Lässt sich der Schreibstil verfälschen um die eigene Anonymität in Textdokumenten zu schützen?

Oren Halvani, Martin Steinebach, Svenja Neitzel*

Fraunhofer SIT, Darmstadt, {Halvani, Steinebach}@SIT.Fraunhofer.de
*TU Darmstadt, Svenja.Neitzel@Freenet.de

Abstract: Die Zahl textueller Daten wächst heutzutage zunehmend, insbesondere aufgrund nutzergenerierter Inhalte im Internet. Zu diesen zählen unter anderem Blogs, Forenbeiträge oder Kommentare, die über unzählige Plattformen verbreitet werden. Wünscht ein Autor hier anonym zu kommunizieren, nutzt er ein oder mehrere Pseudonyme. Schreibstile dagegen verbleiben ungeschützt in den Texten und können mit Hilfe sogenannter Autorschafts-Attributionssysteme bekannten Autoren zugeordnet werden. Aktuelle Systeme erzielen dabei je nach Szenario (Anzahl der Autoren, Qualität der Daten, etc.) gute bis sehr gute Ergebnisse. Wenn die Möglichkeit der Anonymität angestrebt wird, ist folglich eine wichtige Frage, ob und wie Schreibstile in Texten verfälscht werden können, um solche Systeme zu täuschen. In diesem Papier werden zunächst Systeme und deren Komponenten erläutert, mit deren Hilfe Texte hinsichtlich der darin enthaltenen Schreibstile de-anonymisiert werden können. Anschließend wird ein Überblick über manuelle und semi-automatische Gegenmaßnahmen gegeben. Weiterhin werden Möglichkeiten genannt, um eine vollautomatische Anonymisierung der Schreibstile zu realisieren.

1 Einführung

Dank des Internets ist die Zahl existierender Textdaten massiv angestiegen, nicht zuletzt aufgrund nutzergenerierter Inhalte. Dazu zählen vor allem Blogs, Forenbeiträge und Kommentare zu mehr oder weniger brisanten Themen. Autoren, die sich durch die Verwendung von Pseudonymen hinsichtlich ihrer Anonymität sicher fühlen, üben so teilweise sehr offen Kritik. So finden sich beispielsweise auf dem Portal *Jameda.de* ca. $3 \cdot 10^6$ Kommentare von Patienten über deren behandelnde Ärzte. Zum einen können so im Schutz der Anonymität berechtigte Warnungen vor Ärzten mit (zumindest subjektiv empfundenen) schlechten Behandlungsmethoden ausgesprochen werden. Zum anderen besteht die Gefahr, dass die Anonymität bewusst für eine Verfälschung der Bewertungen im positiven oder negativen Sinne missbraucht wird. Diese Gefahr wird noch durch die Möglichkeit verstärkt, mehrere Pseudonyme zu verwenden.

In beiden Fällen wird die sogenannte Autorschafts-Attribution (kurz AA) relevant, deren Ziel es ist, zu einem anonymen Dokument die Autorschaft zuzuordnen. Ein AA-System verlangt neben dem anonymen Dokument als Eingabe Beispieldokumente von in Frage kommenden Autoren, ein Klassifikationsverfahren sowie stilistische Merkmale, mit de-

ren Hilfe die Autoren voneinander unterschieden werden können. Die Erforschung der Möglichkeiten von AA-Systemen wird in verschiedenen Bereichen betrieben. So strebt die Forensik damit eine Zuordnung von Drohbriefen oder Bekennerschreiben zu bekannten Straftätern an. Für die Literaturwissenschaft ist die Zuordnung anonymer Schriften zu bekannten Schriftstellern eine interessante Herausforderung.

Autoren, die Drohungen hinter dem Schutz der Anonymität aussprechen, muss daher bewusst werden, dass sie durch die AA zunehmend in Bedrängnis geraten. Andererseits ist eine Weiterentwicklung dieser Technologie wichtig, um einen Missbrauch der anonymen Kommentierung einzudämmen, indem zumindest erkannt werden kann, dass eine Person unter mehreren Pseudonymen aktiv ist und so Bewertungen verfälscht.

In unserer Arbeit stellen wir das Konzept der AA vor und beschreiben dabei kurz, wie individueller Schreibstil zu einem durch Software nachweisbaren Merkmal wird. Weiterhin gehen wir der sich daraus ableitenden Frage nach, ob sich Autoren durch Methoden der bewussten Stilverfälschung vor einer Aufdeckung schützen können. Dazu wird zunächst in Kapitel 2 der Begriff *Schreibstil* näher erläutert, bevor in Kapitel 3 AA-Systeme und deren Komponenten beschrieben werden. In Kapitel 4 werden Maßnahmen gegen AA-Systeme präsentiert, wobei menschliche und semi-automatische Täuschungsmethoden diskutiert werden. Abschließend werden in Kapitel 5 weiterführende Möglichkeiten genannt, um vollautomatisierte Täuschungsmethoden realisieren zu können.

2 Schreibstil

Schreibstil stellt in diesem Papier ein wichtiges Konzept dar und wird neben dem Begriff Stil synonym verwendet. Da der Begriff jedoch nicht formalisierbar ist, existieren dafür nach [Gan08, Gol07, Sow73] unterschiedliche Auffassungen. Um Stil dennoch greifbar zu machen, wird daher eine Approximation anhand stilistischer Merkmale benutzt. Diese Merkmale werden im Fachjargon als *Features* bezeichnet und sind zentraler Gegenstand der *Stilometrie*, welche die maßgebende Disziplin der AA darstellt. Vereinfacht ausgedrückt: Stil wird anhand von Features charakterisiert, sodass dadurch Autoren voneinander unterschieden werden können. Wichtigste Randbedingung an Features ist jedoch, dass diese unabhängig von Inhalt (z.B. Fachsprache), Kontext (z.B. angesprochenes Publikum) und Funktion (z.B. sich reimende Wörter in einem Gedicht) sind. Daher gilt es, diese drei Komponenten bei der Approximation des Stils stets auszuschließen.

Features formen den wichtigsten Bestandteil von AA-Systemen. Aufgrund ihrer Vielzahl empfiehlt es sich, diese zu kategorisieren und konzeptionell zu betrachten. Tabelle 1 listet insgesamt 20 solcher Feature-Kategorien auf, die teilweise aus [Hal12] entnommen wurden.

3 Autorschafts-Attributionssysteme

Menschliche Leser können den Stil eines Autors erstaunlich leicht erfassen. Sie benötigen dafür weder eine Vergleichsbasis noch spezielle Verfahren. Meist können sie keine exakte Beschreibung des Stils wiedergeben, doch sie würden ihn in anderen Dokumenten des gleichen Autors wiedererkennen und auch Stilinkonsistenzen, z.B. aufgrund mehrerer Autoren bemerken (zumindest wenn nicht versucht wurde, den Schreibstil aneinander anzupassen). Allerdings haben menschliche Leser auch einige Nachteile. Zu diesen zählen neben Zeit und Kosten auch der Aufwand bzw. die Komplexität bei der Bewältigung der Untersuchung großer Textmengen. Daher wird zunehmend an der automatischen Erfassung von Stil geforscht, wobei Features eine zentrale Rolle einnehmen.

Features werden oftmals dafür kritisiert, Stil nur unvollständig zu charakterisieren. In [Sie13] wird als Begründung die nur schwer mögliche Trennung von Inhalt und Stil durch statistische Verfahren genannt. Weiterer Kritikpunkt ist das fehlende Textverständnis. So können z.B. Argumentation, Humor oder Ironie nur schwer statistisch erfasst werden, da diese auf Verständnis und Hintergrundwissen basieren, die im Text nicht explizit vorkommen. Dennoch wurden Features bereits erfolgreich eingesetzt (vgl. [KG06]). Gute Ergebnisse können erzielt werden, wenn Mensch und Maschine interagieren (menschliche Intuition vereinigt mit der Rechenfähigkeit des Computers). Für aussagekräftige Ergebnisse bietet es sich zudem an, mehrere Features in kombinierter Form zu betrachten, wobei eine solche Kombination als *Feature-Set*) bezeichnet wird. Die drei relevantesten Feature-Sets, die im Verlauf dieser Arbeit eine wichtige Rolle einnehmen, lauten: F_1 = *9 Feature-Set*, F_2 = *Synonym-Based Feature-Set* sowie F_3 = *Writeprint Feature-Set*. Tabelle 2 erläutert, welche Features hierbei konkret enthalten sind.

Neben Features stellen Klassifikatoren ebenfalls einen wichtigen Bestandteil von AA-Systemen dar. Ihre Aufgabe ist es, den wahrscheinlichsten Autor eines anonymen Dokuments anhand der Features und einer Menge von Beispieldokumenten bekannter Autoren vorherzusagen. Dabei gilt die Grundannahme, dass sich der wahre Autor in dieser Menge befindet. Klassifikatoren sind Verfahren aus dem Gebiet des Maschinellen Lernens. In der AA sind die am häufigsten verwendeten Klassifikatoren: Support Vector Machines, k-Nearest Neighbors sowie Naive Bayes. Ein genauerer Überblick über Klassifikatoren findet sich in [Kot07]. Heutzutage existieren bereits einige frei zugängliche AA-Systeme. Eines davon ist *JStylo* (erhältlich unter [PSA13]), welches im Folgekapitel betrachtet wird.

4 Möglichkeiten zur Täuschung der Autorschafts-Attribution

Um angesichts der steigenden Nutzung und Leistung von AA-Systeme anonym schreiben zu können, bedarf es einer gezielten Täuschung. Das gilt besonders dann, wenn andere Texte des Autors bereits öffentlich bekannt sind. Prinzipiell existieren zwei Möglichkeiten zur Täuschung der AA: Die **Anonymisierung** (z.B. durch Löschung diskriminierender Features) und die **Imitation** (z.B. durch Nachahmung von Features anderer Autoren). Als mögliche Gründe für die Täuschung der AA fallen einem zunächst zahlreiche kriminelle

Interessen ein. Es existieren jedoch auch legitime Gründe, warum Autoren an solchen Verfahren interessiert sein könnten. Werden z.B. regimekritische Blogger in totalitären Staatssystemen betrachtet, so ist Anonymisierung ein wesentliches Mittel zum Schutz der Blogger und zu einer wahrheitsgemäßen und dennoch sicheren Berichterstattung in Ländern ohne das Recht zur freien Meinungsäußerung. Wenn also im Folgenden von „Fälscher" die Rede ist, sind damit nicht zwangsläufig kriminell gesinnte Menschen gemeint.

4.1 Manuelle Täuschungsmethoden

Menschen sind in der Lage, Stilbrüche zu erkennen, sodass davon ausgegangen werden kann, dass sie ein Gespür für Stil haben. Daher stellt sich die Frage, ob Menschen mit Hilfe dieses Gespürs ihren eigenen Stil verbergen bzw. einen anderen Stil imitieren können. Gewöhnlich tritt dies nur in Zusammenhang mit kriminellen Absichten auf. Ein häufiges Beispiel sind „falsche" Chat-Freundschaften, die unter anderem von Pädophilen initiiert werden. Legale Imitierung anderer Autoren gibt es zudem in speziellen Wettbewerben, die zu Ehren bekannter Schriftsteller stattfinden und Hobby-Autoren dazu aufrufen, in ihren Beiträgen diese zu imitieren (z.B. *International Imitation Hemingway Competition*). Die Einreichungen zu solchen Wettbewerben enthalten wertvolle Informationen zur menschlichen Stil-Imitation und werden daher auch als Forschungs-Korpora verwendet.

Diese Beispiele finden jedoch Menschen gegenüber statt. Daher ist ein Täuschungserfolg auch davon abhängig, wie leicht sich das Opfer täuschen lässt und hat wenig objektive Aussagekraft über die Fähigkeiten des Fälschers. Aussagekräftiger wäre es zu wissen, wie AA-Systeme auf menschliche Imitation oder Anonymisierung von Texten reagieren. Genau dies wurde an der Drexel University Philadelphia mit dem dort entwickelten Programm *JStylo* getestet, welches die vorgestellten Feature-Sets und verschiedene Klassifikatoren standardmäßig bereitstellt. Dort wurde ein Korpus von Dokumenten mehrerer Hobby-Autoren gebildet (*Brennan-Greenstadt Adversarial Corpus*). Diese sollten von ihnen in der Vergangenheit geschriebene Texte beliebiger Thematik einreichen. Außerdem wurden sie gebeten, einen Text in dem sie ihren eigenen Stil bewusst verstecken, sowie einen Text im Stil des US-amerikanischen Roman-Autors Cormac McCarthy zu schreiben. Sie erhielten dazu einen Auszug aus dessen Roman *The Road*, [McC06]. Die verfälschten (also die anonymisierten und imitierten) Texte hatten im Gegensatz zu den echten Texten vorgegebene Themen: Bei dem anonymisierten Text sollten die Probanden ihre Nachbarschaft beschreiben, bei dem imitierten Text ihren bisherigen Tagesablauf wiedergeben.

JStylo erhielt als Trainingsdaten die echten Texte zugeordnet zum jeweiligen Autor. Für die grundsätzliche Genauigkeit[1] von *JStylo* sollten zunächst erneut echte Texte klassifiziert werden. Dabei wurde auch die Anzahl der Autoren variiert. Als Feature-Sets wurden F_1, F_2 und F_3 verwendet, wobei Letzteres die beste Genauigkeit erzielte (auch bei einer AA mit 40 Autoren lag diese noch bei über 80%). F_2 schnitt auch gut ab, lag jedoch stets ca. 5% unter F_3. Das weniger umfangreiche und weniger komplexe F_1 erreichte dagegen deutlich schlechtere Werte, die ab einer Anzahl von 10 Autoren nur noch unter 50%

[1] Genauigkeit bezeichnet hier den Anteil der korrekt klassifizierten Texte unter allen Klassifikationen.

betrugen. Dennoch sollte bedacht werden, dass F_1 in seiner Genauigkeit weit über einer zufälligen Klassifikation liegt (bei 40 Autoren lag die Genauigkeit von F_1 immerhin bei ca. 25% im Gegensatz zu einer zufälligen Klassifikation mit 2,5%). In Anbetracht der Einfachheit von F_1 ist dies ein gutes Ergebnis.

Mit den verfälschten Texten konnten die Hobby-Autoren, die sich vorher nie mit Stilometrie beschäftigt hatten, *JStylo* sehr stark täuschen. Die Genauigkeit bei der Klassifikation der anonymen Texte anhand von F_3 lag nur noch knapp über der der zufälligen Klassifikation. Die Genauigkeiten von F_1 und F_2 lagen entsprechend darunter und ab einer Zahl von 30 möglichen Autoren sogar nur noch knapp über 0%. Die imitierten Texte trieben sogar die Genauigkeit von F_3 unter die der zufälligen Klassifikation. Für einen Güte-Test der Imitate wurden die Trainingsdaten um CormacMcCarthy selbst als potenziellen Autor sowie Textauszüge aus seinen Romanen ergänzt. *JStylo* sollte anschließend erneut den Autor der Imitate bestimmen. Die höchste Genauigkeit in der Klassifikation von Cormac McCarthy als Autor der imitierten Texte erzielte nur F_1: Es lag bei 5 Autoren knapp unter 70% und bei bis zu 30 Autoren noch über 50%. Ähnliche Werte erreichte F_2. Dagegen erzielte F_3 im Schnitt nur halb so hohe Werte wie die anderen beiden. Detaillierte Ergebnisse finden sich in [BAG12]. Der verwendete Korpus dagegen ist unter [PSA13] erhältlich.

Diese Studie zeigt zum einen, dass Menschen ihren Stil auch objektiv messbar verschleiern und dadurch AA-Systeme täuschen können. Die Teilnehmer konnten ihre Texte *JStylo* gegenüber erfolgreich anonymisieren. Dabei war die Imitation eines anderen Stils erfolgreicher als die Anonymisierung durch lediglich Verbergen des eigenen Stils. Zum anderen zeigt die Studie, dass Stilnachahmung funktioniert. Bei genauerer Betrachtung der Ergebnisse fällt jedoch folgendes auf: Vergleicht man die Schaubilder der AA der echten Texte und der AA der imitierten Texte unter Hinzunahme von Cormac McCarthy als potenziellen Autor, so hat sich die Platzierung der einzelnen Feature-Sets bezüglich ihrer erzielten Genauigkeit gerade umgekehrt. Da eine hohe Genauigkeit bei der Zuordnung zu Cormac McCarthy aber gerade eine hohe Täuschungsanfälligkeit des Feature-Sets bedeutet, ergibt dieser Unterschied wieder Sinn: Er unterstreicht die Qualität von F_3 zur besseren Resistenz gegen Täuschungsversuche. Diese guten Eigenschaften sind vor allem auf die Komplexität von F_3 zurückzuführen. So können die Features darin (z.B. einzelne Buchstabenhäufigkeiten) kaum gegenüber einfacheren Features in F_1 (z.B. durchschnittliche Satzlänge) beeinflusst werden. Diese Auffälligkeiten liefern Forschern wiederum Anhaltspunkte, um die Täuschung der AA zu erkennen. Diesem Thema widmet sich z.B. [ABG12]. Allerdings bedeutet das Erkennen einer Täuschung noch lange nicht die Identifikation des wahren Autors. Eine Anonymisierung durch das intuitive Verbergen oder Verändern von Stil ist also in der Regel gewährleistet.

4.1.1 Wie verändern Menschen ihren Stil

Menschen können eher die Werte der Features in F_1 verändern als die in F_3. Welche Features es genau sind, wurde von den Forschern in [BAG12] ebenfalls herausgearbeitet. Zunächst sagten die Probanden selbst, sie verwendeten bei der Anonymisierung eher kürzere und einfachere Sätze und bei der Imitation von Cormac McCarthy eine beschreibendere und düsterere Sprache. Der statistische Vergleich der einzelnen Feature-Werte

in echten und verfälschten Texten unterstützt diese Aussagen zum Teil: Demnach gab es bei der Anonymisierung vor allem Abnahmen der durchschnittlichen Wort-/Satzlänge und Silbenzahl, aber auch eine Zunahme von Adverbien. Bei der Imitation gab es ebenfalls Abnahmen der durchschnittlichen Wort-/Satzlänge und Silbenzahl, sowie weitere Abnahmen von Adjektiven und Adverben und eine Zunahme von Funktionswörtern. Insgesamt schließen die Forscher, dass sich die Komplexität von verfälschten Texten verringert. Diese Ergebnisse sind mit dem Hinweis zu interpretieren, dass die Thematik der echten Texte nicht vorgegeben war, jedoch die anonymisierten und die imitierten Texte jeweils ein vorgegebenes Thema hatten (Beschreibung der Nachbarschaft, Beschreibung des bisherigen Tagesablaufs). Hier war also die geforderte Trennung des Stils von Inhalt, Kontext und Funktion nicht gewährleistet: Alle gefälschten Texte hatten eine beschreibende Funktion, was die Zunahme von Adverben und Adjektiven aus der Funktion heraus begründet. Bei den Imitaten spielt weiterhin der zu imitierende Autor (also der Kontext) eine wichtige Rolle. Hier wären weitere Studien mit themenunabhängigen Korpora nötig, um allgemeingültige Ergebnisse zu erhalten.

4.1.2 Schwierigkeiten und Grenzen

Die Studie in [BAG12] und viele alltägliche Beispiele zeigen, dass sich Menschen ihres persönlichen Stils durchaus bewusst und darüber hinaus in der Lage sind, diesen zu verbergen oder zu verändern. Sie führen damit in erster Linie andere Menschen in die Irre, können jedoch auch eine Anonymisierung im Hinblick auf Computerprogramme erzielen. Die Imitation eines anderen Autors gelingt ihnen nur in Bezug auf einfache Features gut. AA-Systeme können jedoch solche Imitate durch Benutzung komplexerer Feature-Sets entlarven (z.B. einzelne Buchstabenhäufigkeiten). Trotz alledem bleibt der Autor des Imitats anonym. Gerade bei sehr langen oder einer großen Anzahl von Texten können jedoch Schwierigkeiten auftreten: Einen falschen Stil können nur die wenigsten über längere Zeit konsistent einhalten. Zudem ist das Imitieren bzw. Unterdrücken von Stil äußerst anstrengend. Schwieriger wird es noch, wenn existierende Texte nachträglich anonymisiert werden sollen, ohne dabei die Semantik zu verändern.

4.2 Semi-automatische Täuschungsmethoden

Die Grenzen der menschlichen Täuschungsmethoden wecken den Wunsch nach computergestützter Anonymisierung und Stilverfremdung. Programme zur Anonymisierung von Texten müssen nun noch einen Schritt weiter gehen, sodass sie den Text nach erfolgreicher Bearbeitung nicht mehr zuverlässig klassifizieren können. Ein solches Programm wird im weiteren Verlauf vorgestellt. Zunächst werden jedoch zwei weitere Ansätze erläutert.

4.2.1 Übersetzung und Rück-Übersetzung

Schon im Jahr 2000 entstand die Idee, computergestützte Anonymisierung durch maschinelles Übersetzen in eine andere Sprache und Rück-Übersetzung in die Ausgangssprache

zu realisieren. Solche Übersetzungsdienste sind frei verfügbar (z.B. *Google Translate* oder *Bing Translator*) und einfach zu bedienen. Allerdings haben sie einen eher schlechten Ruf bezüglich der Qualität ihrer durchgeführten Übersetzungen. Daher gilt es zu prüfen, ob die Semantik beibehalten wird und falls ja, ob die Schreibstile ausreichend verändert werden können. Eine Studie über die Auswirkung von Übersetzungen auf die AA wurde in [CG12] durchgeführt. Hier wird zunächst herausgestellt, dass maschinelle Übersetzer Spuren in Texten hinterlassen (*Translator-Effect*), woran sogar verschiedene Übersetzer erkannt werden können. Der Übersetzer wird also wie ein zweiter Autor behandelt, der seinerseits Features im Text hinterlässt. Es stellt sich die Frage, ob diese die Features der menschlichen Autoren verstärken, abschwächen oder ob beide Merkmale ungestört in einem Text koexistieren können. Für die Studie wurden Texte[2] den folgenden Übersetzungsfolgen unterzogen: (`en` $\rightarrow$ `de` $\rightarrow$ `en`), (`en` $\rightarrow$ `ja` $\rightarrow$ `en`), sowie (`en` $\rightarrow$ `ja` $\rightarrow$ `de` $\rightarrow$ `en`), mit `en` = Englisch, `de` = Deutsch und `ja` = Japanisch. Als AA-System diente erneut *JStylo* unter Zuhilfenahme von *Google Translate* und *Bing Translator*[3].

Verschiedene Features wurden auf den übersetzten Texten getestet und aus den besten ein eigenes Feature-Set F_4 = **Translation Feature-Set** zusammengestellt. Dieses beinhaltet die erfolgreichsten Features für die Bestimmung des Übersetzers sowie für die AA. Dazu gehören unter anderem Buchstaben Bi- und Trigramme, Wortlänge, Zeichensetzung und Funktionswörter. Eine Studie auf dem gleichen Korpus mit den gleichen Übersetzern, aber den von *JStylo* standardmäßig angebotenen Feature-Sets ist in [BAG12] zu finden. Detaillierte Ergebnisse der AA seitens *JStylo* und die Relevanz einzelner Features aus F_4 finden sich in [CG12]. Auffällig ist unter anderem der große Unterschied der Relevanz von Funktionswörtern bezüglich der AA und der Bestimmung des Übersetzers. Während fürs Letztere der Wert dieses Features sehr relevant ist, scheint es für die AA nur eine kleinere Rolle zu spielen. Als die Funktionswörter jedoch testweise für die AA aus F_4 ausgeschlossen wurden, verschlechterte sich die Genauigkeit im Schnitt um ca. 15%. Daraus schließen die Forscher, dass solche eher übersetzerspezifischen Features, die durch den Translator-Effect in den Text gelangen, die AA nicht bedeutend stören, sondern im Gegenteil sogar unterstützen können.

Insgesamt zeigen die Ergebnisse eine hohe Genauigkeit bezüglich der AA von durchschnittlich 91,54%. Die Ausgangstexte wurden im Schnitt mit 97,62% klassifiziert, was eine Reduktion durch die Übersetzungsvorgänge von ca. 6% bedeutet. Die Genauigkeit reicht dabei von ca. 77% für die Übersetzung mit zwei Zwischensprachen durch Bing bis hin zu 100% mit der Zwischensprache Japanisch und dem gleichen Übersetzer. Die wichtigste Erkenntnis ist, dass die Übersetzungsvorgänge keine Anonymisierung der Ausgangstexte erreichen. Dadurch, dass die niedrigsten Genauigkeiten mit zwei Zwischensprachen erreicht werden konnten (77,78% durch Google und 87,3% durch Bing), ergibt sich die Vermutung, dass die Hinzunahme weiterer Zwischensprachen die AA ungenauer werden lässt. Davor sollte jedoch ein Blick auf die Sinnerhaltung der bereits getesteten Übersetzungsfolgen geworfen werden. In [BAG12] werden hierzu einige Beispielübersetzungen gegeben. Zunächst eine schlechte Übersetzung am Beispiel: *„Just remember that the things you put into your head are there forever, he said.“*, [McC06]:

[2] Hier wurde der *Brennan-Greenstadt Korpus* ebenfalls verwendet (jedoch ohne die verfälschten Texte).

[3] Beide nutzen das gleiche Übersetzungsverfahren, unterscheiden sich jedoch in ihren Übersetzungen.

en → de → en	*„Remember that the things that you are dead set on always there, he said.“*
en → ja → en	*„But things are there forever remember what you put in your head, he said.“*
en → de → jp → en	*„You are dead, that there always is set, please do not forget what he said.“*

Nun eine gute Übersetzung, aber keine Stil-Veränderung anhand des folgenden Beispielsatzes: *„They passed through the city at noon of the day following.“*, [McC06]:

en → de → en	*„They crossed the city at noon the following day.“*
en → ja → en	*„They passed the city at noon the following day.“*
en → de → ja → en	*„They crossed the city at noon the next day.“*

Beide Studien bewerteten das Täuschungsverfahren als insgesamt ungeeignet, da zum einem die Anonymisierung nicht stark genug ist und zum anderen die Semantik des Textes verfälscht wird. In [BAG12] wird hinzugefügt, dass es durchaus gut übersetzte und anonymisierte Sätze in den Übersetzungen gegeben habe, aber dass diese von im Hinblick auf Sinn oder Grad der Anonymisierung unzureichenden Sätzen dominiert wurden und so das Gesamtergebnis unbrauchbar sei. Auch die Hinzunahme weiterer Zwischensprachen scheint angesichts der Sinnverfälschung nicht vielversprechend. Fortschritte im Bereich der maschinellen Übersetzung könnten jedoch eine Verbesserung des Verfahrens zukünftig ermöglichen. Unklar ist, welche Rolle die gewählten Sprachen für das Verfahren spielen. Mit Deutsch wurde eine dem Englischen ähnliche und mit Japanisch eine vollkommen unähnliche Sprache verwendet.

4.2.2 Eliminierung typischer Wörter

Kacmarcik und Gamon forschten in [KG06] an den *Federalist Papers*. Hierbei handelt es sich um eine Kollektion von 85 Artikeln, die 1788 in den USA anonym veröffentlicht wurden. Mittlerweile sind die Autorschaften der meisten Texte eindeutig geklärt. Demnach sind 5 Artikel von John Jay, 51 von Alexander Hamilton, 14 von John Madison und 3 Artikel wurden von Madison und Hamilton gemeinsam geschrieben. Die Autorschaft der verbleibenden 12 Dokumente ist nicht eindeutig geklärt. 1964 führten jedoch Stil-Analysen der 12 Texte zu der Annahme, John Madison sei ihr Verfasser. [4].

In ihrer Studie testeten Kacmarcik und Gamon das systematische Verfälschen bestimmter Features in diesen 12 Texten. Ziel war dabei, die Autorschaft Madisons mit den gleichen stilometrischen Methoden wie von 1964 nicht mehr nachweisbar zu machen und stattdessen die Texte Hamilton zuzuordnen. Die Forscher konzentrierten sich dabei auf die Angleichung der Anzahl von „unterscheidenden“ Wörtern in den 12 zu klassifizierenden Texten an die Anzahl dieser Wörter in Hamiltons Texten. Sie entwickelten einen Algorithmus (aufgeführt in [KG06]), der unterscheidende Wörter identifiziert und den Fälscher anweist,

[4] Die bisher populärste Errungenschaft auf dem Gebiet der *AA*.

wie bestimmte Worthäufigkeiten zu ändern sind, um eine Zuordnung zu Hamiltons Texten zu erreichen. Leider erzielten sie mit der Anpassung der 10 am meisten unterscheidenden Wörter nicht die gewünschte Anonymisierung. Als Grund dafür fanden sie heraus, dass diese insgesamt zu selten vorkamen, als dass sie einen großen Unterschied hätten erzeugen können. Darum erfolgte der nächste Versuch mit den 10 am meisten unterscheidenden Worten, die zusätzlich eine bestimmte Häufigkeit im Text erreichten. Mit dieser Methode gelang es ihnen, alle 12 Artikel so verändern, dass sie durch das gleiche Verfahren wie von 1964 Hamilton zugeordnet wurden. Zur erfolgreichen Manipulation der Texte benötigten Kacmarcik und Gamon pro 1000 Worte durchschnittlich 14,2 Veränderungen. Sie erreichten damit im Schnitt eine Reduzierung der Wahrscheinlichkeit der Autorschaft Madisons von 96,93% auf nur 12,51%. Kacmarcik und Gamon betonen jedoch, dass ihr Verfahren nur eine „seichte Anonymisierung“ ermöglicht hat, welche komplexeren Feature-Sets leider nicht standhält. Dennoch konnten sie in [KG06] zeigen, dass stilometrische Faktoren computergestützt beeinflusst werden können. Sie halten die Weiterentwicklung dieser Techniken zu umfangreicheren Programmen zur Täuschung der AA für realistisch.

4.2.3 Anonymouth

Anonymouth (erhältlich unter [PSA13]) ist ein Programm zur Anonymisierung von Texten. Es wurde ebenfalls an der Drexel University Philadelphia entwickelt und ist in [MAC+12] ausführlich beschrieben. *Anonymouth* interagiert mit *JStylo* und hat das Ziel, diesen zu täuschen. Als Input werden dabei das zu verfälschende Dokument $\mathcal{D}$, andere Dokumente des gleichen Autors $\mathbb{D}_{same}$ sowie Dokumente anderer Autoren $\mathbb{D}_{others}$ verlangt. Anschließend müssen verschiedene Feature-Sets ausgewählt werden, anhand derer $\mathcal{D}$ analysiert wird. Dabei legen $\mathbb{D}_{others}$ fest, wie die Werte der Features sein sollten und $\mathbb{D}_{same}$ wie sie möglichst nicht sein sollten. *Anonymouth* gibt dann für jedes Feature den Ist- und Soll-Wert an. Für komplexe Feature-Sets findet eine Priorisierung der Features statt, d.h. es werden z.B. nur die 5 Features angezeigt, deren Ist-Wert am meisten vom Soll-Wert abweicht. Darüber hinaus werden die Stellen visualisiert, an denen das Feature in $\mathcal{D}$ auftritt und der Nutzer angewiesen, an einigen dieser Stellen Änderungen vorzunehmen. Anschließend wird $\mathcal{D}$ erneut analysiert. Der Prozess wiederholt sich solange, bis der Nutzer ihn abbricht oder die gewünschte Genauigkeit erreicht ist. In jeder Iteration bekommt der Nutzer das Ergebnis der AA angezeigt. Diese schrittweise Anonymisierung ist notwendig, da sich viele Features gegenseitig beeinflussen (Anzahl der Sätze, durchschnittliche Satzlänge, etc.) und $\mathcal{D}$ daher stets re-klassifiziert werden muss. Die Anonymisierung gilt als erfolgreich, falls der wahre Autor anhand des gewählten Feature-Sets nur noch mit einer kleineren Wahrscheinlichkeit als bei einer zufälligen Zuordnung seitens *JStylo* vorhergesagt werden kann.

Die Entwickler selbst bezeichnen *Anonymouth* nur als ersten Schritt in die Richtung computergestützer Anonymisierung von Texten. *Anonymouth* nimmt wie bereits erwähnt keine vollautomatische Anonymisierung vor, sondern gibt nur Anweisungen. Das Ändern von $\mathcal{D}$ geschieht weiterhin manuell durch den Nutzer, wodurch sich einige Schwächen ergeben. So können dem Nutzer bestimmte Änderungen nicht zugemutet werden (z.B. das Ändern von n-Grammen oder einzelnen Buchstabenhäufigkeiten, vgl. [MAC+12]). Wei-

terhin kann die gegenseitige Beeinflussung vieler Features und die dadurch verbundene iterative Anonymisierung den Vorgang sehr in die Länge ziehen. Eine Benutzbarkeitsstudie in [MAC+12] ergab, dass einige Nutzer ihren Text in 30-60 Minuten anonymisieren konnten, während anderen Nutzern die auf eine Stunde begrenzte Zeit jedoch nicht reichte, um den gewünschten Grad der Anonymisierung zu erzielen. Außerdem verlangt *Anonymouth* vom Nutzer einen Korpus, die Auswahl von Feature-Sets und ein Klassifikator, was wiederum Hintergrundwissen voraussetzt.

5 Zusammenfassung und Ausblick

Die Weiterentwicklung von AA-Systemen ermöglicht zunehmend, öffentlich zugängliche vermeintlich anonyme Texte ihren Autoren zuzuordnen. Dies kann eine Beeinträchtigung der Privatsphäre der Autoren zur Folge haben. Aus diesem Grund wird zunehmend mehr an der Täuschung dieser Systeme geforscht. Verschiedene Ansätze und Studien zu diesem neuen Forschungsgebiet wurden in dieser Arbeit vorgestellt. Die beschriebenen Grenzen haben aufgezeigt, dass menschliche Täuschungsmethoden impraktikabel erscheinen. Dagegen bieten semi-automatische Methoden bessere Ergebnisse im Hinblick auf die Täuschung von AA-Systemen. Aber auch hier zeigen sich Grenzen. Sollen z.B. viele Texte (eventuell sogar simultan) anonymisiert werden, so sind diese Methoden ungeeignet, da hier immer noch der Mensch die Stiländerung selbst durchführen müsste. Vollautomatisierte Täuschungsmethoden stellen eine Alternative dar, die diese Problematik beheben und weitere Einschränkungen aufheben würde. Sie können Stiländerung ohne dem Benutzer durchführen und setzen kein linguistisches Hintergrundwissen voraussetzen.

Eine Möglichkeit zur Realisierung vollautomatisierter Täuschungsmethoden sind Natural Language Watermarking Verfahren, welche beispielsweise in [HSWZ13, Klo14] vorgeschlagen werden. Zwar verfolgen diese einen anderen Zweck (Einbettung verdeckter Nachrichten), ermöglichen jedoch automatisierte Textumformungen, die auf allen sprachlichen Ebenen (phonemisch, morphologisch, lexikalisch, syntaktisch aber auch semantisch) Änderungen durchführen. Diese Änderungen haben stilistische Verzerrungen zur Folge, was wiederum verhelfen kann, AA-Systeme zu täuschen. Oberste Priorität jedoch ist, die Semantik des Textes möglichst weitgehend zu erhalten. Andernfalls wäre der Text nach den Umformungen unlesbar. Im Rahmen zweier Studien zeigte sich in [HSWZ13] mit 89 bzw. in [Klo14] mit 42 Teilnehmern, dass deren vorgeschlagene Natural Language Watermarking Verfahren die Semantik deutschsprachiger Ausgangstexte nur geringfügig verzerrten, sodass die Mehrheit der Teilnehmer die Umformungen nicht signifikant wahrnehmen konnten. In Zukunft gilt es, die Tauglichkeit dieser Verfahren zum Zwecke der Anonymisierung des persönlichen Schreibstils zu analysieren in der Hoffnung, AA-Systeme dadurch erfolgreich täuschen zu können.

6 Danksagung

Diese Arbeit wurde unterstützt vom CASED - Center for Advanced Security Research Darmstadt (www.cased.de), gefördert vom Hessischen Ministerium für Wissenschaft und Kunst unter dem LOEWE-Förderprogramm.

Literatur

[ABG12] S. Afroz, M. Brennan und R. Greenstadt. Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In *Security and Privacy (SP), 2012 IEEE Symposium on*, Seiten 461–475, 2012.

[BAG12] Michael Brennan, Sadia Afroz und Rachel Greenstadt. Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Trans. Inf. Syst. Secur.*, 15(3):12:1–12:22, November 2012.

[CG12] A. Caliskan und R. Greenstadt. Translate Once, Translate Twice, Translate Thrice and Attribute: Identifying Authors and Machine Translation Tools in Translated Text. In *2012 IEEE Sixth International Conference on Semantic Computing (ICSC)*, Seiten 121–125, 2012.

[CH07] Jonathan H. Clark und Charles J. Hannon. A Classifier System for Author Recognition Using Synonym-Based Features. In *Proceedings of the artificial intelligence 6th Mexican international conference on Advances in artificial intelligence*, MICAI'07, Seiten 839–849, Berlin, Heidelberg, 2007. Springer-Verlag.

[Gan08] Prof. Dr. Christina Gansel. Philologische Methoden, Ernst-Moritz-Arndt-Universität Greifswald Philosophische Fakultät (Institut fr Deutsche Philologie), 2008. Letzter Zugriff: 13.02.2014.

[Gol07] Felix Golcher. Ein Einblick in die statistische Stilometrie, 2007.

[Hal12] Oren Halvani. Autorenschaftsanalyse im Kontext der Attributation, Verifikation und intrinsischen Exploration. Master thesis, Technische Universität Darmstadt, 2012.

[HSWZ13] Oren Halvani, Martin Steinebach, Patrick Wolf und Ralf Zimmermann. Natural Language Watermarking for German Texts. In ACM, Hrsg., *Proceedings of the 1st ACM Workshop on Information Hiding and Multimedia Security, June 17-19, 2013 Montpellier, France*, 2013.

[KG06] Gary Kacmarcik und Michael Gamon. Obfuscating Document Stylometry to Preserve Author Anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, Seiten 444–451, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[Klo14] Peter Kloeckner. *Phonemische, Lexikalische und Syntaktische Natural-Language-Watermarking-Verfahren*. Bachelor thesis, Technische Universität Darmstadt, 2014.

[Kot07] S. B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, Seiten 3–24, Amsterdam, The Netherlands, The Netherlands, 2007. IOS Press.

[MAC+12] AndrewW.E. McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman und Rachel Greenstadt. Use Fewer Instances of the Letter i": Toward Writing Style Anonymization. In Simone Fischer-Hbner und Matthew Wright, Hrsg., *Privacy Enhancing Technologies*, Jgg. 7384 of *Lecture Notes in Computer Science*, Seiten 299–318. Springer Berlin Heidelberg, 2012.

[MB11] Rachel Greenstadt Michael Brennan, Sadia Afroz. Deceiving Autorship Attribution. Bericht, Drexel University Philadelphia, 2011.

[McC06] C. McCarthy. *The Road*. Oprah's Book Club. Vintage Books, 2006.

[PSA13] PSAL. Drexel University's Privacy, Security, and Automation Laboratory. JStylo-Anonymouth Webseite: https://psal.cs.drexel.edu/index.php/JStylo-Anonymouth, 2013. Letzter Zugriff: 13.02.2014.

[Sie13] Martin Siefkes. Stil und Gesellschaft - Plädoyer für eine allgemeine Stilistik. 2013.

[Sow73] Bernhard Sowinski. *Deutsche Stilistik*. Fischer-Taschenbücher. Fischer Taschenbuch Verlag, 1973.

7 Anhang

Feature-Kategorie	**Kurzbeschreibung / Beispiele**
Interpunktionszeichen	(,), [,], !, ?, ;, :, . . .
Buchstaben	A-Z, Ä, Ö, Ü, a-z, ä, ö, ü, ß
Buchstaben n-Gramme	Textbeispiel $\xrightarrow{n=2}$ {Te, ex, xt, tb, be, . . .}, $\xrightarrow{n=3}$ {Tex, ext, xtb, tbe, . . .}, . . .
Präfixe	**Text**beispiel (Vorsilbe)
Infixe	Text**bei**spiel (innerer Wortbestandteil)
Suffixe	Textbei**spiel** (Nachsilbe)
Funktionswörter	Artikel (der, das, einer, eines, . . .), Konjunktionen (und, oder, . . .), . . .
Anglizismen	Wortentlehnungen (z.B. Mail, Newsletter, Chat, Meeting, Update, . . .)
Neologismen	Kunstwörter (z.B. Abmahnwelle, Nerd, googeln, verschlimmbessern, . . .)
Wort n-Gramme	Ein kleines Textbeispiel $\xrightarrow{n=2}$ {(Ein kleines), (kleines Textbeispiel)}
Kollokationen	Häufig vorkommende Wortverbindungen (z.B. *starker Tobak*)
Wortarten	Adjektive, Interjektion, Numerale, Substantive, . . .
Wortart n-Gramme	(Artikel-Adjektiv-Nomen), (Pronomen-Nomen-Artikel), . . .
Phrasen/Redewendungen	Redensarten (z.B. *aus dem Nähkästchen plaudern*)
Satz-Anfänge/Endungen	Satzanfang(Nomen), Satzende(finites Verb), . . .
Wort-Komplexität	Wörter bestimmter Länge, Wörter mit x Vokalen
Satz-Komplexität	Sätze bestimmter Länge, Vorfeld/Mittelfeld/Nachfeld-Komplexitäten, . . .
Text-Komplexität	Funktionswort-Dichte, Koreferenzketten, . . .
Verständlichkeits-Indizes	*Gunning Fog Readability Index*, *Flesch-Kincaid Reading Ease*, . . .
Grammatikalische Fehler	Falsche Verwendung von Genus, Kasus, Kommata, . . .

Tabelle 1: Eine Auswahl von 20 Feature-Kategorien, teilweise aus [Hal12] entnommen.

Feature-Set	**Enthaltene Features**
F_1 = *9 Feature-Set*	Enthält unter anderem x = die Anzahl nur einmal vorkommender Wörter, Verhältnis von x zur Anzahl aller Wörter, durchschnittliche Silbenzahl pro Wort, durchschnittliche Satzlänge, Anzahl von Zeichen, Buchstaben und Sätze sowie *Gunning Fog Readability Index* und *Flesch-Kincaid Readability Ease*, [BAG12].
F_2 = *Synonym-Based Feature-Set*	Enthält die Anzahl der möglichen Synonymen von Wörtern. Je mehr Synonyme existieren, desto charakteristischer ist das Wort für den Stil. Berücksichtigt wird dabei außerdem die jeweilige Worthäufigkeit im zu testenden Text und in der Vergleichsbasis, [CH07].
F_3 = *Writeprint Feature-Set*	Häufigkeiten von spezifischen Zeichen/Symbolen, Interpunktionszeichen, Ziffern, Buchstaben, Wörtern, Funktionswörtern sowie Wortarten. Weiterhin: Anzahl aller Zeichen, kurzer Worte sowie aller Wörter. Prozentualer Anteil von Ziffern, großgeschriebenen Buchstaben sowie gängigen Zeichen Bi- und Trigrammen. Verhältnis von Hapax und Dis legomena (Wörter, die nur einmal bzw. zweimal in einem Text vorkommen), sowie durchschnittliche Wortlänge, [BAG12].

Tabelle 2: Die drei relevantesten Feature-Sets im Rahmen dieser Arbeit.