

Wer zwitschert denn da?

Autorenschaftsattributions mittels stilistischer Merkmale für kurze Social-Media-Nachrichtentexte

Katharina Luger¹ und Jörg Schmittwilken ²


Abstract: Zur Bekämpfung von Computerkriminalität sowie zur Wahrung der Informationssicherheit ist es vielfach notwendig, die Autorenschaft von Texten zu kennen oder zu ermitteln. Gerade die Zuordnung anonymer Texte zu einer möglichen Autorin oder einem möglichen Autor ist in diesem Kontext ein häufig zu lösendes Problem. Beispielsweise muss im Rahmen der Ermittlungsarbeit zu Hass-Kommentaren die Menge möglicher Autor:innen bestenfalls auf eine Person reduziert werden. In diesem Beitrag wird ein Modell zur Autorenschaftsattributions vorgestellt, das mithilfe von maschinellem Lernen aus einem Datensatz mit den Tweets von 915 Twitter-Accounts gelernt wurde. Das Modell basiert auf Support-Vector-Machines. Der Fokus des Beitrags richtet sich auf das Feature-Engineering, also der Erstellung sowie der Auswahl von Merkmalen, auf denen das Modell basiert. Es werden Feature sowie andere Modellparameter vorgestellt, die eine Klassifikationsgenauigkeit von bis zu 63% erzielen.

Keywords: Informationssicherheit, Computerkriminalität, Autorenschaftsattributions, Maschinelles Lernen, Support-Vector-Machine, stilistische Merkmale

1 Motivation

Die Zuordnung der Autorenschaft eines Textes zu einer Person ist auch in vielen Kontexten der öffentlichen Verwaltung interessant und in Bezug auf die Informationssicherheit von höchster Bedeutung. So stehen beispielsweise Strafverfolgungsbehörden im Rahmen der Ermittlungsarbeit von Computerkriminalität häufig vor dem Problem, anonyme Texte einer Absenderin oder einem Absender zuzuordnen zu müssen – sei es bei anonymen oder extremistischen (Hass)postings in Internetforen

¹ Absolventin im Studiengang Verwaltungsinformatik der Hochschule des Bundes für öffentliche Verwaltung, katharina.luger@vit-bund.de

² Hochschule des Bundes für öffentliche Verwaltung, Studiengang Verwaltungsinformatik, Gescherweg 100, 48161 Münster, schmittwilken@vit-bund.de,  <https://orcid.org/0009-0009-4424-3008>

[AC05], im Chatverlauf beschlagnahmter Smartphones oder gar bei der Analyse des Codes von Schadsoftware [Ka19].

Auch im Rahmen der Plagiatserkennung ist die Identifikation der Autor:in der Originalquelle eine hilfreiche Information. Ferner kann im Rahmen der Betrugsbekämpfung die Herleitung von Information über die Autorenschaft eines Textes (z. B. einer Phishing-Mail oder einer gefälschten Bewertung) bei der Verbrechensaufklärung helfen.

Diese Zuschreibung von Texten zu deren Verfasser:innen wird als Autorenschaftsattributions (engl. authorship attribution) bezeichnet und ist Gegenstand dieses Beitrags. Viele Arbeiten zur Autorenschaftsattributions verwenden Textkorpusse mit langen Texten wie Dokumenten, Reden, E-Mails o.ä. [Sc13; BMA13]. Diese eignen sich aufgrund der Vielzahl der in den Texten enthaltenen Merkmale besonders gut und erzielen hohe Genauigkeiten bei der Zuweisung der Autor:innen.

Kennzeichnend für Social-Media-Nachrichten, die im Fokus dieses Beitrags liegen, ist jedoch ihre Kürze, das häufige Ignorieren grammatikalischer Regeln sowie die starke Verwendung von Emojis, sodass die Attribution der Autorenschaft bei diesen Texten schwieriger und weniger stark untersucht ist [Ro16; Sc13; BMA13].

Im Fokus dieses Beitrags steht die Forschungsfrage, ob geeignete Features konstruiert werden können, durch die eine Autorenschaftsattributions auch bei sehr kurzen Social-Media-Texten möglich ist.

Im Folgenden wird nach der Darstellung der verwandten Arbeiten in Abschnitt 2 ein auf Support-Vector-Machines basierendes Modell zur Autorenschaftsattributions vorgestellt, das für Texte des Kurznachrichtendienstes Twitter (s. g. Tweets) optimiert ist (Abschnitt 3). Der Fokus wird anschließend auf das Feature-Engineering, also die Auswahl der verwendeten Merkmale (Abschnitt 4) und die Güte der Klassifikation (Abschnitt 5) gelegt. In Abschnitt 6 wird die Arbeit zusammengefasst und es wird ein Ausblick gegeben.

2 Verwandte Arbeiten

Autorenschaftsattributions ist seit vielen Jahren Gegenstand der Forschung im Bereich des Maschinellen Lernens. In diesem Teilgebiet der künstlichen Intelligenz werden mithilfe von Lernalgorithmen Modelle aus Trainingsdaten gelernt. Diese Modelle können dann zur Klassifikation neuer Daten verwendet werden. Die Arbeiten im Bereich der Autorenschaftsattributions unterscheiden sich im Wesentlichen hinsichtlich der eingesetzten Lernmethode sowie der verwendeten Trainingsdaten. Zudem unterscheiden sich die Ansätze in Art, Anzahl und Umfang der verwendeten Feature. Eine Übersicht möglicher Verfahren zur Autorenschaftsattributions geben [Ro16; St09].

Etabliert sind für diese Anwendung insbesondere die leistungsfähigeren überwachten Lernalgorithmen wie Deep Learning auf Basis künstlicher neuronaler Netze, Random-

Forrests oder Support-Vector-Machines (SVM). Aber auch einfache statistische Ansätze wie Naive Bayes sowie N-Gramme werden zu diesem Zweck verwendet.

[BNJ03] stellen die *latent dirichlet allocation* (LDA) als generatives, probabilistisches Modell eines Textkorpus vor, das es ermöglicht, die latent im Text beinhalteten Themen zu identifizieren. Hierzu modellieren sie die Themen durch eine charakteristische, multinominale Verteilung von Worten.

[AMM17] zeigen ein Konzept zur Autorenschaftsattributions von Werken der arabischen Poesie. Sie setzen zur Klassifikation Naive Bayes sowie SVM ein. [Di03] wählen zur Identifikation der Autoren deutscher Zeitungsartikel ebenfalls SVM. Beim Feature-Engineering kommen unter anderem Part-of-speech-Tagging sowie N-Gramme zum Einsatz. Sie erzielen mit dem Ansatz eine Trefferquote von 60-80%. [Sc13] stellen die Konzepte der *signatur* sowie *flexible pattern* kurzer Social-Media-Nachrichten vor, mit deren Hilfe sie die Autorenschaft dieser Texte attributieren können. Sie schlagen die Verwendung dieser Metriken z.B. zur Klassifikation mithilfe von SVM vor und erzielen hiermit Genauigkeiten von bis zu 70%. [BMA13] stellen die Autorenschaftsattributions von Tweets im Rahmen forensischer Analysen vor. Hierzu verwenden sie insbesondere stilometrische Merkmale zur Klassifikation mit SVM und erzielen Genauigkeiten von bis zu 90%.

[CS03] vergleichen die Verwendung von N-Grammen und naive Bayes Ansätzen zur Autorenschaftsattributions sowie zur Bestimmung des Topics von langen Texten.

3 Methodik

Die Vorgehensweise folgt einem häufig anzutreffenden Vorgehen im Bereich der Autorenschaftsattributions. Als Datengrundlage kam eine Sammlung aller öffentlichen Tweets von 915 berühmten Twitter-Nutzer:innen zum Einsatz. Der User Ahmed Shahriar Sakib veröffentlichte diesen Datensatz unter der Lizenz „*for educational purposes only*“ [Sa22] auf der Internetseite Kaggle [Sa22].

Um die Arbeit mit dem vorliegenden Datensatz zu vereinfachen, werden einige Vorannahmen getroffen. Diese umfassen, dass jeder Tweet nur von einem Autor bzw. einer Autorin verfasst wurde, jeder Account nur von einer Person geführt wird und diese nicht versucht hat, den Schreibstil zu verfälschen.

3.1 Data Preparation

Zunächst wurde der Datensatz in einigen Schritten vorverarbeitet. Hierbei wurden unter anderem Zeichenketten, die nicht Teil des ursprünglichen Tweets sind, entfernt. Ein Beispiel hierfür ist der Zeitstempel der Veröffentlichung.

Zudem enthalten Social-Media-Nachrichten eine Reihe an domänenspezifischen Elementen wie Hashtags, Emojis, Hyperlinks und Referenzen, die einerseits Informationen über das Twitter-Verhalten einer Person liefern, andererseits durch detaillierte Analysen auch zu ungenaueren Ergebnissen führen können. [LWD] argumentieren beispielsweise, dass ein Großteil der Zeichen von Replies bereits vorgegeben ist, weshalb besonders bei kurzen Texten der Anteil an nicht selbst verfassten Zeichen sehr hoch sein kann. [Ro16] führen an, dass die Gefahr von Fehlzuordnungen von Tweets bei einer detaillierten Betrachtung sehr hoch ist, da überwachte Verfahren des maschinellen Lernens sehr viel Wert auf eine bestimmte User-Referenz legen, falls diese häufiger in Tweets vorkommt. Zudem fanden [LWD] bei ihren Untersuchungen zu Hashtags und Replies heraus, dass detaillierte Informationen der Hashtags kaum zur Verbesserung der Genauigkeit beitragen. Da derselbe Link meist nicht in mehreren Tweets vorkommt, werden alle Links bei [Ro16] und [Sc13] durch einen entsprechenden Tag ersetzt. Selbes gilt für Nummern, Datumsangaben und Uhrzeit [Ro16; Sc13]. Dieses Vorgehen soll auch hier Anwendung finden, indem alle eben genannten Komponenten durch die entsprechenden Tags `REF`, `HASH`, `URL`, `NUM`, `DATE` und `TIME` ersetzt wurden.

Besonders bei Emojis liegt die Vermutung nahe, dass Personen bestimmte Präferenzen bezüglich Art und Häufigkeit ihrer Verwendung besitzen. Diese Elemente sind im vorliegenden Datensatz in Form von UTF-8-Bytecode angegeben. Eine detaillierte Untersuchung der Emojis lag hier allerdings nicht im Fokus, weshalb auf eine aufwendige, differenzierte Betrachtung verzichtet wurde. Aus diesem Grund wurden hier alle alleinstehenden UTF-8-Bytecodes durch den Tag `EMOJ` ersetzt und somit nur ein Näherungswert der Verwendungshäufigkeit betrachtet.

Da nicht alle Tweets für eine Schreibstilanalyse relevant sind, mussten zudem einige Texte aus dem Korpus entfernt werden. Wie in verwandten Arbeiten wurde auf sämtliche Retweets verzichtet, da es sich hierbei üblicherweise um fremde Nachrichten handelt [Ro16; Sc13]. Zudem ist die Identifikation der Sprache nicht Teil der hiesigen Betrachtungen, weshalb auf eine entsprechende sprachliche Differenzierung verzichtet wurde. Um nur englische Tweets zu untersuchen, wurde ein Großteil der anderssprachigen Texte ausgeschlossen. [Ro16] sowie [Sc13] entfernen zudem kurze Nachrichten, die zu wenige Informationen über den Stil der verfassenden Person enthalten. Auch hier wurden deshalb alle Tweets gelöscht, die weniger als drei Komponenten beinhalten. Hierbei bilden Worte, Satzzeichen und Tags jeweils einzelne Komponenten.

Ebenfalls zu berücksichtigen ist der Zeitpunkt der Veröffentlichung, da sich der Stil einer Person mit der Zeit verändern kann. Um eine mögliche Stiländerung über die Jahre zu berücksichtigen, fanden nur Tweets der letzten 24 Monate Verwendung. Zuletzt wurden nur Profile verwendet, die nach der Vorverarbeitung noch mehr als 300 Tweets besitzen und sämtliche Usernamen mit einer Nummer ersetzt.

3.2 Text Representation Strategy

In der Stilanalyse werden aus einem Text sogenannte Features / Merkmale extrahiert (Feature-Extraktion). Diese stellen den Stil einer Person näherungsweise dar [HS14]. In der Literatur wird eine Vielzahl an möglichen Feature-Typen untersucht und diskutiert. Diese werden in der Regel in verschiedenen Kategorien zusammengefasst. [AC08] unterscheiden beispielsweise zwischen lexikalischen, syntaktischen, strukturellen, inhaltspezifischen und idiosynkratischen Merkmalen.

Um ein möglichst genaues Modell zu erhalten, müssen geeignete Feature-Typen gewählt werden, da nicht alle für jeden Anwendungszeck sinnvoll sind. In der Regel werden mehrere dieser ausgewählt und in einem Feature-Set zusammengefasst.

Um die Eignung der Merkmale auch bei besonders kurzen Texten zu prüfen, wurden die gewählten Feature-Typen für jeden Text einzeln extrahiert und zu einem Vektor zusammengefasst. Hier wird also der sogenannte instanzbasierte Ansatz verfolgt, bei dem jeder Text einen Vektor bildet, der den Schreibstil der verfassenden Person repräsentiert [St09].

3.3 Feature-Selektion

Da einige Feature-Typen eine sehr hohe Anzahl an Dimensionen annehmen können, wird nach Erhebung der Vektoren in der Regel eine Feature-Selektion durchgeführt. Hierbei wird die Zahl an Features reduziert und somit die Performance eines Klassifikators verbessert [Sa20, S. 83].

Eine Strategie der Feature-Reduktion ist die Betrachtung der Auftrittshäufigkeit eines Merkmals, da besonders häufig auftretende Features besser in der Lage sind, stilistische Veränderungen zu erfassen. Bei Betrachtung der hier erhobenen Vektoren fällt auf, dass einige Feature-Typen eine besonders hohe Dimensionalität zur Folge haben, einige Merkmale aber kaum Aussagekraft aufweisen. Bei Funktionswörtern wurden deshalb alle Worte entfernt, die nicht oder nur einmal im gesamten Trainingsdatensatz vorkommen. [Ro16] entfernen bei N-Grammen sämtliche Features, die nur einmal im Datensatz zu finden sind, da diese in zukünftigen Texten wahrscheinlich nicht noch einmal auftauchen. Dieses Vorgehen wurde hier übernommen.

Bei Zeichen-N-Grammen wurde die Feature-Anzahl zusätzlich noch einmal auf die 10.000 am häufigsten vorkommenden Tetragramme reduziert. Dieser Feature-Typ hatte nach dem vorherigen Verarbeitungsschritt immer noch eine signifikant hohe Zahl an Dimensionen, was die zur Verfügung stehende Rechenkapazität zur Parameterwahl der Support Vector Machines und zum Erlernen des Klassifikators überstieg.

Insgesamt konnte eine Vielzahl an Merkmalen ausgeschlossen werden. Dabei mussten keine besonders relevant erscheinenden Informationen entfernt werden, insbesondere im Fall von Funktionswörtern, Wort- und Part-of-speech-N-Grammen.

3.4 Maschinelles Lernen

Im nächsten Schritt werden die erhobenen Vektoren einer Methode des maschinellen Lernens übergeben und somit ein Klassifikator erlernt. Analog zu u.a. [Ro16], [Sc13] und [BMA13] kamen auch hier SVMs zum Einsatz. [AC05; zitiert nach Zh06] führen an, dass sich SVMs für rauschende Daten eignen und mit hoch dimensionalen Vektoren umgehen können. Deshalb eignen sie sich besonders für die Analyse von Online-Nachrichten [AC05]. Verglichen mit Deep Learning, das vor allem in den letzten Jahren im Bereich der Autorschaftsanalyse zunehmend an Popularität gewann, eignen sich SVMs besser für den hier betrachteten Einsatzzweck. [Ro21] stellten bei ihren Untersuchungen an russischen Texten fest, dass SVMs besonders dann eine signifikant bessere Leistung erbringen, wenn nur eine begrenzte Textlänge zur Verfügung steht, da neuronale Netze mehr Trainingsdaten benötigen, um informative Merkmale aus dem Text zu extrahieren. Die verfügbaren Daten könnten in realen Situationen [Ro21] – wie auch in den hier durchgeführten Experimenten, in denen relativ wenige und besonders kurze Texte (max. 280 Zeichen) pro Autor:in verwendet werden – nicht ausreichen, um exakte Ergebnisse mit Deep Learning zu erzielen [Ro21].

Die SVM gehört zu den sogenannten überwachten Lernverfahren, welche sich dadurch auszeichnen, dass die Daten gelabelt sind, die Zuordnung von Autor:in und Text also bereits vorliegt [AC08]. Im Schritt des Lernverfahrens wird nun versucht, die Trainingsdaten optimal zu trennen. Im einfachsten Fall (2 Autor:innen) wird im zweidimensionalen Raum eine lineare Trennlinie gesucht, welche die Datenpunkte bzw. Textinstanzen der Personen teilt und somit zwei Klassen bildet [Sa20, S. 123]. Bei Prüfung unbekannter Instanzen, sollten diese vom Klassifikator idealerweise der richtigen Klasse zugeordnet werden.

Zum Trainieren und Testen der Modelle werden Datensätze üblicherweise in Test- und Trainingsdaten unterteilt. Um trotz kurzer Texte möglichst gute Ergebnisse zu erzielen und gleichzeitig noch genug Autor:innen mit einer ausreichenden Anzahl an Instanzen zu behalten, wurde ein Verhältnis von 4:1 (240 Trainings- und 60 Test-Tweets pro Account) festgesetzt.

4 Auswahl der Merkmale

In der vorliegenden Arbeit bestand der Fokus darin, Feature-Typen zu ermitteln, die für die Anwendung im Social-Media-Bereich besonders sinnvoll erscheinen. Eine Übersicht des erarbeiteten Feature-Sets ist in Tab. 1 abgebildet.

Lexikalische Merkmale, welche die Verwendung einzelner Worte und Zeichen bzw. Zeichenketten untersuchen, scheinen besonders im Bereich von Social Media sinnvoll. Aufgrund der weniger strengen Vorgaben bezüglich Grammatik und Rechtschreibung, wurden hier besonders große Unterschiede und Eigenheiten zwischen Personen vermutet.

Satzzeichen werden teils versehentlich falsch verwendet oder vergessen, aber von einigen Autor:innen auch bewusst als Stilmittel genutzt. So finden sich in manchen Tweets Auffälligkeiten wie z.B. „!!!“. Aus ähnlichen Gründen wurde das Verhältnis zwischen Groß- und Kleinschreibung berücksichtigt. Daneben ist die durchschnittliche Wortlänge in der Literatur ein häufig aufgeführter Feature-Typ. Der Anteil an langen Worten entscheidet in der Regel mit darüber, wie komplex ein Text wirkt und ist auch von der Intention der verfassenden Person abhängig [Sa20, S. 31]. Da diese Faktoren auch im Social-Media-Bereich eine Rolle spielen, floss dieser Feature-Typ ebenfalls in die Betrachtung ein. Umgesetzt wurde dieser analog zum Ansatz von z.B. [We21] und [AC08], in welchem die durchschnittliche Anzahl an Zeichen pro Wort herangezogen wird.

Merkmalskategorie	Feature-Typen
Lexikalisch	<ul style="list-style-type: none"> • Verhältnis Satzzeichen zu Zeichen gesamt • Durchschnittliche Wortlänge • Verhältnis Groß- zu Kleinschreibung • Zeichen-Tetragramme ($n=4$)
Syntaktisch	<ul style="list-style-type: none"> • Häufigkeit Funktionswörter • POS-Monogramme ($n=1$) • POS-Bigramme ($n=2$)
Strukturell	<ul style="list-style-type: none"> • Anzahl Worte pro Nachricht • Anzahl Absätze
Inhaltsspezifisch	<ul style="list-style-type: none"> • Wort-Monogramme ($n=1$)
Social-Media-spezifisch	<ul style="list-style-type: none"> • Verhältnis Hashtags zu Token • Verhältnis Referenzen zu Token • Verhältnis URLs zu Token • Verhältnis Emoticons zu Token

Tab. 1: Übersicht der gewählten Feature-Kategorien und -Typen

Laut [Sa] sind in der Autorenschaftsattributions Zeichen-N-Gramme das erfolgreichste Feature und werden bei der Analyse von besonders kurzen bzw. Social-Media-Texten u.a. von [AC08], [Sh17], [Sc13] sowie [LWD] verwendet. Bei sogenannten N-Grammen wird der Text als Gruppen von Worten, Zeichen oder Tags dargestellt [Br17]. Die hier verwendeten Zeichen-Tetragramme stellen die Tweets als Sammlung von Zeichenketten mit der Länge 4 dar. Der Satz „*Das ist ein Beispielsatz.*“ würde durch die folgende Menge von Zeichen-Tetragrammen dargestellt: {Das_; as_i; s_is; _is; ...; satz; atz.;}. Leerzeichen wurden hier durch _ gekennzeichnet.

Warum Zeichen-N-Gramme so effektiv sind, ist nicht vollständig geklärt. [Sa] fanden allerdings heraus, dass die Erfassung von Präfixen und Suffixen, welche Informationen

über die Morphologie eines Wortes liefern, und Zeichensetzung, besonders zur Effektivität dieses Feature-Typen beiträgt. [St09] sieht einen Vorteil bei Zeichen-N-Grammen im Social-Media-Bereich darin, dass diese von rauschenden Daten z.B. in Form von Grammatik- oder Zeichensetzungsfehlern, nicht übermäßig beeinträchtigt werden. Zudem können Zeichen-N-Gramme Präferenzen bezüglich Groß- und Kleinschreibung wie z.B. CamelCase sowie Zeichensetzung, beispielsweise Smileys in Form von „;“)“, erfassen [Ro16].

[Ro16] untersuchten ebenfalls Twitter Texte mittels Zeichen-N-Gramme und kamen aufgrund ihrer guten Klassifikationsergebnisse zu dem Schluss, dass dieser Feature-Typ sehr wichtig für die Autorenschaftsanalyse von Social-Media-Texten ist. Sie konzentrierten sich hierbei auf Tetragramme, da kleinere n redundante Informationen und größere Werte ähnliche Hinweise wie die von Wort-N-Grammen erfassen würden [Ro16]. Deshalb wurden hier ebenfalls Wort-Tetragramme in das Feature-Set aufgenommen.

Aufgrund der Vorteile von Wort- und POS-N-Gramme, welche auch auf Social-Media-Texte übertragbar sind, und der Feststellung von [Ro16], dass Zeichen-Tetragramme, Wort-Unigramme und POS-Uni- und POS-Bigramme 93% der Wichtigkeit ihres Feature-Sets ausmachten, wurden diese in die hiesigen Untersuchungen eingeschlossen. Sie funktionieren nach dem gleichen Prinzip wie Zeichen-N-Gramme. Anstelle von n-langen Zeichenketten, werden einzelne Wort oder POS-Tags, welche die zugehörige Wortart eines Begriffes abbilden, zur Darstellung der Texte verwendet.

Die Erfassung von Funktionswörtern hat sich in der Literatur ebenfalls als erfolgreich herausgestellt. Diese besitzen eine rein grammatikalische Bedeutung und werden lediglich genutzt, um Inhaltswörter zu verbinden [Be18]. [Ro16] sehen Funktionswörter im Bereich von Social-Media-Texten als besonders hilfreich an, da diese auch in kurzen Texten mit hoher Wahrscheinlichkeit auftreten. In der vorliegenden Untersuchung wird der Ansatz und die Wortliste von [We21] übernommen, welche eine Liste von 815 englischen Worten nutzten. Die Liste umfasste zum Zeitpunkt des Downloads 851 Worte und wurde von der Internetseite countwordsfree.com heruntergeladen.

5 Ergebnisse

Mittels Testdatensatz konnten die trainierten SVMs geprüft werden. Hierbei wurden zunächst eine Autorenschaftsattributions mit nur einem Feature-Typ, dann mit einer kompletten Feature-Kategorie und zuletzt mit dem kompletten Feature-Set mit steigender Anzahl an Autor:innen durchgeführt. Ermittelt wurde jeweils der Wert accuracy (Genauigkeit), welcher das Verhältnis zwischen der Anzahl korrekter Vorhersagen und der Anzahl aller Vorhersagen darstellt. Anhand der Ergebnisse erfolgte eine Einschätzung, wie gut sich die Feature-Typen für eine Autorenschaftsanalyse von kurzen Social-Media-Texten eignen.

Anzahl Autor:innen	Social-Media-spezifisch	Inhaltsspezifisch	Strukturell	Syntaktisch	Lexikalisch
5	0,4367	0,5033	0,3667	0,4233	0,6733
10	0,3350	0,4850	0,2350	0,4217	0,6167
20	0,2125	0,3667	0,1408	0,2575	0,4992

Tab. 2: Übersicht der Genauigkeit (accuracy) der Autorenschaftsattribution auf Basis der verschiedenen Feature Kategorien anhand von n zufälligen Autor:innen.

Die Ergebnisse der Autorenschaftsattribution mit allen Features innerhalb der jeweiligen Kategorie sind in Tab. 2 eingetragen. Demnach schneiden die strukturellen Merkmale erkennbar am schlechtesten ab. Dies lässt vermuten, dass strukturelle Eigenheiten aufgrund der Kürze der Texte, keine signifikanten Unterschiede aufweisen und Schreibstile von Autor:innen mit diesen Merkmalen nicht zuverlässig voneinander abgrenzbar sind.

Mit Social-Media-spezifischen Features konnten zwar bessere Ergebnisse erzielt werden, trotzdem wurden bei 20 Autor:innen nur knapp über 20 % der Texte richtig zugeordnet. Somit konnten Informationen zu Eigenheiten im Nutzungsverhalten solcher Elemente erfasst werden, im Vergleich zu anderen Feature-Kategorien schnitt diese Kategorie jedoch deutlich schlechter ab. Fraglich ist, ob eine genauere Erfassung von Emojis eine signifikante Verbesserung erzielt hätte.

Ähnlich verhält es sich mit syntaktischen Merkmalen, deren Ergebnisse hilfreich sein können, sich als alleinige Kategorie jedoch nicht eignen. Wie auch von [Ro16] festgestellt, schnitten die Wort-Monogramme (hier inhaltsspezifischen Merkmale) im Vergleich sehr gut ab. Anhand dieser konnten bei 20 Autor:innen immer noch über 35 % der Tweets richtig zugeordnet werden. Dies lässt darauf schließen, dass Personen bei kurzen Texten eine Präferenz für bestimmte Worte aufweisen. Somit kann empfohlen werden, diesen Feature-Typen bei der Analyse von Social-Media-Texten ergänzend einzusetzen.

Anzahl Autor:innen	Interpunktion	Durchschnittliche Wortlänge	Groß-/Kleinschreibung	Zeichen-Tetragramme
5	0,3033	0,2300	0,3533	0,6700
10	0,2217	0,1700	0,1467	0,6100
20	0,1083	0,0825	0,0825	0,4933

Tab. 3: Übersicht der Genauigkeit (accuracy) der Autorenschaftsattribution mit n Autor:innen und aller lexikalischen Feature-Typen

Auffällig gute Ergebnisse wurden mithilfe der lexikalischen Merkmale erreicht. Hier ist allerdings zu bemerken, dass vor allem die Zeichen-Tetragramme von Bedeutung waren.

In Tab. 3 ist zu sehen, dass die anderen Feature-Typen dieser Kategorie schlechte Ergebnisse erzielten. Unter Verwendung von Zeichen-Tetragramme als alleinstehender Feature-Typ kann bereits eine Vielzahl an Nachrichten richtig zugeordnet werden, weshalb deren Berücksichtigung bei der Analyse von Social-Media-Texten besonders hilfreich ist. Dies stellen auch [Ro16] fest. Bei einem Vergleich mit den in Tab. 4 dargestellten Ergebnissen, welche unter Verwendung des gesamten Feature-Sets erzielt wurden, schneiden Zeichen-Tetragramme ähnlich gut ab.

Abschließend ist festzustellen, dass anhand des komplette Feature-Sets, trotz besonders kurzer Texte, immer noch viele Instanzen richtig zugeordnet werden konnten. Tab. 4 zeigt, dass bei 20 Kandidat:innen fast 50 % und bei 100 Autor:innen noch über 30 % aller Tweets korrekt zugeordneten wurden.

Anzahl Autor:innen	Genauigkeit (accuracy)
10	0,6367
20	0,4967
50	0,3690
100	0,3110

Tab. 4: Übersicht der Ergebnisse der Autorenschaftsattribuion mit n Autor:innen und dem gesamten Feature-Set.

6 Zusammenfassung und Ausblick

In diesem Beitrag wurde eine Menge von 14 Feature-Typen (Tab. 1) aufgezeigt, die mit Support-Vector-Machines verwendet werden können, um ein Modell zur Autorenschaftsattribuion zu trainieren. Zum Lernen der Modelle wurde jeweils ein Trainingsdatensatz mit insgesamt 300 Twitter-Kurznachrichten von 5-100 unterschiedlichen Autor:innen verwendet. In Abhängigkeit der Anzahl und des Typs der verwendeten Features konnten mit dem Modell eine Prädiktionsgenauigkeiten von bis zu 63% erreicht werden. Somit ist die Zuschreibung eines Tweets zu einer Autorin oder einem Autoren mit der dargestellten Genauigkeit möglich.

Dieses Modell kann damit im Rahmen der Verbrechensbekämpfung und zur Strafverfolgung wertvolle Ansätze für die Ermittlungsarbeit bieten und auch im Rahmen eines Gerichtsverfahrens zur Überführung von Täterinnen und Tätern beitragen. Der präventive Einsatz dieses Modells wäre ebenfalls denkbar. Hier könnte beispielsweise im Rahmen der Schulsozialarbeit bereits bei Vorliegen von Chat-Nachrichten, die als Vorstufe von Mobbing einzuordnen sind, eine gezielte Ansprache der identifizierten Absender:innen erfolgen und so eine Eskalation hin zum Mobbing vermieden werden.

Die vorgestellten Features zeigen, dass die Forschungsfrage beantwortet werden konnte. Eine Autorenschaftsattributions ist unter Berücksichtigung sehr kurzen Länge der Texte mit guter Genauigkeit möglich.

Es wurde dargestellt, dass die Klassifikationsgenauigkeit stark von der Auswahl geeigneter Feature abhängt. Es haben sich Feature die 14 vorgestellten Feature-Typen als sehr gut geeignet erwiesen. Zukünftig sollten weitere Feature konzipiert und getestet werden. Hier sollte gerade im Kontext von Social-Media-Texten und Kurznachrichten das Augenmerk auf die Verwendung von Emoticons gelegt werden.

Kritisch wurde die Annahme bezüglich des Trainingsdatensatzes hinterfragt. Beim gewählten Datensatz der Celebrity-Tweets konnte nicht sicher beantwortet werden, ob die Tweets von einer Person oder einem Social-Media-Team erstellt wurden. So ist fraglich, ob die guten Vorhersageergebnisse erzielt werden konnten, weil sich die Tweets eines Accounts sehr ähnlich sind oder ob die Ergebnisse gut sind, obwohl sich die Tweets stark unterscheiden. Hier müssten zukünftig ebenfalls weitere Untersuchungen angestellt werden. Die Beschaffung eines geeigneten Datensatzes zur Verifikation dieser Hypothese könnte jedoch aufwendig sein, da er vermutlich für diesen Zweck erstellt werden müsste.

Zusammenfassend ist festzustellen, dass mit dem vorgestellten Modell der Autorenschaftsattributions ein Beitrag zur Informationssicherheit und auch zur Verfolgung von Computerkriminalität geleistet werden kann.

Literaturverzeichnis

- [AC05] Abbasi, Ahmed; Chen, Hsiu-chin; Applying Authorship Analysis to Extremist-Group Web Forum Messages. IEEE Intelligent Systems 20(5), 67-75, in: Intelligent Systems, IEEE, 20, 2005, S. 67–75.
- [AC08] Abbasi, Ahmed; Chen, Hsiu-chin; Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace, in: ACM Transactions on Information Systems, 26, 2008, S. 1–29.
- [AMM17] Ahmed, Al-Falahi; Mohamed, Ramdani; Mostafa, Bellafkih; Machine Learning for Authorship Attribution in Arabic Poetry, in: 20103751, 6, 2017, S. 42–46.
- [Be18] Beare; Kenneth; Content and Function Words, 02.10.2018, <https://www.thoughtco.com/content-and-function-words-1211726>. Abgerufen am 23.10.2022.
- [BMA13] Bhargava, Mudit; Mehndiratta, Pulkit; Asawa, Krishna; Stylometric Analysis for Authorship Attribution on Twitter, in: Vasudha Bhatnagar, Srinath Srinivasa (Hrsg.), Big Data Analytics: Second International Conference, BDA 2013, Mysore, India, December 16-18, 2013, Proceedings, Springer, Cham, 2013, S. 37–47.
- [BNJ03] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I.; Latent Dirichlet allocation, in: 0003-6951, 3, 2003, S. 993–1022.

-
- [Br17] Brownlee, Jason; A Gentle Introduction to the Bag-of-Words Model, 07.08.2017, <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>. Abgerufen am 17.10.2022.
 - [CS03] Clement, Ross; Sharp, David; Ngram and Bayesian Classification of Documents for Topic and Authorship, in: 0268-1145, 18, 2003, S. 423–447.
 - [Di03] Diederich, Joachim; Kindermann, Jörg; Leopold, Edda; Paass, Gerhard; Authorship Attribution with Support Vector Machines, in: 1573-7497, 19, 2003, S. 109–123.
 - [HS14] Halvani, Oren; Steinebach, Martin; Autorschaftsanalyse — die Illusion der Anonymität, in: Wirtschaftsinformatik & Management, 6, 2014, S. 33–43.
 - [Ka19] Kalgutkar, Vaibhavi; Kaur, Ratinder; Gonzalez, Hugo; Stakhanova, Natalia; Matyukhina, Alina; Code Authorship Attribution: Methods and Challenges, in: 0360-0300, 52, 2019, 3:1-3:36.
 - [LWD] Layton, Robert; Watters, Paul; Dazeley, Richard; Authorship Attribution for Twitter in 140 Characters or Less, in: 2010 Second Cybercrime and Trustworthy Computing Workshop, 2010, S. 1–8.
 - [Ro16] Rocha, Anderson; Scheirer, Walter; Forstall, Christopher; Cavalcante, Thiago; Theophilo, Antonio; Shen, Bingyu; Carvalho, Ariadne; Stamatatos, Efstathios; Authorship Attribution for Social Media Forensics, in: IEEE Transactions on Information Forensics and Security, 12, 2016, S. 5.
 - [Ro21] Romanov, Aleksandr; Kurtukova, Anna; Shelupanov, Alexander; Fedotova, Anastasia; Goncharov, Valery; Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks, in: Future Internet, 13, 2021.
 - [Sa] Sapkota, Upendra; Bethard, Steven; Montes, Manuel; Solorio, Thamar; Not All Character N-grams Are Created Equal: A Study in Authorship Attribution, in: , Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, Association for Computational Linguistics, 2015, S. 93–102.
 - [Sa20] Savoy, Jacques; Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling, Cham. Springer International Publishing AG, 2020.
 - [Sa22] Sakib, Ahmed Shahriar; Top 1000 Twitter Celebrity Tweets And Embeddings: Tweets and Embeddings of most followed celebrity twitter accounts, Juli 2022, <https://www.kaggle.com/datasets/ahmedshahriarsakib/top-1000-twitter-celebrity-tweets-embeddings>. Abgerufen am 12.10.2022.
 - [Sc13] Schwartz, R.; Tsur, O.; Rappoport, A.; Koppel, Moshe; Authorship attribution of micro-messages, in: EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2013, S. 1880–1891.
 - [Sh17] Shrestha, Prasha; Sierra, Sebastian; González, Fabio; Montes, Manuel; Rosso, Paolo; Solorio, Thamar; Convolutional Neural Networks for Authorship Attribution of Short Texts, 2017.
 - [St09] Stamatatos, Efstathios; A Survey of Modern Authorship Attribution Methods, in:

JASIST, 60, 2009, S. 538–556.

- [We21] Weerasinghe, Janith; Singh, Rhia; Greenstadt, Rachel; Feature vector difference based authorship verification for open-world settings, in: CEUR Workshop Proceedings, 2936, 2021, S. 2201–2207.
- [Zh06] Zheng, Rong; Li, Jiexun; Chen, Hsiu-chin; Huang, Zan; A framework for authorship identification of Online messages: Writing-style features and classification techniques, in: JASIST, 57, 2006, S. 378–393.