term und einer Klasse berücksichtigt die relative Termfrequenz, wobei nur die Dokumente einbezogen werden, die der Klasse zugehörig sind. Der Gleichung (5.1) folgend wurde für jeden Term t_i im Vokabular des Korpus ein Klassen-Term-Gewicht w_{ij} berechnet durch Aufsummierung der einzelnen Dokument-Term-Gewichte.

$$w_{ij} = \sum_{k:d_k \in P_j} \log_2(1 + \frac{tf_{ik}}{len(d_k)}) \tag{5.1}$$

P_j bezeichnet die Sammlung an Dokumenten, die zu Klasse p_j gehört, tf_{ik} ist die Anzahl an Vorkommen des Terms t_i im Dokument d_k und bei $len(d_k)$ handelt es sich um die Länge des Dokuments d_k.

Anschließend wurde eine Normalisierung der Gewichte vorgenommen, um tp_{ij} zu erhalten, die pro Term t_i diesen als Vektor darstellen: $t_i = \langle tp_{i1}, ..., tp_{in} \rangle$. Aufgrund der Unausgeglichenheit der Daten des verwendeten Korpus wurde die Normalisierung nach Gleichung (5.2) über die Summe der Term-Gewichte in der Klasse vorgenommen, anstatt wie es bei ausbalancierten Daten möglich ist, über die Gewichte des Terms in den anderen Klassen.

$$tp_{ij} = \frac{w_{ij}}{\sum_{i=1}^{TERME} w_{ij}} \tag{5.2}$$

2. Dokument-Repräsentation (Berechnung von Dokumentvektoren)

Nach der Berechnung der Termvektoren wurden aus ihnen Dokumentvektoren gebildet, welche den SOA entsprechen. Sie stellen die Beziehungen zwischen Dokumenten und Klassen dar. Pro Dokument wurden sie aus der Summe derjenigen Termvektoren, deren Terme im Dokument enthalten sind, errechnet, wobei die Termvektoren vorher mit der relativen Termfrequenz von Term t_i im Dokument d_k gewichtet wurden. Die folgende Gleichung (5.3) stellt dieses Vorgehen dar.

$$d_k = \sum_{t_i \in D_k} \frac{tf_{ik}}{len(d_k)} \cdot t_i \tag{5.3}$$

Hierbei steht D_k für die Gesamtheit der Terme, die in Dokument d_k enthalten sind.

Bei diesem Feature hängt der Wert von allen Dokumenten in einer Klasse ab, was dazu führt, dass für dasselbe Dokument in den unterschiedlich zusammengesetzten Datensätzen diese mehrfach berechnet werden mussten.

Die verschiedenen SOA-Gruppen wurden einzeln und zusätzlich in unterschiedlichen Kombinationen getestet. Weiterhin wurde ihre Nützlichkeit in Verbindung mit den statistischen Features untersucht, welche im folgenden Abschnitt 5.2.3 vorgestellt werden.

5.2.3 Statistische Features

Die zusätzlichen statistischen Merkmale wurden teilweise auch im Hinblick auf die Cross-Genre-Evaluation gewählt, da sie im Gegensatz zu Wort-n-Grammen themen- und genreunabhängig sind und stattdessen eher den Schreibstil erfassen. Die im Folgenden dargestellten Features wurden zu sechs Feature-Gruppen zusammengefasst, welche in sämtlichen Kombinationen in den Experimenten getestet wurden.

Part-of-Speech

Die POS-Tag-n-Gramme bieten eine Abbildung von der Satzstruktur der einzelnen Autoren. Aufgrund der Annahme, dass

1. bestimmte Wortarten von Personen bestimmter Altersgruppen häufiger oder seltener verwendet werden als von anderen, da dies im Bereich der Genderbestimmung bereits festgestellt wurde [39], und
2. sich solch ein Verhalten auch in Texten eines anderen Genres widerspiegeln würde,

wurden zusätzlich die Anteile der universellen POS-Tags im Text als Features berechnet.

Wortlänge

Mit der Annahme, dass jüngere Menschen seltener Worte großer Länge und häufiger Abkürzungen (insbesondere informelle) verwenden als ältere Personen, wurde die Wortlänge als weiteres Feature hinzugezogen. Dafür wurden

1. der Anteil von Worten bestimmter Zeichenanzahl n mit Werten für n von 1 bis 9 sowie $n \geq 10$ und
2. die durchschnittliche Wortlänge im Dokument

berechnet und als zwei verschiedene Feature-Gruppen in den Experimenten behandelt. Für diese Features wurden apostrophierte Ausdrücke zu einzelnen Termen ohne Apostroph umgewandelt (z.B. *here's* wird zu *here* und *s*), da dieses Vorgehen auch die gekürzten Worte beachtet und keine unwirklichen Wortlängen erfasst werden, wie es der Fall wäre, wenn die apostrophierten Ausdrücke als ein Wort betrachtet werden.

Großbuchstaben

Es wurde vermutet, dass auch die Verwendung von Großbuchstaben nach Alter variieren könnte (z.B. dass eher jüngere Menschen ganze Worte oder Sätze in Großbuchstaben schreiben). Daraus ergaben sich zwei weitere Features: der Anteil an Worten, die nur in Großbuchstaben geschrieben wurden und der Anteil an Großbuchstaben im Text.

Satzzeichen

Auch Satzzeichen wurden als möglicher Altersgruppenindikator angesehen, unter der Annahme, dass ältere Menschen eher (korrekte) Satzzeichen verwenden. So wurde zum einen der Anteil an Satzzeichen (ausschließlich .,*!?*) berechnet, einmal im Verhältnis zu Satzzeichen und Worten und einmal im Verhältnis zu Satzzeichen und Buchstaben.
Zum anderen wurde auch der Anteil an Satzzeichen-Abfolgen (z.B. *!?*, *!!!!*) berechnet, wobei eine Abfolge eine Mindestlänge von zwei aufweist und der Anteil im Verhältnis zu allen Satzzeichen-Abfolgen inklusive einfach vorkommenden Satzzeichen steht.

Emoticons

Auch für Emoticons bestand die Annahme, dass sie häufiger von jüngeren Menschen genutzt werden, weshalb der Anteil an Emoticons im Text als ein weiteres Feature hinzugezogen wurde. Als Emoticon wurde jeder Term beginnend mit einem *:* oder *;* betrachtet.

5.3 Klassifikation und Evaluation

Im Rahmen dieser Arbeit wurden mit drei Datengruppen (vgl. Kapitel 4, Tabellen 4.1 und 4.2) Klassifikationen durchgeführt:

1. 3-Klassen-Klassifikation mit Blog-Trainingsdaten 1 und Blog-Testdaten 1
2. 2-Klassen-Klassifikation mit Blog-Trainingsdaten 1b und Blog-Testdaten 1b
3. 2-Klassen-Klassifikation mit Blog-Trainingsdaten 1b und Chat-Testdaten 2

Zum Training wurden jeweils Blog-Daten verwendet, abhängig von den Testdaten entweder die Daten, in denen drei verschiedene Klassen gelabelt sind (Datengruppe 1) oder die mit zwei gelabelten Klassen (Datengruppen 2 und 3).

Anhand der ersten beiden Datengruppen wurden die Experimente durchgeführt, um die beste Feature-Zusammensetzung sowohl für eine Mehrklassen- als auch für eine binäre Klassifizierung zu finden. Die dritte Datengruppe wurde zur finalen Evaluation der besten

Modelle genutzt, um festzustellen, ob ein Modell, welches auf die Altersbestimmung eines Autors anhand seiner Blogbeiträgen trainiert wurde, auch zur Erkennung des Alters von Autoren von Chatnachrichten verwendet werden kann und somit im Bereich der Grooming-Detektion hilfreich wäre.

Für die Klassifikationen wurden zwei verschiedene Techniken gewählt, die sich in ihrer Klassifikationsweise sowie dem Zeitaufwand unterscheiden: der RF-Algorithmus und eine SVM. Für die SVM wurde sich aufgrund bereits erfolgreicher Anwendung bei AP-Tasks [58, 62, 111] entschieden, RF wurde aufgrund seiner Effizienz und da er ebenfalls bereits erfolgreich beim AP angewendet wurde [55, 105], hinzugezogen.

Für das Baseline-Modell und sämtliche Modelle, die n-Gramme als Features nutzten, wurde der schnellere RF-Klassifikator verwendet, ebenso für alle weiteren Experimente. Zusätzlich wurde die SVM[4] mit der Mehrklassen-Strategie one-versus-one als zweite Klassifikationstechnik hinzugezogen für Experimente mit geringerer Feature-Anzahl, d.h. diejenigen, bei denen die SOA als Features verwendet wurden. Zur Bestimmung des C-Wertes für die SVM pro SOA-Typ wurde eine Grid-Search durchgeführt mit einer 3-fold Cross-Validation.

Die berechneten Maße zur Evaluation umfassen Precision, Recall und das F1-Maß pro Klasse, sowie den durchschnittlichen macro-F1-Wert, welcher als Hauptkennzahl betrachtet wurde. Um eine gewisse Vergleichbarkeit mit den Ansätzen, vorgestellt in Kapitel 3 (insbesondere im Abschnitt 3.4), zu schaffen, wurde auch die Accuracy berechnet, obwohl sie für unausgeglichene Daten wie die Blog-Daten weniger geeignet ist. Für den ausbalancierten Cross-Genre-Testkorpus hingegen gibt die Accuracy einen besseren Überblick über die Güte der Klassifikation.

[4] Verwendet wurde die SVC-Implementierung von *scikit-learn* (https://scikit-learn.org).

6 Ergebnisse und Diskussion

In diesem Kapitel sind die Ergebnisse dieser Arbeit zusammengestellt. Sie sind geordnet nach der Art der genutzten Features und nach Mehrklassen- bzw. binärer Klassifikation unterteilt. Der letzte Abschnitt betrachtet die Ergebnisse auf die Cross-Genre-Daten. Hauptsächlich wird das macro-F1-Maß zum Vergleich der unterschiedlichen Versuche betrachtet, da es Recall und Precision aller Klassen vereint.

6.1 Ergebnisse der n-Gramm-Experimente

Dieser Abschnitt stellt die Ergebnisse zu den Experimenten mit den verschiedenen n-Grammen dar, unterteilt nach 3-Klassen- und 2-Klassen-Klassifikation. Diese Klassifikationen wurden mit RF vorgenommen.

6.1.1 3-Klassen-Klassifikation mit n-Grammen

Für die extrahierten n-Gramme der Worte, Zeichen, universellen (POS1) und detaillierten POS-Tags (POS2) sowie der Satzzeichen sind in Tabelle 6.1 für alle getesteten einzelnen und kombinierten n-Gramme die macro-F1-Werte dargestellt. Pro n-Gramm-Typ/n-Wert-Kombination wurden neben dem gesamten Feature-Set auch verschiedene Teilmengen, ausgewählt mittels χ^2-Feature-Selektion, für die Klassifikation verwendet. Die Tabelle enthält jeweils nur das beste erzielte Ergebnis pro n-Gramm-Typ und Wert für n, welches in den meisten Fällen mit einer reduzierten Feature-Menge erzielt wurde.

Die Ergebnisse des macro-F1-Wertes bewegen sich in einem engen Bereich zwischen 0,3228 für die Zeichen-Bigramme und 0,3311 für die Wort-Trigramme. Dabei wurde die Baseline von 0,3253 der Wort-Unigramme von den meisten anderen Versuchen übertroffen oder zumindest erreicht.

Für die Wort-n-Gramme wird deutlich, dass jedes Feature-Set bessere Ergebnisse lieferte als die Wort-Unigramme, was zeigt, dass mehr differenzierende Aspekte in der syntaktischen Verwendung von Worten liegen als in der bloßen relativen Häufigkeit der einzelnen genutzten Worte.

Die beiden POS-Tag-Features betreffend fällt auf, dass die POS2-Tags grundsätzlich bessere macro-F1-Werte erzielten als die POS1-Tags. Weiterhin ist zu erkennen, dass die besseren Ergebnisse eher mit einzelnen und zweifachen n-Grammen erzielt wurden als mit drei- oder vierfachen, welche die Baseline teilweise nicht erreichten.

Fast alle Versuche mit Zeichen-n-Grammen erreichten einen macro-F1-Wert von über 0,327. Eine der Ausnahmen bildeten hierbei die Zeichen-Bigramme mit dem insgesamt schlechtesten Wert. Dies kann sich dadurch erklären lassen, dass diese n-Gramme als einzelnes

Tabelle 6.1: Übersicht über die macro-F1-Werte, die mit den verschiedenen n-Gramm-Typen und Werten für n bei der 3-Klassen-Klassifikation mit RF erreicht wurden. Der Baseline-Wert ist kursiv gekennzeichnet. Der jeweils höchste Wert pro n-Gramm-Typ ist fett hervorgehoben.

$n =$	**Worte**	**Zeichen**	**POS1**	**POS2**	**Satzzeichen**
$\{1\}$	*0,3253*		0,3274	0,3277	**0,3271**
$\{2\}$	0,3273	0,3228	**0,3275**	**0,3290**	
$\{3\}$	**0,3311**	0,3273	0,3247	0,3271	
$\{4\}$		0,3272	0,3261	0,3276	
$\{5\}$		0,3279			
$\{1,2\}$	0,3265		0,3257	0,3268	
$\{2,3\}$	0,3289	0,3253	0,3257	0,3271	
$\{3,4\}$		0,3276	0,3260	0,3270	
$\{4,5\}$		0,3272			
$\{1,2,3\}$	0,3269		0,3250	0,3260	
$\{2,3,4\}$		0,3265	0,3246	0,3249	
$\{3,4,5\}$		0,3275			
$\{1,2,3,4\}$			0,3253	0,3256	
$\{2,3,4,5\}$		**0,3289**			

Feature-Set zu wenige Klassen abgrenzende Informationen enthalten, wie es bei einzelnen Zeichen auch angenommen wird. Die Verteilung der Buchstaben könnte viel eher einen Hinweis auf die Dokumentensprache geben als auf die Altersgruppe des Autors, was für diesen Task aber nicht nützlich erscheint und außerdem alle Texte in Englisch verfasst sind.

Anders war jedoch die Annahme betreffend der Satzzeichen, die sich in den Ergebnissen auch bestätigte. Den Ergebnissen zufolge kann durchaus davon ausgegangen werden, dass die Satzzeichen-Nutzung in Zusammenhang mit dem Alter des Autors steht und ein gutes Feature darstellt.

Neben dem insgesamt besten Ergebnis der Wort-Trigramme, gehörten auch die POS2-Bigramme mit einem macro-F1-Wert von 0,329 sowie die Zeichen-{2,3,4,5}- und Wort-{2,3}-Gramme mit einem Wert von jeweils 0,3289 zu den besseren Ergebnissen. In Tabelle 6.2 sind jeweils die Kombinationen pro n-Gramm-Typ mit den höchsten erreichten macro-F1-Werten dargestellt.

Die Resultate lassen darauf schließen, dass POS-Tags, insbesondere die detaillierten POS2-Tags, durchaus eine Wort- oder Zeichen-n-Gramm-Repräsentation ersetzen können und damit eine Feature-Variante bieten, welche weniger vom konkreten Inhalt abhängt und dadurch möglicherweise geeigneter bei Cross-Domain-Problemen (insbesondere Cross-Topic) ist. Zudem weist die Textrepräsentation mittels POS-Tags eine geringere Dimensionalität auf, da es weniger einzigartige POS-Tags als Worte oder Zeichen gibt.

Für Recall und Precision und somit auch für den F1-Score ergaben sich pro Klasse Werte, die in Korrelation mit dem Anteil der Klasse im Datensatz stehen. So wies die unterrepräsentierte Klasse der *10s* deutlich geringere Werte auf als die anderen beiden Klassen und zeigte die niedrigste Erkennungsrate von maximal 0,0276 mit den Wort-Trigrammen.

Tabelle 6.2: Übersicht über die n-Gramm-Auswahl pro Typ mit dem höchsten macro-F1-Wert bei drei Klassen. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes.

	Worte	**Zeichen**	**POS1**	**POS2**	**Satzzeichen**
$n =$	$\{3\}$	$\{2,3,4,5\}$	$\{2\}$	$\{2\}$	$\{1\}$
mit χ^2 ausgewählter Anteil	0,8	0,8	0,5	0,5	1,0
macro-F1	**0,3311**	0,3289	0,3275	0,3290	0,3271
Accuracy	0,4683	0,4684	0,4681	**0,4693**	0,4592
Recall					
- *10s*	**0,0276**	0,0191	0,0175	0,0186	0,0253
- *20s*	0,4350	0,4604	0,4564	0,4579	**0,4668**
- *30s*	**0,5437**	0,5288	0,5310	0,5320	0,5079
Precision					
- *10s*	0,0829	0,0854	0,0854	**0,0868**	0,0690
- *20s*	0,3643	0,3651	0,3646	**0,3667**	0,3638
- *30s*	0,5674	**0,5693**	0,5665	0,5673	0,5658
F1-Score					
- *10s*	**0,0414**	0,0313	0,0290	0,0306	0,0371
- *20s*	0,3965	0,4072	0,4054	0,4072	**0,4089**
- *30s*	**0,5553**	0,5288	0,5482	0,5491	0,5352

Zudem ist erkennbar, dass der Recall im Vergleich zur Precision pro Klasse über die Feature-Sets hinweg stärker variierte. Das zeigt, dass der Anteil an korrekt klassifizierten Datenpunkten unter den Vorhersagen zwar recht stabil war, die korrekte Detektion der einzelnen Altersgruppen jedoch mehr von den gewählten Features abhing. Der höchste F1-Wert für die *10s* und *30s* wurde mit den Worten erreicht, für die *20s* mit den Satzzeichen.

6.1.2 2-Klassen-Klassifikation mit n-Grammen

Die Ergebnis-Darstellung der n-Gramm-Versuche für die binäre Klassifikation in Tabelle 6.3 wurde genauso gehandhabt wie für die 3-Klassen-Klassifikation. Auch hier wurden in den meisten Fällen die besten Resultate mit einer Teilmenge der Features erreicht.

Die Baseline der Wort-Unigramme lag bei einem macro-F1-Wert von 0,4913 und wurde von den anderen Versuchen nur in fünf Fällen übertroffen. Die erzielten macro-F1-Werte bewegten sich in einem Bereich von 0,4807 für die Zeichen-{3,4}-Gramme bis 0,4934 für die Wort-{1,2,3}-Gramme.

Auffallend ist, dass die Wort- und Satzzeichen-n-Gramme insgesamt die besten Werte erzielten und die Zeichen-n-Gramme am schlechtesten abschnitten. Das lässt darauf schließen, dass die bloße Verwendung von Worten, die syntaktische Anordnung von ihnen sowie die Nutzung von Satzzeichen, sich bei Minderjährigen und Erwachsenen unterscheidet und mehr zur Differenzierung beiträgt als die Anordnung von Buchstaben und anderen Zeichen.

Tabelle 6.3: Übersicht über die macro-F1-Werte, die mit den verschiedenen n-Gramm-Typen und Werten für n bei der 2-Klassen-Klassifikation mit RF erreicht wurden. Der Baseline-Wert ist kursiv gekennzeichnet. Der jeweils höchste Wert pro n-Gramm-Typ ist fett hervorgehoben.

$n =$	**Worte**	**Zeichen**	**POS1**	**POS2**	**Satzzeichen**
$\{1\}$	*0,4913*		**0,4914**	**0,4907**	**0,4933**
$\{2\}$	0,4884	0,4862	0,4880	0,4880	
$\{3\}$	0,4932	0,4861	0,4862	0,4879	
$\{4\}$		0,4874	0,4873	0,4856	
$\{5\}$		0,4878			
$\{1,2\}$	0,4930		0,4867	0,4873	
$\{2,3\}$	0,4904	0,4868	0,4878	0,4877	
$\{3,4\}$		0,4807	0,4873	0,4878	
$\{4,5\}$		0,4859			
$\{1,2,3\}$	**0,4934**		0,4867	0,4873	
$\{2,3,4\}$		0,4878	0,4860	0,4871	
$\{3,4,5\}$		**0,4878**			
$\{1,2,3,4\}$			0,4867	0,4872	
$\{2,3,4,5\}$		0,4865			

Die beiden POS-Tag-Feature-Sets unterschieden sich in ihren Ergebnissen weniger stark. Zwischen den beiden wird der höchste macro-F1-Wert von 0,4914 von den POS1-Unigrammen erzielt, insgesamt lagen die Werte der POS2-n-Gramme aber meist etwas über denen der POS1-n-Gramme.

Die ausführlicheren Ergebnisse der n-Gramm-Sets pro Feature-Gruppe mit dem jeweils höchsten macro-F1-Wert sind in Tabelle 6.4 dargestellt.

Tabelle 6.4: Übersicht über die n-Gramm-Auswahl pro Typ mit dem höchsten macro-F1-Wert bei binärer Klassifikation. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes.

	Worte	**Zeichen**	**POS1**	**POS2**	**Satzzeichen**
$n =$	$\{1,2,3\}$	$\{3,4,5\}$	$\{1\}$	$\{1\}$	$\{1\}$
mit χ^2 ausgewählter Anteil	0,5	1,0	0,8	0,5	1,0 und 0,65
macro-F1	**0,4934**	0,4878	0,4914	0,4907	0,4933
Accuracy	0,9179	**0,9253**	0,9195	0,9201	0,9119
Recall					
- *Minderjährige*	0,0180	0,0079	0,0146	0,0135	**0,0214**
- *Erwachsene*	0,9854	**0,9941**	0,9874	0,9881	0,9787
Precision					
- *Minderjährige*	0,0849	**0,0915**	0,0802	0,0787	0,0701
- *Erwachsene*	**0,9304**	0,9303	0,9303	0,9303	0,9302
F1-Score					
- *Minderjährige*	0,0297	0,0145	0,0248	0,0231	**0,0328**
- *Erwachsene*	0,9571	**0,9612**	0,9580	0,9583	0,9538

Es wird deutlich, dass die korrekte Erkennung der Klasse *Minderjährige* viel schlechter gelang als die der Klasse *Erwachsene*. Das kann in erster Linie mit der starken Datenunausgeglichenheit zusammenhängen. Der höchste Recall sowie F1-Score für Kinder lag bei 0,0214 bzw.

0,0328 und wurde mit den Satzzeichen-Unigrammen erzielt. Gleichzeitig zeigt diese Feature-Gruppe aber auch den geringsten Recall und F1-Score für die Klasse *Erwachsene* sowie die niedrigste Precision für beide Klassen. Durch dieses Feature werden also im Vergleich zu den anderen mehr Datenpunkte, die Erwachsene repräsentieren, als Kinder klassifiziert.

Die Zeichen-n-Gramme weisen zwar den geringsten Recall für Minderjährige auf, aber dafür die höchste Precision. Die Wahrscheinlichkeit, dass ein Datenpunkt, klassifiziert als Kind, auch tatsächlich ein Kind repräsentiert, ist für diese Feature-Gruppe also am höchsten. Zudem ergaben die Zeichen-n-Gramme den höchsten F1-Score für *Erwachsene*, was auch daran lag, dass insgesamt die meisten Datenpunkte als diese überrepräsentierte Klasse klassifiziert wurden.

6.2 Ergebnisse der SOA-Experimente

In diesem Abschnitt werden die Ergebnisse der Experimente mit den SOA-Features (siehe Abschnitt 5.2.2) dargestellt und besprochen. Die Klassifikationen mit den SOA wurden sowohl mit RF als auch mit einer SVM durchgeführt. Zu Beginn wurden die einzelnen SOA-Typen allein getestet, zusätzlich in verschiedenen Kombinationen und schließlich in Verbindung mit den statistischen Features (siehe Abschnitt 5.2.3). Die vier untersuchten SOA-Typen umfassen Worte, Worte inklusive Satzzeichen (bezeichnet als Worte+), universelle (POS1) und detaillierte POS-Tags (POS2).

6.2.1 3-Klassen-Klassifikation mit SOA

In den nachfolgenden Tabellen sind die Ergebnisse der Mehrklassen-Klassifikation mit den vier unterschiedlichen SOA-Typen unterschieden nach Klassifikator zusammengefasst. Es fällt auf, dass es sowohl Unterschiede zwischen den verwendeten Features als auch den zwei Klassifikationstechniken gab.

Bei den Experimenten mit RF, dargestellt in Tabelle 6.5, schnitten deutlich die Ergebnisse erzielt mit den POS2-SOA am besten ab. Sie zeigten den höchsten macro-F1-Wert von 0,3376 und wiesen mit einer Ausnahme auch die höchsten Precision-Werte pro Klasse auf. Für *10s* und *30s* ergaben sich damit zudem die höchsten Recall- und F1-Werte.

Bei den Versuchen mit der SVM, deren Ergebnisse in Tabelle 6.6 zusammengefasst sind, wird der höchste macro-F1-Wert von 0,3299 mit den Wort-SOA erreicht. Die besten F1-Scores verteilen sich über die SOA-Typen, wobei die Klasse der *20s* insgesamt am häufigsten mit den Wort-SOA, die Klasse der *30s* mit den POS1-SOA und die *10s* mit den POS2-SOA richtig erkannt wurde. Auffallend ist zudem die recht große Spanne des Recalls der *10s* und *20s* zwischen den verschiedenen SOA-Typen.

Tabelle 6.5: Ergebnisse der 3-Klassen-Klassifikation mit den verschiedenen SOA-Features und RF. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes.

	Worte	**Worte+**	**POS1**	**POS2**
macro-F1	0,3313	0,3324	0,3317	**0,3376**
Accuracy	0,4627	0,4557	0,4606	**0,4668**
Recall				
- *10s*	0,0439	0,0574	0,0518	**0,0580**
- *20s*	0,3990	**0,4052**	0,3879	0,3890
- *30s*	0,5548	0,5368	0,5572	**0,5667**
Precision				
- *10s*	0,0743	0,0731	0,0657	**0,0752**
- *20s*	0,3618	0,3615	0,3638	**0,3677**
- *30s*	0,5635	0,5655	0,5665	**0,5717**
F1-Score				
- *10s*	0,0552	0,0643	0,0579	**0,0655**
- *20s*	0,3795	**0,3821**	0,3755	0,3781
- *30s*	0,5591	0,5508	0,5618	**0,5692**

Tabelle 6.6: Ergebnisse der 3-Klassen-Klassifikation mit den verschiedenen SOA-Features und SVM. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes.

	Worte	**Worte+**	**POS1**	**POS2**
macro-F1	**0,3299**	0,3267	0,3214	0,3175
Accuracy	**0,4285**	0,4201	0,4109	0,4023
Recall				
- *10s*	0,1030	0,1357	0,1684	**0,1920**
- *20s*	**0,4303**	0,3729	0,3203	0,3003
- *30s*	0,4673	0,4852	**0,4985**	0,4932
Precision				
- *10s*	**0,0704**	0,0692	0,0659	0,0670
- *20s*	**0,3660**	0,3603	0,3620	0,3599
- *30s*	0,5623	0,5645	**0,5651**	0,5629
F1-Score				
- *10s*	0,0837	0,0916	0,0947	**0,0993**
- *20s*	**0,3956**	0,3665	0,3399	0,3275
- *30s*	0,5105	0,5219	**0,5297**	0,5258

Kombinationen von SOA-Features

Mit dem Gedanken die besten F1-Scores in einem Modell zu vereinen (besonders in Bezug auf die Ergebnisse der SVM) wurden weiterhin Experimente durchgeführt, bei denen die einzelnen SOA-Typen in unterschiedlichen Kombinationen getestet wurden. Dabei wurde auf die Verknüpfung der Worte- und Worte+-SOA verzichtet, da die Worte+-Features lediglich eine Erweiterung der Worte-SOA sind und angenommen wurde, dass sich keine neuen Erkenntnisse aus ihrer Verbindung ergeben.

Die Ergebnisse für die Kombinationen der SOA-Typen klassifiziert mit RF sind in Tabelle 6.7 dargestellt. Den höchsten macro-F1-Wert von 0,3351 erreichte die Verknüpfung der POS1- und POS2-SOA. Entgegen der Erwartungen wurde aber mit keiner Kombination der macro-F1-Wert der POS2-SOA von 0,3376 (siehe Tabelle 6.5) übertroffen.

Die F1-Scores für die Klassen *20s* und *30s* konnten jeweils mit unterschiedlichen SOA-Zusammensetzungen etwas erhöht werden, aber nicht mit einem einzigen Feature-Set. So wurde mit der Verknüpfung von Wort- und POS2-SOA der höchste F1-Score von 0,3942 für die *20s* erreicht und mit der POS1-POS2-Verbindung für die *30s* mit 0,5749. Der F1-Score für die *10s*-Klasse hat sich jedoch mit allen Kombinationen im Vergleich zu den einzelnen SOA-Features verschlechtert.

Tabelle 6.7: Ergebnisse der 3-Klassen-Klassifikation mit Kombinationen der SOA-Features mit RF. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes.

	W, P1	**W, P2**	**W+, P1**	**W+, P2**	**W, P1, P2**	**W+, P1, P2**	**P1, P2**
macro-F1	0,3266	0,3322	0,3315	0,3332	0,3304	0,3312	**0,3351**
Accuracy	0,4633	0,4692	0,4654	0,4704	0,4703	0,4693	**0,4742**
Recall							
- *10s*	0,0264	0,0310	**0,0417**	0,0372	0,0259	0,0343	0,0377
- *20s*	0,4211	**0,4232**	0,3945	0,3989	0,4204	0,3965	0,3943
- *30s*	0,5439	0,5524	0,5626	0,5693	0,5568	0,5693	**0,5788**
Precision							
- *10s*	0,0614	0,0663	**0,0694**	0,0674	0,0621	0,0643	0,0681
- *20s*	0,3621	0,3689	0,3632	0,3673	0,3676	0,3640	**0,3704**
- *30s*	0,5630	0,5685	0,5660	0,5690	0,5681	0,5691	**0,5710**
F1-Score							
- *10s*	0,0370	0,0422	**0,0521**	0,0476	0,0366	0,0448	0,0486
- *20s*	0,3894	**0,3942**	0,3782	0,3825	0,3922	0,3795	0,3820
- *30s*	0,5533	0,5603	0,5643	0,5692	0,5624	0,5692	**0,5749**

Legende: W = Worte, W+ = Worte und Satzzeichen, P1 = universelle POS-Tags, P2 = detaillierte POS-Tags

Die Ergebnisse von der Klassifikation mit SVM sind in Tabelle 6.8 zu finden. Wie auch bei der Klassifikation mit RF erreichte die Kombination von POS1- und POS2-SOA den höchsten macro-F1-Wert. Mit 0,3215 lag dieser jedoch unter dem besten Wert der einzelnen SOA von 0,3299 für die Wort-SOA (siehe Tabelle 6.6).

Die besten F1-Scores für die Klassen *10s* und *30s* vereinten sich in der Verbindung von Wort- und POS1-SOA und stellten eine Erhöhung dar. Die Klassifizierung der *20s* dagegen verschlechterte sich, mit einem F1-Höchstwert von 0,3478 im Vergleich zu 0,3956.

SOA mit statistischen Features

Weiterhin wurden Experimente durchgeführt, bei denen die SOA-Typen in Verbindung mit unterschiedlichen Kombinationen der statistischen Features getestet wurden, um die Resultate der Verwendung einzelner SOA-Typen zu verbessern. Die statistischen Features wurden

Tabelle 6.8: Ergebnisse der 3-Klassen-Klassifikation mit Kombinationen der SOA-Features mit SVM. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes.

	W, P1	**W, P2**	**W+, P1**	**W+, P2**	**W, P1, P2**	**W+, P1, P2**	**P1, P2**
macro-F1	0,3082	0,3185	0,3109	0,3169	0,3154	0,3184	**0,3215**
Accuracy	0,3959	**0,4085**	0,3991	0,4079	0,4059	0,4082	0,4078
Recall							
- *10s*	**0,2337**	0,1881	0,2241	0,1757	0,1886	0,1937	0,1700
- *20s*	0,2269	0,2865	0,2390	0,2899	0,2738	0,2807	**0,3364**
- *30s*	**0,5237**	0,5134	0,5227	0,5118	0,5169	0,5159	0,4826
Precision							
- *10s*	0,0680	0,0681	0,0683	0,0647	0,0655	**0,0687**	0,0677
- *20s*	0,3545	0,3576	0,3554	0,3546	0,3557	0,3584	**0,3601**
- *30s*	0,5630	0,5641	0,5630	**0,5651**	0,5644	0,5640	0,5636
F1-Score							
- *10s*	**0,1054**	0,0999	0,1047	0,0946	0,0972	0,1015	0,0968
- *20s*	0,2767	0,3181	0,2858	0,3190	0,3094	0,3148	**0,3478**
- *30s*	**0,5426**	0,5375	0,5421	0,5371	0,5396	0,5389	0,5199

Legende: W = Worte, W+ = Worte und Satzzeichen, P1 = universelle POS-Tags, P2 = detaillierte POS-Tags

in den thematischen Gruppen zusammengefasst, wie sie in Abschnitt 5.2.3 beschrieben sind. Das ergab 63 mögliche Kombinationen, von denen pro SOA-Typ jeweils die fünf mit dem höchsten macro-F1-Wert in den beiden Ergebnistabellen zusammengefasst sind, um eine mögliche Tendenz der am meisten differenzierenden Features zu erkennen.

Tabelle 6.9: Dargestellt sind die fünf höchsten erreichten macro-F1-Werte für jeden SOA-Typen und die entsprechenden Features, die zusätzlich zu den SOA-Features verwendet wurden, bei der 3-Klassen-Klassifikation mit RF. Der insgesamt höchste Wert ist fett hervorgehobenen.

Worte	Features	E	W2, S	W2	G, S, E	W1
	macro-F1	0,3332	0,3315	0,3311	0,3311	0,3307
Worte+	Features	S	S, E	W2	G, S	G
	macro-F1	0,3344	0,3334	0,3329	0,332	0,3315
POS1	Features	S	G	S, E	P, W1, E	W1, W2, G, S
	macro-F1	0,3312	0,3306	0,3302	0,3297	0,3297
POS2	Features	E	G, S	W2, S	G	S
	macro-F1	**0,3361**	0,3337	0,3331	0,3331	0,3324

Legende: P = Anteile der POS-Tags, W1 = Anteile von Worten bestimmter Zeichenlänge, W2 = durchschnittliche Wortlänge, G = Großbuchstaben-Features, S = Satzzeichen-Features, E = Anteil an Emoticons

Die jeweils höchsten macro-F1-Werte bei der Klassifikation mit RF (siehe Tabelle 6.9) wurden mit einem einzelnen zusätzlichen Feature-Paket erreicht, und zwar einerseits den Emoticons für die Worte- und POS2-SOA und andererseits den Satzzeichen-Features für die Worte+- und POS1-SOA.

Die zusätzlichen statistischen Features brachten im Vergleich zu den einzelnen Worte- und Worte+-SOA eine leichte Verbesserung des macro-F1-Wertes von rund 0,002. Die Werte der beiden POS-SOA wurden nicht übertroffen durch die Hinzunahme weiterer Features. Außerdem überstieg keine Kombination den anfänglichen Bestwert von 0,3376 der einzelnen POS2-SOA. Besonders die Satzzeichen-Features sind in diesen besten Kombinationen häufig vertreten, aber auch die Großbuchstaben- und Emoticon-Features. Seltener kamen die Anteile der POS-Tags sowie die durchschnittliche Wortlänge zum Einsatz.

Auch die Klassifikation mittels SVM wurde für die SOA in Verbindung mit den statistischen Features getestet. Die Ergebnisse dafür mit den höchsten macro-F1-Werten sind in Tabelle 6.10 dargestellt. Es ist auffallend, dass die Maximal-Werte durch Hinzunahme von mindestens drei weiteren Feature-Sets erreicht wurden. Den höchsten macro-F1-Wert von 0,3334 erzielte dabei die Kombination aus Wort-SOA, Anteilen von Worten bestimmter Zeichenlänge sowie Großbuchstaben-, Satzzeichen- und Emoticon-Features. Diese vier statistischen Feature-Gruppen waren in unterschiedlichen Konstellationen für all diese besten Ergebnisse verantwortlich, wohingegen die Anteile der POS-Tags sowie die durchschnittliche Wortlänge nicht zum Einsatz kamen.

Tabelle 6.10: Dargestellt sind die fünf höchsten erreichten macro-F1-Werte für jeden SOA-Typen und die entsprechenden Features, die zusätzlich zu den SOA-Features verwendet wurden, bei der 3-Klassen-Klassifikation mit SVM. Der insgesamt höchste Wert ist fett hervorgehobenen.

Worte	Features	W1, G, S, E	G, S, E	G, S	W1, G, S	W1, S, E
	macro-F1	**0,3334**	0,3325	0,3323	0,3319	0,3317
Worte+	Features	W1, G, S, E	G, S, E	G, S	W1, G, S	W1, S, E
	macro-F1	0,3332	0,3327	0,3326	0,3323	0,3311
POS1	Features	G, S, E	W1, S, E	G, S	W1, G, S, E	W1, S
	macro-F1	0,3333	0,3322	0,3320	0,3314	0,3313
POS2	Features	G, S, E	G, S	W1, S, E	W1, S	W1, G, S, E
	macro-F1	0,3327	0,3322	0,3321	0,3315	0,3312

Legende: P = Anteile der POS-Tags, W1 = Anteile von Worten bestimmter Zeichenlänge, W2 = durchschnittliche Wortlänge, G = Großbuchstaben-Features, S = Satzzeichen-Features, E = Anteil an Emoticons

6.2.2 2-Klassen-Klassifikation mit SOA

Nach der Mehrklassen- wurde auch die binäre Klassifikation mit den SOA-Features untersucht. Die Ergebnisse der einzelnen SOA-Typen sind in den Tabellen 6.11 und 6.12 zusammengefasst, geteilt nach Klassifikationstechnik.

Die Ergebnisse der RF-Klassifikation lagen sehr nah beieinander, der höchste macro-F1-Wert von 0,5 wurde mit den Worte-SOA erzielt. Diese wiesen auch die höchste Precision pro Klasse auf. Die meisten korrekten Klassifikationen wurden zwar bei der Verwendung der POS1-SOA gemacht und dieser Versuch wies den höchsten Recall für die Klasse *Erwachsene* auf, jedoch gleichzeitig den niedrigsten für die Klasse *Minderjährige*. Der höchste F1-Score für die Klasse *Minderjährige* wurde mit den Worte+-SOA erreicht, für die Klasse *Erwachsene* mit

Tabelle 6.11: Ergebnisse der 2-Klassen-Klassifikation mit den verschiedenen SOA-Features und RF. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes.

	Worte	**Worte+**	**POS1**	**POS2**
macro-F1	**0,5000**	0,4985	0,4945	0,4996
Accuracy	0,8899	0,8830	**0,8925**	0,8909
Recall				
- *Minderjährige*	0,0490	**0,0529**	0,0372	0,0473
- *Erwachsene*	0,9531	0,9453	**0,9566**	0,9542
Precision				
- *Minderjährige*	**0,0726**	0,0677	0,0604	0,0719
- *Erwachsene*	**0,9303**	0,9301	0,9298	**0,9303**
F1-Score				
- *Minderjährige*	0,0585	**0,0594**	0,0460	0,0571
- *Erwachsene*	0,9416	0,9376	**0,9430**	0,9421

den POS1-SOA, wobei diese Features auch den niedrigsten F1-Score der jeweils anderen Klasse aufwiesen. Das zeigt, dass eine Verbesserung bei der Erkennung einer Klasse mit einer Verschlechterung der Erkennung der anderen Klasse einherging.

Tabelle 6.12: Ergebnisse der 2-Klassen-Klassifikation mit den verschiedenen SOA-Features und SVM. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes.

	Worte	**Worte+**	**POS1**	**POS2**
macro-F1	**0,4826**	0,4722	0,4721	0,4722
Accuracy	**0,7695**	0,7296	0,7307	0,7286
Recall				
- *Minderjährige*	0,1779	0,2241	0,2207	**0,2264**
- *Erwachsene*	**0,8139**	0,7675	0,7689	0,7663
Precision				
- *Minderjährige*	0,0669	0,0675	0,0669	**0,0678**
- *Erwachsene*	0,9295	0,9295	0,9293	**0,9296**
F1-Score				
- *Minderjährige*	0,0973	0,1037	0,1027	**0,1043**
- *Erwachsene*	**0,8679**	0,8409	0,8416	0,8401

Bei der Klassifikation mit der SVM, deren durchschnittliche Ergebnisse unter denen der RF-Klassifikation lagen, wurde der höchste macro-F1-Wert von 0,4826 ebenfalls mit den Worte-SOA erzielt, wobei im Vergleich mit den anderen SOA-Typen die meisten Erwachsenen aber die wenigsten Minderjährigen korrekt klassifiziert wurden. Diese Recall-Werte setzten sich zudem stärker von den anderen, die sehr nah beieinander lagen, ab. Die meisten Minderjährigen wurden mit den POS2-SOA korrekt erkannt, die auch den höchsten F1-Score für diese Klasse erzielten. Dagegen wurde der höchste F1-Score für die Erwachsenen mit den Worte-SOA erreicht.

Kombinationen von SOA-Features

Auch für die binäre Klassifikation wurden in der Hoffnung auf die Vereinigung der besten F1-Scores in einem Modell die Kombinationen von SOA-Typen getestet. Die Ergebnisse, die dabei mit dem RF-Klassifikator erzielt wurden, sind in Tabelle 6.13 dargestellt, die mit der SVM erzielten Ergebnisse in Tabelle 6.14.

Tabelle 6.13: Ergebnisse der 2-Klassen-Klassifikation mit Kombinationen der SOA-Features mit RF. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes.

	W, P1	**W, P2**	**W+, P1**	**W+, P2**	**W, P1, P2**	**W+, P1, P2**	**P1, P2**
macro-F1	0,4932	0,4937	0,4903	**0,4971**	0,4954	0,4930	0,4947
Accuracy	0,9072	0,9041	0,9097	0,9064	0,9094	**0,9102**	0,9070
Recall							
- *Minderjährige*	0,0242	0,0270	0,0186	**0,0304**	0,0259	0,0220	0,0265
- *Erwachsene*	0,9734	0,9699	0,9765	0,9722	0,9757	**0,9769**	0,9731
Precision							
- *Minderjährige*	0,0640	0,0632	0,0561	**0,0757**	0,0742	0,0666	0,0687
- *Erwachsene*	0,9300	0,9300	0,9299	**0,9304**	0,9303	0,9301	0,9302
F1-Score							
- *Minderjährige*	0,0351	0,0379	0,0279	**0,0434**	0,0384	0,0330	0,0382
- *Erwachsene*	0,9512	0,9495	0,9526	0,9508	0,9525	**0,9529**	0,9511

Legende: W = Worte, W+ = Worte und Satzzeichen, P1 = universelle POS-Tags, P2 = detaillierte POS-Tags

Mit RF liegen die macro-F1-Werte in einem Bereich von 0,4903 bis 0,4971, wobei der höchste Wert mit der Verknüpfung von Worte+- und POS2-SOA erreicht wurde. Diese Kombination stellte auch den höchsten F1-Score und Recall für die Klasse *Minderjährige* sowie die höchste Precision für beide Klassen. Eine fast identische Precision von jeweils rund 0,930 auf die Klasse *Erwachsene* wurde von allen Kombinationen erzielt. Im Vergleich zu den Ergebnissen der einzelnen SOA (siehe Tabelle 6.11) wurden die macro-F1-Werte außer der für die POS1-SOA nicht überschritten und es ist keine Verbesserung im F1-Score der Klasse *Minderjährige* erreicht worden.

Bei der SVM-Klassifikation wurden insgesamt niedrigere macro-F1-Werte erreicht im Vergleich zu der RF-Klassifikation. Es wurden zwar mehr Minderjährige richtig klassifiziert, was auch zu einem höheren Recall und F1-Score dieser Klasse bei den meisten Feature-Kombinationen führte. Gleichzeitig wurden jedoch weniger Erwachsene korrekt eingeordnet, was sich in den oftmals geringeren Recall- und F1-Scores widerspiegelte. Insgesamt konnte der macro-F1-Höchstwert der einzeln genutzten SOA (siehe Tabelle 6.12) nicht erhöht werden.

SOA mit statistischen Features

Die besten Ergebnissen der Experimente mit Kombinationen von SOA-Typen und weiteren statistischen Features, die in den Tabellen 6.15 und 6.16 zusammengefasst sind, zeigen, dass der RF-Klassifikator die besten Resultate in Verbindung mit ein bis zwei weiteren statistischen Feature-Gruppen erreichte. Dabei waren unter dieser Auswahl an besten Ergebnissen vor

Tabelle 6.14: Ergebnisse der 2-Klassen-Klassifikation mit Kombinationen der SOA-Features mit SVM. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes.

	W, P1	**W, P2**	**W+, P1**	**W+, P2**	**W, P1, P2**	**W+, P1, P2**	**P1, P2**
macro-F1	0,4581	0,4639	0,4623	0,4681	0,4650	0,4718	**0,4788**
Accuracy	0,6823	0,7027	0,6936	0,7192	0,7073	0,7227	**0,7518**
Recall							
- *Minderjährige*	**0,2798**	0,2528	0,2703	0,2297	0,2461	0,2399	0,2010
- *Erwachsene*	0,7125	0,7364	0,7254	0,7560	0,7419	0,7589	**0,7931**
Precision							
- *Minderjährige*	0,0681	0,0672	0,0688	0,0660	0,0668	**0,0695**	0,0680
- *Erwachsene*	0,9295	0,9292	0,9298	0,9290	0,9291	**0,9301**	0,9297
F1-Score							
- *Minderjährige*	0,1095	0,1061	**0,1097**	0,1025	0,1050	0,1077	0,1016
- *Erwachsene*	0,8067	0,8217	0,8150	0,8336	0,8250	0,8358	**0,8560**

Legende: W = Worte, W+ = Worte und Satzzeichen, P1 = universelle POS-Tags, P2 = detaillierte POS-Tags

allem die Satzzeichen-Features, die durchschnittliche Wortlänge sowie die Großbuchstaben-Features vertreten. Der höchste macro-F1-Wert von 0,4986 wurde mit den POS2-SOA in Verbindung mit den Großbuchstaben- und Satzzeichen-Features erreicht. Jedoch konnten die Resultate der einzeln verwendeten SOA-Typen (siehe Tabelle 6.11) bis auf die der POS1-SOA nicht erhöht werden.

Tabelle 6.15: Dargestellt sind die fünf höchsten erreichten macro-F1-Werte für jeden SOA-Typen und die entsprechenden Features, die zusätzlich zu den SOA-Features verwendet wurden, bei der 2-Klassen-Klassifikation mit RF. Der insgesamt höchste Wert ist fett hervorgehobenen.

Worte	Features	E	W2, S	W2	W2, G	S
	macro-F1	0,4978	0,4975	0,4966	0,4958	0,4951
Worte+	Features	W2, S	W2	S	G, S	W2, G, S
	macro-F1	0,4977	0,4977	0,4956	0,4951	0,4951
POS1	Features	E	S	W2, S	G	P, W1, G, S
	macro-F1	0,4950	0,4950	0,4938	0,4929	0,4927
POS2	Features	G, S	S	W2, G, S	W2	W2, G
	macro-F1	**0,4986**	0,4978	0,4970	0,4952	0,4950

Legende: P = Anteile der POS-Tags, W1 = Anteile von Worten bestimmter Zeichenlänge, W2 = durchschnittliche Wortlänge, G = Großbuchstaben-Features, S = Satzzeichen-Features, E = Anteil an Emoticons

Bei der Klassifikation mit der SVM (siehe Tabelle 6.16) wurden die besten Ergebnisse mit tendenziell mehr zusätzlichen Features erzielt. Die meist genutzten Feature-Gruppen waren wie bei der RF-Klassifikation die der Satzzeichen und Großbuchstaben, aber auch die Anteile von Worten bestimmter Länge. Für die besten Ergebnisse kam die durchschnittliche Wortlänge gar nicht zum Einsatz. Der höchste macro-F1-Wert von 0,4998 wurde mit den POS2-SOA in Verbindung mit Großbuchstaben-, Satzzeichen- und Emoticon-Features erzielt. Zudem konnten sämtliche Ergebnisse der einzeln genutzten SOA (siehe Tabelle 6.12) durch die Hinzunahme statistischer Features verbessert werden.

Tabelle 6.16: Dargestellt sind die fünf höchsten erreichten macro-F1-Werte für jeden SOA-Typen und die entsprechenden Features, die zusätzlich zu den SOA-Features verwendet wurden, bei der 2-Klassen-Klassifikation mit SVM. Der insgesamt höchste Wert ist fett hervorgehobenen.

Worte	Features	W1, G, S, E	W1, G, S	P, W1, G, S, E	P, W1, G, S	W1, S
	macro-F1	0,4989	0,4985	0,4984	0,4982	0,4976
Worte+	Features	W1, G, S, E	W1, S	W1, G, S	P, W1, G, S	P, W1, G, S, E
	macro-F1	0,4993	0,4984	0,4984	0,4984	0,4951
POS1	Features	W1, S	G, S, E	G, S	W1, G, S	W1, G, S, E
	macro-F1	0,4996	0,4996	0,4991	0,4988	0,4988
POS2	Features	G, S, E	G, S	W1, S	W1, G, S, E	W1, G, S
	macro-F1	**0,4998**	0,4995	0,4995	0,4992	0,4988

Legende: P = Anteile der POS-Tags, W1 = Anteile von Worten bestimmter Zeichenlänge, W2 = durchschnittliche Wortlänge, G = Großbuchstaben-Features, S = Satzzeichen-Features, E = Anteil an Emoticons

6.3 Zusammenfassung der Ergebnisse

In diesem Abschnitt werden die besten Ergebnisse der vorherigen Abschnitte zusammengefasst sowie Beobachtungen bezüglich der verschiedenen Features und Klassifikationstechniken diskutiert.

Grundsätzlich konnten die Baseline-Werte in den Experimenten übertroffen werden, wobei es für die 3-Klassen-Klassifikation in mehr Fällen gelungen ist als bei der 2-Klassen-Klassifikation. Insgesamt haben die Modelle mit Features basierend auf SOA sowohl bei der 3- als auch bei der 2-Klassen-Klassifikation besser abgeschnitten als die n-Gramm-basierten Modelle. Die Reduktion des Feature-Raumes bzw. eine niedrig-dimensionale Datenrepräsentation zeigte sich damit als nützlich.

Die Ergebnisse der Modelle, die den höchsten macro-F1-Wert aufweisen pro Klassifikationstechnik und jeweils für die 3- und 2-Klassen-Klassifikation sind in Tabelle 6.17 zusammengefasst.

Abhängig von der Klassifikationstechnik waren entweder einfache SOA (RF) oder SOA in Verbindung mit zusätzlichen statistischen Features (SVM) am erfolgreichsten. Dabei sind es entweder die Wort- oder POS2-SOA, welche die Grundlage bildeten. Dies führt zu dem Schluss, dass POS-Tags, aber speziell detailliertere, Wort-Repräsentationen ersetzen können. Das hat insbesondere den Vorteil einer niedrigeren Dimensionalität des Feature-Raums in Bezug auf n-Gramm-Repräsentationen, da Textdaten durch weniger unterschiedliche POS-Tags als Worte repräsentiert werden. Auch die Berechnung der POS-Tag-SOA erfordert dadurch weniger Rechenoperationen, weil der Ausgangspunkt insgesamt weniger unterschiedliche Terme sind.

Weiterhin fällt auf, dass die Klassifikationstechniken unterschiedlich klassifizierten, was besonders an den Recall- und F1-Werten pro Klasse erkennbar ist. So nahm die SVM grundsätzlich mehr korrekte Vorhersagen der unterrepräsentierten Klasse (*10s* bzw. *Minderjährige*)

Tabelle 6.17: Übersicht über die insgesamt besten Ergebnisse der 3- und 2-Klassen-Klassifikation jeweils mit RF und SVM.

	3-Klassen-Klassifikation			**2-Klassen-Klassifikation**	
Klassifikator Features	RF POS2-SOA	SVM Worte-SOA mit W1, G, S, E	Klassifikator Features	RF Worte-SOA	SVM POS2-SOA mit G, S, E
macro-F1	0,3376	0,3334	macro-F1	0,5000	0,4998
Accuracy	0,4668	0,4360	Accuracy	0,8899	0,8106
Recall			Recall		
- *10s*	0,0580	0,1081	- *Minderjährige*	0,0490	0,1599
- *20s*	0,3890	0,3995	- *Erwachsene*	0,9531	0,8594
- *30s*	0,5667	0,4997			
Precision			Precision		
- *10s*	0,0752	0,0762	- *Minderjährige*	0,0726	0,0787
- *20s*	0,3677	0,3626	- *Erwachsene*	0,9303	0,9317
- *30s*	0,5717	0,5655			
F1-Score			F1-Score		
- *10s*	0,0655	0,0894	- *Minderjährige*	0,0585	0,1055
- *20s*	0,3781	0,3802	- *Erwachsene*	0,9416	0,8941
- *30s*	0,5692	0,5306			

Legende: W1 = Anteile von Worten bestimmter Zeichenlänge, G = Großbuchstaben-Features, S = Satzzeichen-Features, E = Anteil an Emoticons

vor als RF, gleichzeitig aber auch weniger von der Klasse mit den meisten Datenpunkten (*30s* bzw. *Erwachsene*). Die Precision-Werte pro Klasse waren dagegen recht ähnlich zwischen den Klassifikatoren, was zeigt, dass unter den Vorhersagen ungefähr gleich große Anteile pro Klasse korrekt klassifiziert wurden.

Insgesamt wurden mit RF etwas höhere macro-F1-Werte als mit SVM erreicht. Möglicherweise ist RF für die verwendeten Daten oder extrahierten Features besser geeignet. Denn auch bei López-Santillán et al. [68] sind unterschiedliche Accuracy-Werte für Klassifikationen mit SVM und RF (und zwei weiteren Klassifikationsmethoden) festzustellen. Welche der Techniken die höchste Accuracy aufwies hing bei ihrem Ansatz von dem genutzten Datensatz ab. Auch von Meina et al. [55], die Teilnehmer bei dem AP-Task der PAN 2013 [23] waren, wurden SVM und RF (und weitere Techniken) gegeneinander getestet. RF erzielte dabei mit den verwendeten Features die höchste Accuracy und auch im Vergleich mit den anderen Teams eines der besten Ergebnisse.

6.4 Ergebnisse auf die Cross-Genre-Daten

Die Modelle für die Cross-Genre-Klassifikation wurden auf dem binären Blog-Datensatz trainiert. Als Features und Klassifikatoren wurden die fünf Kombinationen ausgewählt, die bei der binären Klassifikation der Blog-Daten die höchsten macro-F1-Werte erzielten. Diese sind:

- Modell 1: RF mit SOA der Worte
- Modell 2: SVM mit SOA der detaillierten POS-Tags sowie Großbuchstaben-Features, Satzzeichen-Features und Anteil der Emoticons

- Modell 3: RF mit SOA der detaillierten POS-Tags
- Modell 4: SVM mit SOA der universellen POS-Tags sowie Anteile von Worten mit bestimmter Zeichenlänge und Satzzeichen-Features
- Modell 5: SVM mit SOA der Worte+ sowie Anteile von Worten mit bestimmter Zeichenlänge, Großbuchstaben-Features, Satzzeichen-Features und Anteil der Emoticons

Die Ergebnisse dieser Klassifikationen sind in Tabelle 6.18 dargestellt. Unter den Ergebnissen der fünf Modelle sticht Modell 4 besonders hervor. Nicht nur, weil es den höchsten macro-F1-Score von 0,551 sowie die höchste Accuracy von 0,5539 erzielte, sondern auch, da es als einziges Modell mehr Kinder als Erwachsene korrekt klassifizierte. Die übrigen Modelle identifizierten deutlich mehr Erwachsene als Minderjährige richtig wie es auch bei den Blog-Daten der Fall war, was in gewisser Weise die unausgeglichene Datenverteilung der Trainingsdaten widerspiegelt, da die Cross-Genre-Daten ausbalanciert sind. Anteilig wurden dennoch mehr Minderjährige und weniger Erwachsene korrekt klassifiziert als bei der 2-Klassen-Klassifikation mit den Single-Genre-Daten.

Tabelle 6.18: Ergebnisse der Cross-Genre-Klassifikation mit den fünf ausgewählten Modellen. Die fett hervorgehobenen Werte kennzeichnen den höchsten erreichten Wert des entsprechenden Evaluierungsmaßes. Die erste Zeile zeigt zum Vergleich die macro-F1-Werte, die mit den Modellen in der Single-Genre-Klassifikation erzielt wurden.

	Modell 1	**Modell 2**	**Modell 3**	**Modell 4**	**Modell 5**
Single-Genre macro-F1	0,5000	0,4998	0,4996	0,4996	0,4993
macro-F1	0,5045	0,4347	0,4446	**0,5510**	0,4790
Accuracy	0,5356	0,4661	0,5022	**0,5539**	0,4847
Recall					
- *Minderjährige*	0,2850	0,2303	0,1803	**0,6334**	0,3800
- *Erwachsene*	0,7862	0,7019	**0,8241**	0,4743	0,5895
Precision					
- *Minderjährige*	**0,5713**	0,4358	0,5061	0,5465	0,4807
- *Erwachsene*	0,5237	0,4770	0,5013	**0,5641**	0,4874
F1-Score					
- *Minderjährige*	0,3803	0,3013	0,2658	**0,5868**	0,4244
- *Erwachsene*	**0,6286**	0,5680	0,6234	0,5153	0,5336

Die festgestellte schlechtere Vorhersage der Minderjährigen im Vergleich zu den Erwachsenen (ausgenommen Modell 4) kann neben der Unausgeglichenheit der Trainingsdaten auch mit dem Ursprung der Daten zusammenhängen. Dass es sich bei den Opfern um Erwachsene handelt, die vorgaben Kinder zu sein und deren Schreibweise versuchten zu imitieren, stellt für diesen Task der Altersdetektion eine Herausforderung und potenzielle Fehlerquelle dar, da sich der Schreibstil von Pseudo-Kindern und echten Kindern unterscheiden kann. Jedoch gibt es ansonsten keine öffentlich verfügbaren Datensätze zur Cybergrooming-Detektion.

Modelle 1 und vor allem 4 erzielten außerdem einen höheren macro-F1-Wert auf die Cross-Genre-Daten als auf die Single-Genre-Daten, was für eine gewisse Robustheit der Modelle gegenüber veränderten Daten und anderen Datenverteilungen spricht. Die übrigen Modelle verzeichneten hingegen eine Verschlechterung von bis zu 0,065 des Vergleichswertes.

Die Ergebnisse der fünf Modelle lassen kaum Trends bezüglich besonders nützlicher Features und Klassifikationstechniken erkennen. Im Hinblick auf die Klassifikatoren lässt sich feststellen, dass RF grundsätzlich mehr Datenpunkte als *Erwachsene* klassifizierte als die SVM. In Bezug auf die Features kann die anfängliche Vermutung, dass für eine Cross-Genre-Klassifikation die Verwendung von POS-Tags zur Feature-Generierung besser als die Verwendung von Worten geeignet ist, in soweit bestätigt werden, dass das Modell 4 mit dem höchsten macro-F1-Wert die universellen POS-Tags als Feature-Grundlage nutzte. Allerdings basierten die Modelle 2 und 3 ebenfalls auf POS-Tags und sie erreichten die niedrigsten macro-F1-Werte. Eine klare Tendenz ist demnach nicht zu erkennen.

Im Hinblick auf den Task der Grooming-Erkennung anhand des Alters liegt der Fokus darauf so viele Groomer, also Erwachsene, wie möglich zu detektieren. Das ist mit Modell 3 am besten gelungen, womit über 80% der Groomer als *Erwachsene* klassifiziert wurden. Gleichzeitig wurden jedoch die wenigsten Opfer als *Minderjährige* erkannt. Werden als Teil eines Sicherheitssystems die Chats, in denen Erwachsene, also potenzielle Groomer, erkannt wurden, an einen Moderator zur genaueren Inspektion übermittelt, ist allerdings ein hoher Recall wichtiger als eine hohe Precision [24]. Denn durch die anschließende Begutachtung werden falsch-positive Einordnungen nachträglich aussortiert, wohingegen eine hohe Falsch-Negativ-Rate dafür sorgt, dass viele potenzielle Straftaten gar nicht erst entdeckt werden.

Die 5.958 Chatpartner der Grooming-Daten wurden unabhängig voneinander klassifiziert. Wenn ein System aber zur Erkennung von Kommunikation zwischen Kindern und Erwachsenen dienen soll, ist die korrekte Altersklassifikation beider Chatpartner relevant. Dafür gibt es in Tabelle 6.19 eine Übersicht darüber, wie viele der 2.979 Chats pro Modell vollständig korrekt klassifiziert, d.h. bei wie vielen Chats beide Chatpartner der richtigen Altersgruppe zugeordnet wurden. Mit rund 24% der Chats gelang diese Aufgabe Modell 4 am besten, wohingegen Modell 3 mit ca. 14% der Chats die wenigsten korrekt vorhersagte. Das zeigt, dass Modelle abhängig vom geplanten Einsatzbereich besser oder schlechter geeignet sind.

Tabelle 6.19: Übersicht über die genaue Anzahl sowie den Anteil an korrekt klassifizierten Chats der 2.979 Chats der Grooming-Daten.

	Modell 1	**Modell 2**	**Modell 3**	**Modell 4**	**Modell 5**
Anzahl an korrekt erkannten Chats	657	442	421	719	628
Anteil an korrekt erkannten Chats	22,1%	14,8%	13,5%	24,1%	21,1%

Auch wenn die erzielten Resultate für den aktiven Einsatz der Modelle nicht ausreichend sind, zeigt sich, dass durch die Alterserkennung ein Beitrag zur Grooming-Detektion geleistet werden kann. Mit solch einem System zur Altersbestimmung von Nutzern einer Online-Plattform kann die Anzahl an zu überprüfenden potenziellen Grooming-Fällen verringert werden durch den direkten Ausschluss von Gesprächen zwischen Gleichaltrigen, egal ob Kinder oder Erwachsene. Da zwar die Kommunikation zwischen einem Kind und einem Erwachsenen nicht immer ins Grooming-Spektrum eingeordnet werden kann, bieten Systeme zur Altersbestimmung aber zumindest eine Vorselektion der Daten, die die Arbeitslast von Strafverfolgungsbehörden verringern kann.

7 Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurden Versuche im Hinblick auf die Alterserkennung von Blog-Autoren anhand ihrer geschriebenen Texte vorgenommen. Es galt die Autoren in drei Altersgruppen einzuordnen. Da im forensischen Kontext insbesondere die Unterscheidung zwischen Minderjährigen und Erwachsenen eine wichtige Rolle spielt, wurde zudem diese Art der Differenzierung betrachtet und die Modelle auch für eine binäre Klassifikation getestet.

Es wurden Experimente mit drei Feature-Arten durchgeführt, die n-Gramme, die SOA und statistische Features. Zudem wurden zwei Klassifikationstechniken verwendet, RF und SVM. Die SOA erzielten im Vergleich zu den n-Grammen bessere Ergebnisse und konnten teilweise durch die Hinzunahme bestimmter statistischer Features noch erhöht werden. Die Ergebnisse erzielt mit RF lagen außerdem geringfügig über denen der SVM.

Der beste macro-F1-Wert belief sich bei der Mehrlassen-Klassifikation auf 0,3376, bei der binären Klassifikation auf 0,5. Diese Werte wurden mit SOA basierend auf detaillierten POS-Tags bzw. Worten erreicht.

Final wurden die besten binär klassifizierenden Modelle für die Klassifikation von Cross-Genre-Daten verwendet. Bei diesen Daten handelt es sich um Chatnachrichten aus Grooming-Fällen, die einen relevanten Aspekt für die Strafverfolgung darstellen. Der höchste macro-F1-Wert von 0,551 überschritt das Ergebnis desselben Modells auf die Single-Genre-Daten unter Verwendung der folgenden Features: SOA basierend auf universellen POS-Tags in Verbindung mit Anteilen von Worten bestimmter Zeichenlänge, Großbuchstaben- und Satzzeichen-Features sowie der Anteil an Emoticons.

Die Verwendung von Altersbestimmung zur Erkennung von Grooming-Fällen kann demnach durchaus hilfreich sein, auch wenn die hier erzielten Ergebnisse für den aktiven Einsatz der Modelle nicht ausreichend sind. Um die Sicherheit der Vorhersagen zu erhöhen könnten auch verschiedene Modelle auf die Daten angewendet werden. Denn wie in dieser Arbeit festgestellt wurde, klassifizieren verschiedene Klassifikationstechniken durchaus unterschiedlich. So hat die SVM stets mehr Kinder korrekt klassifiziert als RF. Wohingegen mit RF höhere macro-F1-Werte erzielt wurden. Zudem unterschied sich auch die Feature-Auswahl der besten Modelle pro Klassifikator.

Da die SOA als niedrig-dimensionale Repräsentationsweise besser für die Altersbestimmung funktionierte als die herkömmlichen n-Gramme, kann in Zukunft nach weiteren nützlichen textuellen Einheiten für die SOA gesucht werden. Auch die Forschung im Hinblick auf die Verwendung von POS-Tags scheint sinnvoll zu sein, da sie bezüglich der Ergebnisse dieser Arbeit dem Nutzung von Worten nicht nachstanden.

Die hier getesteten Altersbestimmungsmodelle konnten auch auf die Grooming-Detektion übertragen werden. Jedoch ist nicht jeder Erwachsene ein Groomer, weshalb diese Methode eher als Vorselektion der Daten eingesetzt werden sollte. In Zukunft kann an Modellen gearbeitet werden, die neben einem solchen Altersbestimmungsmodul ein nachgeschaltetes

Grooming-Detektionsmodul enthalten, wie es in ähnlicher Form bereits von van de Loo et al. [24] vorgestellt wurde. Das Problem, dass Groomer sich häufig als Kinder ausgeben, muss in der Erstellung der Modelle auch beachtet werden, denn es kann eine größere Herausforderung darstellen diese Personen korrekt zu klassifizieren.

Zusammenfassend kann festgehalten werden, dass solche Altersbestimmungssysteme einen großen Dienst zur Erhöhung der Sicherheit im Internet leisten können, da sie effizienter arbeiten und viel mehr Inhalte überprüfen können als es für die Beamten der Strafverfolgung auf manuelle Weise möglich wäre. Auch wenn die Ergebnisse nicht immer korrekt sind, kann die Vorselektion der Daten hilfreich sein, um die finale Überprüfung durch Menschen zu ermöglichen.

Literaturverzeichnis

[1] Statista, Hrsg. „Anteil der Internetnutzer in Deutschland in den Jahren 2001 bis 2023 [Graph]“. Basierend auf dem D21-Digital-Index erhoben durch die Initiative D21 https://initiatived21.de/. (2024), Adresse: https://de.statista.com/statistik/daten/studie/13070/umfrage/entwicklung-der-internetnutzung-in-deutschland-seit-2001/ (besucht am 26. 04. 2024).

[2] Statista, Hrsg. „Generationen in Deutschland zu den beliebtesten Aktivitäten im Internet im Jahr 2023 [Graph]“. Basierend auf der Allensbacher Markt- und Werbeträger-Analyse - AWA 2023 durchgeführt von dem IfD Allensbach https://www.ifd-allensbach.de/. (2023), Adresse: https://de.statista.com/statistik/daten/studie/1425154/umfrage/umfrage-zu-den-aktivitaeten-im-internet-nach-generationen/ (besucht am 01. 05. 2024).

[3] S. Argamon, S. Dhawle, M. Koppel und J. W. Pennebaker, „Lexical Predictors of Personality Type“, in *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, (USA), 2005.

[4] A. Gasparetto, M. Marcuzzo, A. Zangari und A. Albarelli, „A Survey on Text Classification Algorithms: From Text to Predictions“, *Information*, Jg. 13, Nr. 2, 2022. DOI: 10.3390/info13020083. Adresse: https://www.mdpi.com/2078-2489/13/2/83.

[5] G. Tsoumakas und I. Katakis, „Multi-Label Classification: An Overview“, *International Journal of Data Warehousing and Mining*, Jg. 3, Nr. 3, S. 1–13, 2007, ISSN: 1548-3924. DOI: 10.4018/jdwm.2007070101.

[6] C. Catal und M. Nangir, „A sentiment classification model based on multiple classifiers“, *Applied Soft Computing*, Jg. 50, S. 135–141, 2017, ISSN: 15684946. DOI: 10.1016/j.asoc.2016.11.022.

[7] S. Zavrak und S. Yilmaz, „Email spam detection using hierarchical attention hybrid deep learning method“, *Expert Systems with Applications*, Jg. 233, S. 120 977, 2023, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.120977.

[8] D. W. Castro, E. Souza, D. Vitório, D. Santos und A. L. Oliveira, „Smoothed n-gram based models for tweet language identification: A case study of the Brazilian and European Portuguese national varieties“, *Applied Soft Computing*, Jg. 61, S. 1160–1172, 2017, ISSN: 15684946. DOI: 10.1016/j.asoc.2017.05.065.

[9] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes und D. Brown, „Text Classification Algorithms: A Survey“, *Information*, Jg. 10, Nr. 4, 2019. DOI: 10.3390/info10040150.

[10] Z.-H. Zhou, *Machine Learning*. Singapore: Springer Singapore, 2021, ISBN: 978-981-15-1966-6. DOI: 10.1007/978-981-15-1967-3.

[11] K. Sparck Jones, „A Statistical Interpretation of Term Specifity and its Application in Retrieval“, *Journal of Documentation*, Jg. 28, Nr. 1, S. 11–21, 1972, ISSN: 0022-0418. DOI: 10.1108/eb026526.

[12] C. Zhai und S. Massung, *Text Data Management and Analysis, A Practical Introduction to Information Retrieval and Text Mining* (ACM Books). New York, NY, USA: Association for Computing Machinery and Morgan & Claypool, 2016, Bd. 12, ISBN: 9781970001174. DOI: 10.1145/2915031.

[13] R. Bellman, *Dynamic programming*. Princeton, NJ: Princeton University Press, 1957.

[14] J. Li et al., „Feature Selection: A Data Perspective“, *ACM Computing Surveys*, Jg. 50, 2016. DOI: 10.1145/3136625.

[15] P. Dhal und C. Azad, „A comprehensive survey on feature selection in the various fields of machine learning“, *Applied Intelligence*, Jg. 52, Nr. 4, S. 4543–4581, 2022, ISSN: 0924-669X. DOI: 10.1007/s10489-021-02550-9.

[16] J. Devlin, M.-W. Chang, K. Lee und K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), J. Burstein, C. Doran und T. Solorio, Hrsg., Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, S. 4171–4186. DOI: 10.18653/v1/N19-1423.

[17] C. Cortes und V. Vapnik, „Support-Vector Networks“, *Machine Learning*, Jg. 20, Nr. 3, S. 273–297, 1995, ISSN: 08856125. DOI: 10.1023/A:1022627411411.

[18] T. K. Ho, „Random decision forests“, in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, (Montreal, Que., Canada), IEEE Comput. Soc. Press, 1995, S. 278–282, ISBN: 0-8186-7128-9. DOI: 10.1109/ICDAR.1995.598994.

[19] B. Murauer und G. Specht, „Generating Cross-Domain Text Classification Corpora from Social Media Comments“, in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, F. Crestani et al., Hrsg., Ser. Lecture Notes in Computer Science, Bd. 11696, Cham: Springer International Publishing, 2019, S. 114–125, ISBN: 978-3-030-28576-0. DOI: 10.1007/978-3-030-28577-7_7.

[20] E. Stamatatos, „On the robustness of authorship attribution based on character n-gram features“, *Journal of Law and Policy*, Jg. 21, S. 421–439, 2013, 01.

[21] L. Kaati, E. Lundeqvist, A. Shrestha und M. Svensson, „Author Profiling in the Wild“, in *2017 European Intelligence and Security Informatics Conference (EISIC)*, (Athens, Greece), IEEE, 2017, S. 155–158, ISBN: 978-1-5386-2385-5. DOI: 10.1109/EISIC.2017.32.

[22] S. Argamon und P. Juola, „Overview of the International Authorship Identification Competition at PAN 2011“, in *Notebook Papers of CLEF 2011 Labs and Workshops*, (Amsterdam, The Netherlands), V. Petras, P. Forner und P. D. Clough, Hrsg., Bd. 1177, CEUR-WS.org, 2011, ISBN: 978-88-904810-1-7. Adresse: http://ceur-ws.org/Vol-1177.

[23] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos und G. Inches, „Overview of the Author Profiling Task at PAN 2013“, in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, (Valencia, Spain), P. Forner, R. Navigli und D. Tufis, Hrsg., CEUR-WS.org, 2013, ISBN: 978-88-904810-3-1.

[24] J. van de Loo, G. de Pauw und W. Daelemans, „Text-Based Age and Gender Prediction for Online Safety Monitoring“, *International Journal of Cyber-Security and Digital Forensics*, Jg. 5, Nr. 1, S. 46–60, 2016, ISSN: 2305-0012. DOI: 10.17781/P002012.

[25] M. Ashcroft, L. Kaati und M. Meyer, „A Step Towards Detecting Online Grooming – Identifying Adults Pretending to be Children“, in *2015 European Intelligence and Security Informatics Conference*, (Manchester, UK), IEEE, 2015, S. 98–104, ISBN: 978-1-4799-8657-6. DOI: 10.1109/EISIC.2015.41.

[26] K. Siva et al., „Prevention of Emotional Entrapment of Children on Social Media“, in *2021 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, (Hyderabad, India), IEEE, 2021, S. 95–100, ISBN: 978-1-6654-1559-0. DOI: 10.1109/ICETCI51973.2021.9574068.

[27] D. Smahel et al., *EU Kids Online 2020, Survey results from 19 countries*, EU Kids Online, 2020. DOI: 10.21953/lse.47fdeqj01ofo. Adresse: https://www.lse.ac.uk/media-and-communications/assets/documents/research/eu-kids-online/reports/EU-Kids-Online-2020-10Feb2020.pdf (besucht am 10. 04. 2024).

[28] C. E. Chaski, „Author Identification In The Forensic Setting“, in *The Oxford Handbook of Language and Law*, L. M. Solan und P. M. Tiersma, Hrsg., Oxford University Press, 2012, S. 490–503, ISBN: 0199572127. DOI: 10.1093/oxfordhb/9780199572120.013.0036.

[29] F. Rangel et al., „Overview of the 2nd Author Profiling Task at PAN 2014“, in *Working Notes Papers of the CLEF 2014 Evaluation Labs*, (Sheffield, UK), L. Cappellato, N. Ferro, M. Halvey und W. Kraaij, Hrsg., Ser. CEUR Workshop Proceedings, Bd. 1180, CEUR-WS.org, 2014. Adresse: http://ceur-ws.org/Vol-1180/.

[30] F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein und W. Daelemans, „Overview of the 3rd Author Profiling Task at PAN 2015“, in *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers*, (Toulouse, France), L. Cappellato, N. Ferro, G. Jones und E. San Juan, Hrsg., Ser. CEUR Workshop Proceedings, Bd. 1391, CEUR-WS.org, 2015.

[31] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast und B. Stein, „Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, Bd. 1609, CEUR-WS.org, 2016.

[32] F. Rangel, P. Rosso, M. Potthast und B. Stein, „Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter“, in *Working Notes Papers of the CLEF 2017 Evaluation Labs*, (Dublin, Ireland), L. Cappellato, N. Ferro, L. Goeuriot und T. Mandl, Hrsg., Ser. CEUR Workshop Proceedings, Bd. 1866, 2017. Adresse: http://ceur-ws.org/Vol-1866/.

[33] F. Rangel, M. Montes-y-Gómez, M. Potthast und B. Stein, „Overview of the 6th Author Profiling Task at PAN 2018: Cross-domain Authorship Attribution and Style Change Detection“, in *CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers*, (Avignon, France), Linda Cappellato, Nicola Ferro, Jian-Yun Nie und Laure Soulier, Hrsg., Ser. CEUR Workshop Proceedings, Bd. 2125, CEUR-WS.org, 2018. Adresse: http://ceur-ws.org/Vol-2125/.

[34] Y. HaCohen-Kerner, D. Mughaz, H. Beck und E. Yehudai, „Words as classifiers of documents according to their historical period and the ethnic origin of their authors“, *Cybernetics and Systems*, Jg. 39, Nr. 3, S. 213–228, 2008, ISSN: 0196-9722. DOI: 10.1080/01969720801944299.

[35] Y. HaCohen-Kerner, H. Beck, E. Yehudai und D. Mughaz, „Stylistic feature sets as classifiers of documents according to their historical period and ethnic origin“, *Applied Artificial Intelligence*, Jg. 24, Nr. 9, S. 847–862, 2010, ISSN: 0883-9514. DOI: 10.1080/08839514.2010.514197.

[36] Y. HaCohen–Kerner, H. Beck, E. Yehudai, M. Rosenstein und D. Mughaz, „Cuisine: Classification using stylistic feature sets and/or name–based feature sets“, *Journal of the American Society for Information Science and Technology*, Jg. 61, Nr. 8, S. 1644–1657, 2010. DOI: 10.1002/asi.21350.

[37] P. Juola und R. H. Baayen, „A Controlled-corpus Experiment in Authorship Identification by Cross-entropy“, *Literary and Linguistic Computing*, Jg. 20, Nr. Suppl, S. 59–67, 2005, ISSN: 0268-1145. DOI: 10.1093/llc/fqi024.

[38] M. Koppel, J. Schler und K. Zigdon, „Determining an author's native language by mining a text for errors“, in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, (Chicago, Illinois, USA), R. Grossman, R. Bayardo und K. Bennett, Hrsg., New York, NY, USA: ACM, 2005, S. 624–628, ISBN: 159593135X. DOI: 10.1145/1081870.1081947.

[39] J. Schler, M. Koppel, S. Argamon und J. W. Pennebaker, „Effects of age and gender on blogging“, in *AAAI Spring Symposium: Computational approaches to analyzing weblogs*, (Stanford, California, 2006), N. Nicolov, F. Salvetti, M. Liberman und J. H. Martin, Hrsg., Ser. Technical report / American Association for Artificial Intelligence SS, Bd. 6, Menlo Park, Calif.: AAAI Press, 2006, S. 199–205, ISBN: 978-1-57735-264-8.

[40] S. Goswami, S. Sarkar und M. Rustagi, „Stylometric Analysis of Bloggers' Age and Gender“, *Proceedings of the International AAAI Conference on Web and Social Media*, Jg. 3, Nr. 1, S. 214–217, 2009. DOI: 10.1609/icwsm.v3i1.13992.

[41] M. Abdul-Mageed, C. Zhang, A. Rajendran, A. R. Elmadany, M. Przystupa und L. Ungar, *Sentence-Level BERT and Multi-Task Learning of Age and Gender in Social Media*, 2019. Adresse: https://arxiv.org/abs/1911.00637.

[42] I. Bilan und D. Zhekova, „CAPS: A Cross-genre Author Profiling System—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016.

[43] K. Bougiatiotis und A. Krithara, „Author Profiling using Complementary Second Order Attributes and Stylometric Features—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016.

[44] F. L. Cruz, R. Haro R. und F. J. Ortega, „ITALICA at PAN 2013: An Ensemble Learning Approach to Author Profiling—Notebook for PAN at CLEF 2013“, in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, (Valencia, Spain), P. Forner, R. Navigli und D. Tufis, Hrsg., CEUR-WS.org, 2013, ISBN: 978-88-904810-3-1.

[45] D. Dichiu und I. Rancea, „Using Machine Learning Algorithms for Author Profiling in Social Media—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016.

[46] P. Gencheva et al., „PANcakes Team: A Composite System of Genre-Agnostic Features For Author Profiling—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016. Adresse: http://ceur-ws.org/Vol-1609/.

[47] P. Modaresi, M. Liebeck und S. Conrad, „Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016. Adresse: http://ceur-ws.org/Vol-1609/.

[48] L. Gillam, „Readability for author profiling?—Notebook for PAN at CLEF 2013“, in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, (Valencia, Spain), P. Forner, R. Navigli und D. Tufis, Hrsg., CEUR-WS.org, 2013, ISBN: 978-88-904810-3-1.

[49] A. Poulston, M. Stevenson und K. Bontcheva, „Topic Models and n-gram Language Models for Author Profiling, Notebook for PAN at CLEF 2015“, in *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers*, (Toulouse, France), L. Cappellato, N. Ferro, G. Jones und E. San Juan, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2015.

[50] A. Zahid, A. Sampath, A. Dey und G. Farnadi, „Cross-genre Age and Gender Identification in Social Media—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016. Adresse: http://ceur-ws.org/Vol-1609/.

[51] L. Flekova und I. Gurevych, „Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media—Notebook for PAN at CLEF 2013“, in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, (Valencia, Spain), P. Forner, R. Navigli und D. Tufis, Hrsg., CEUR-WS.org, 2013, ISBN: 978-88-904810-3-1.

[52] A. P. López-Monroy, M. Montes-y-Gómez, H. J. Escalante, L. Villaseñor-Pineda und E. Villatoro-Tello, „INAOE's participation at PAN'13: Author Profiling task—Notebook for PAN at CLEF 2013“, in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, (Valencia, Spain), P. Forner, R. Navigli und D. Tufis, Hrsg., CEUR-WS.org, 2013, ISBN: 978-88-904810-3-1.

[53] R. Bakkar Deyab, J. Duarte und T. Gonçalves, „Author Profiling Using Support Vector Machines—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016.

[54] S. Argamon, M. Koppel, J. W. Pennebaker und J. Schler, „Mining the Blogosphere: Age, gender and the varieties of self-expression“, *First Monday*, Jg. 12, Nr. 9, 2007. DOI: 10.5210/fm.v12i9.2003.

[55] M. Meina et al., „Ensemble-based Classification for Author Profiling Using Various Features—Notebook for PAN at CLEF 2013“, in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, (Valencia, Spain), P. Forner, R. Navigli und D. Tufis, Hrsg., CEUR-WS.org, 2013, ISBN: 978-88-904810-3-1.

[56] C. Peersman, W. Daelemans und L. van Vaerenbergh, „Predicting age and gender in online social networks“, in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, (Glasgow, Scotland, UK), I. Cantador, F. M. Carrero, J. C. Cortizo, P. Rosso, M. Schedl und J. A. Troyano, Hrsg., New York, NY, USA: ACM, 2011, S. 37–44, ISBN: 9781450309493. DOI: 10.1145/2065023.2065035.

[57] M. J. Garciarena Ucelay et al., „Profile-based Approach for Age and Gender Identification - Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016. Adresse: http://ceur-ws.org/Vol-1609/.

[58] M. A. Rahman und Y. A. Akter, „Multi-lingual Author Profiling: Predicting Gender and Age from Tweets!“, in *Image Processing and Capsule Networks*, Ser. Advances in Intelligent Systems and Computing, J. I.-Z. Chen, J. M. R. S. Tavares, S. Shakya und A. M. Iliyasu, Hrsg., Bd. 1200, Cham: Springer International Publishing, 2021, S. 505–513, ISBN: 978-3-030-51858-5. DOI: 10.1007/978-3-030-51859-2_46.

[59] I. Ameer, G. Sidorov und R. M. A. Nawab, „Author profiling for age and gender using combinations of features of various types“, *Journal of Intelligent & Fuzzy Systems*, Jg. 36, Nr. 5, S. 4833–4843, 2019, ISSN: 10641246. DOI: 10.3233/JIFS-179031.

[60] M. Agrawal und T. Gonçalves, „Age and Gender Identification using Stacking for Classification - Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016.

[61] M. A. Álvarez-Carmona, A. P. López-Monroy, M. Montes-y-Gómez, L. Villaseñor-Pineda und H. J. Escalante, „INAOE's participation at PAN'15: Author Profiling task, Notebook for PAN at CLEF 2015“, in *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers*, (Toulouse, France), L. Cappellato, N. Ferro, G. Jones und E. San Juan, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2015.

[62] M. Busger op Vollenbroek et al., „GronUP: Groningen User Profiling—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016.

[63] I. Markov, H. Gómez-Adorno, G. Sidorov und A. Gelbukh, „Adapting Cross-Genre Author Profiling to Language and Corpus—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016.

[64] A. P. López-Monroy, M. Montes-y-Gómez, H. J. Escalante und L. Villaseñor-Pineda, „Using Intra-Profile Information for Author Profiling—Notebook for PAN at CLEF 2014“, in *Working Notes Papers of the CLEF 2014 Evaluation Labs*, (Sheffield, UK), L. Cappellato, N. Ferro, M. Halvey und W. Kraaij, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2014.

[65] R. Bayot und T. Gonçalves, „Author Profiling using SVMs and Word Embedding Averages - Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016.

[66] R. Bayot und T. Gonçalves, „Multilingual author profiling using word embedding averages and SVMs“, in *2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*, (Chengdu, China), IEEE, 2016, S. 382–386, ISBN: 978-1-5090-3298-3.

[67] T. Mikolov, K. Chen, G. Corrado und J. Dean, *Efficient Estimation of Word Representations in Vector Space*, 2013. Adresse: http://arxiv.org/pdf/1301.3781v3.

[68] R. López-Santillán, M. Montes-y-Gómez, L. C. González-Gurrola, G. Ramírez-Alonso und O. Prieto-Ordaz, „Richer Document Embeddings for Author Profiling tasks based on a heuristic search“, *Information Processing & Management*, Jg. 57, Nr. 4, S. 102 227, 2020, ISSN: 03064573. DOI: 10.1016/j.ipm.2020.102227.

[69] P. Bojanowski, E. Grave, A. Joulin und T. Mikolov, *Enriching Word Vectors with Subword Information*, 2016. Adresse: http://arxiv.org/pdf/1607.04606v2.

[70] A. Pavan, A. Mogadala und V. Varma, „Author Profiling Using LDA and Maximum Entropy - Notebook for PAN at CLEF 2013“, in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, (Valencia, Spain), P. Forner, R. Navigli und D. Tufis, Hrsg., CEUR-WS.org, 2013, ISBN: 978-88-904810-3-1.

[71] O. Pimas, A. Rexha, M. Kröll und R. Kern, „Profiling Microblog Authors Using Concreteness and Sentiment—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016.

[72] W.-Y. Lim, J. Goh und V. L. L. Thing, „Content-centric age and gender profiling - Notebook for PAN at CLEF 2013“, in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, (Valencia, Spain), P. Forner, R. Navigli und D. Tufis, Hrsg., CEUR-WS.org, 2013, ISBN: 978-88-904810-3-1.

[73] S. Ashraf, H. R. Iqbal und R. M. A. Nawab, „Cross-Genre Author Profile Prediction Using Stylometry-Based Approach—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016.

[74] E. Brunet, *Le Vocabulaire de Jean Giraudoux: structure et évolution : statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la langue française* (Travaux de linguistique quantitative). Slatkine, 1978.

[75] A. Honoré, „Some simple measures of richness of vocabulary“, *Association for Literary and Linguistic Computing Bulletin*, Jg. 7, Nr. 2, S. 172–177, 1979.

[76] H. S. Sichel, „On a distribution law for word frequencies“, *Journal of the American Statistical Association*, Jg. 70, Nr. 351a, S. 542–547, 1975.

[77] E. H. Simpson, „Measurement of Diversity“, *Nature*, Jg. 163, Nr. 4148, S. 688, 1949. DOI: 10.1038/163688a0.

[78] G. U. Yule, *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press, 1944.

[79] J. P. Kincaid, R. P. Fishburne Jr., R. L. Rogers und B. S. Chissom, *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*, Institute for Simulation and Training, University of Central Florida, 1975. DOI: 10.21236/ada006655. Adresse: https://stars.library.ucf.edu/istlibrary/56.

[80] E. A. Smith und R. J. Senter, *Automated Readability Index*, Aerospace Medical Research Laboratories, 1967. Adresse: https://apps.dtic.mil/sti/tr/pdf/AD0667273.pdf.

[81] C.-H. Björnsson, *Läsbarhet*. Stockholm: Liber, 1968, ISBN: 99-0346315-4.

[82] M. Coleman und T. L. Liau, „A computer readability formula designed for machine scoring“, *Journal of Applied Psychology*, Jg. 60, Nr. 2, S. 283–284, 1975, ISSN: 0021-9010. DOI: 10.1037/h0076540.

[83] R. Flesch, „A new readability yardstick“, *Journal of Applied Psychology*, Jg. 32, Nr. 3, S. 221–233, 1948, ISSN: 0021-9010. DOI: 10.1037/h0057532.

[84] G. H. McLaughlin, „SMOG Grading - A New Readability Formula“, *Journal of Reading*, Jg. 12, Nr. 8, S. 639–646, 1969.

[85] R. Gunning, „The Fog Index After Twenty Years“, *Journal of Business Communication*, Jg. 6, Nr. 2, S. 3–13, 1969, ISSN: 0885-2456. DOI: 10.1177/002194366900600202.

[86] E. Dale und J. S. Chall, „A Formula for Predicting Readability“, *Educational Research Bulletin*, Jg. 27, Nr. 1, S. 11–20+28, 1948.

[87] D. M. Blei, A. Y. Ng und M. I. Jordan, „Latent Dirichlet Allocation“, *Journal of Machine Learning Research*, Jg. 3, S. 993–1022, 2003, ISSN: 1532-4435.

[88] K. Santosh, R. Bansal, M. Shekhar und V. Varma, „Author Profiling: Predicting Age and Gender from Blogs—Notebook for PAN at CLEF 2013“, in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, (Valencia, Spain), P. Forner, R. Navigli und D. Tufis, Hrsg., CEUR-WS.org, 2013, ISBN: 978-88-904810-3-1.

[89] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer und R. Harshman, „Indexing by Latent Semantic Analysis“, *Journal of the American Society for Information Science and Technology*, Jg. 41, Nr. 6, S. 391–407, 1990. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[90] P. Paatero, „Least squares formulation of robust non-negative factor analysis“, *Chemometrics and Intelligent Laboratory Systems*, Jg. 37, Nr. 1, S. 23–35, 1997, ISSN: 01697439. DOI: 10.1016/S0169-7439(96)00044-5.

[91] C. Suman, P. Kumar, S. Saha und P. Bhattacharyya, „Gender Age and Dialect Recognition using Tweets in a Deep Learning Framework - Notebook for FIRE 2019“, in *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation*, (Kolkata, India), P. Mehta, P. Rosso, P. Majumder und M. Mitra, Hrsg., 2019.

[92] E. R. D. Weren, V. P. Moreira und José P. M. de Oliveira, „Exploring Information Retrieval features for Author Profiling—Notebook for PAN at CLEF 2014“, in *Working Notes Papers of the CLEF 2014 Evaluation Labs*, (Sheffield, UK), L. Cappellato, N. Ferro, M. Halvey und W. Kraaij, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2014. Adresse: http://ceur-ws.org/Vol-1180.

[93] E. R. D. Weren, „Information Retrieval Features for Personality Traits—Notebook for PAN at CLEF 2015“, in *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers*, (Toulouse, France), L. Cappellato, N. Ferro, G. Jones und E. San Juan, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2015.

[94] E. Moreau und C. Vogel, „Style-based Distance Features for Author Profiling—Notebook for PAN at CLEF 2013“, in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, (Valencia, Spain), P. Forner, R. Navigli und D. Tufis, Hrsg., CEUR-WS.org, 2013, ISBN: 978-88-904810-3-1.

[95] D. Pinto, H. Jiménez-Salazar und P. Rosso, „Clustering Abstracts of Scientific Texts Using the Transition Point Technique“, in *Computational Linguistics and Intelligent Text Processing*, D. Hutchison et al., Hrsg., Ser. Lecture Notes in Computer Science, Bd. 3878, Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, S. 536–546, ISBN: 978-3-540-32205-4. DOI: 10.1007/11671299_55.

[96] S. Mechti, M. Jaoua, R. Faiz, H. Bouhamed und L. Hadrich Belguith, „Author Profiling: Age Prediction Based on Advanced Bayesian Networks“, *Research in Computing Science*, Jg. 110, Nr. 1, S. 129–137, 2016, ISSN: 1870-4069. DOI: 10.13053/rcs-110-1-11.

[97] R. M. Ortega-Mendoza, A. P. López-Monroy, A. Franco-Arcega und M. Montes-y-Gómez, „Emphasizing personal information for Author Profiling: New approaches for term selection and weighting“, *Knowledge-Based Systems*, Jg. 145, S. 169–181, 2018, ISSN: 09507051. DOI: 10.1016/j.knosys.2018.01.014.

[98] D. Radha und P. Chandra Sekhar, „A Feature Selection Technique-Based Approach for Author Profiling“, in *Intelligent Systems and Sustainable Computing*, Ser. Smart Innovation, Systems and Technologies, V. S. Reddy, V. K. Prasad, D. N. Mallikarjuna Rao und S. C. Satapathy, Hrsg., Bd. 289, Singapore: Springer Nature Singapore, 2022, S. 583–591, ISBN: 978-981-19-0010-5. DOI: 10.1007/978-981-19-0011-2_52.

[99] S. Argamon, M. Koppel, J. W. Pennebaker und J. Schler, „Automatically profiling the author of an anonymous text“, *Communications of the ACM*, Jg. 52, Nr. 2, S. 119–123, 2009, ISSN: 0001-0782. DOI: 10.1145/1461928.1461959.

[100] M. Kocher und J. Savoy, „UniNE at CLEF 2016: Author Profiling—Notebook for PAN at CLEF 2016“, in *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, (Évora, Portugal), K. Balog, L. Cappellato, N. Ferro und C. Macdonald, Hrsg., Ser. CEUR Workshop Proceedings, CEUR-WS.org, 2016. Adresse: http://ceur-ws.org/Vol-1609/.

[101] C. Zhang und M. Abdul-Mageed, „BERT-Based Arabic Social Media Author Profiling“, in *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation*, (Kolkata, India), P. Mehta, P. Rosso, P. Majumder und M. Mitra, Hrsg., 2019.

[102] E. Alzahrani, M. Al Qurashi und L. Jololian, „Comparative Analysis of the Use of Pre-Trained Models to Profile Authors' Ages and Genders", in *2022 2nd International Conference on Computing and Machine Intelligence (ICMI)*, (Istanbul, Turkey), IEEE, 2022, S. 1–7, ISBN: 978-1-6654-7483-2. DOI: 10.1109/ICMI55296.2022.9873677.

[103] W. W. Cohen, „Fast Effective Rule Induction", in *Proceedings of the Twelfth International Conference on Machine Learning*, (Tahoe City, Kalifornien), Elsevier, 1995, S. 115–123, ISBN: 9781558603776.

[104] A. Basile, G. Dwyer, M. Medvedeva, J. Rawee, H. Haagsma und M. Nissim, „Simply the Best: Minimalist System Trumps Complex Models in Author Profiling", in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Ser. Lecture Notes in Computer Science, P. Bellot et al., Hrsg., Bd. 11018, Cham: Springer International Publishing, 2018, S. 143–156, ISBN: 978-3-319-98931-0. DOI: 10.1007/978-3-319-98932-7_14.

[105] S. S. Reddy Seelam, S. Kumar, C. M. Gopi und R. T. Raghunadha, „A New Term Weight Measure for Gender and Age Prediction of the Authors by analyzing their Written Texts", in *2018 IEEE 8th International Advance Computing Conference (IACC)*, (Greater Noida, India), IEEE, 2018, S. 150–156, ISBN: 978-1-5386-6678-4. DOI: 10.1109/IADCC.2018.8692092.

[106] K. Kavuri und M. Kavitha, „A Term Weight Measure based Approach for Author Profiling", in *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, (Chennai, India), IEEE, 2022, S. 275–280, ISBN: 978-1-6654-8385-8. DOI: 10.1109/ICESIC53714.2022.9783526.

[107] Statista, Hrsg. „Anteil der Internetnutzer nach Altersgruppen in Deutschland in den Jahren 1997 bis 2023 [Graph]". Basierend auf der ARD/ZDF-Onlinestudie 1997 bis 2023 https://www.ard-zdf-onlinestudie.de/. (2023), Adresse: https://de.statista.com/statistik/daten/studie/36149/umfrage/anteil-der-internetnutzer-in-deutschland-nach-altersgruppen-seit-1997/ (besucht am 07. 01. 2024).

[108] F. Rangel, P. Rosso, A. Charfi, W. Zaghouani, B. Ghanem und J. Sánchez-Junquera, „Overview of the Track on Author Profiling and Deception Detection in Arabic", in *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation*, (Kolkata, India), P. Mehta, P. Rosso, P. Majumder und M. Mitra, Hrsg., 2019.

[109] Y. Sun et al., „Author Profiling in Arabic Tweets: An Approach based on Multi-Classification with Word and Character Features", in *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation*, (Kolkata, India), P. Mehta, P. Rosso, P. Majumder und M. Mitra, Hrsg., 2019.

[110] G. Inches und F. Crestani, „Overview of the International Sexual Predator Identification Competition at PAN-2012", in *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, (Rome, Italy), P. Forner, J. Karlgren und C. Womser-Hacker, Hrsg., CEUR-WS.org, 2012, ISBN: 978-88-904810-3-1.

[111] J. Tam und C. H. Martell, „Age Detection in Chat", in *2009 IEEE International Conference on Semantic Computing*, (Berkeley, CA, USA), IEEE, 2009, S. 33–39, ISBN: 978-1-4244-4962-0. DOI: 10.1109/ICSC.2009.37.

Eidesstattliche Erklärung

Hiermit versichere ich – Maria Starke – an Eides statt, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach Publikationen oder Vorträgen anderer Autoren entnommen sind, habe ich als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt oder anderweitig veröffentlicht.

[redacted] Mai 2024 [redacted]

Ort, Datum Maria Starke