Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Mediensysteme

Paarweise Autorenschaftsverifikation von kurzen Texten

Masterarbeit

Fabian Loose

1. Gutachter: Prof. Dr. Benno Stein
Betreuer: Martin Potthast

Datum der Abgabe: 17.05.2011

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weimar, den 17.05.2011

.....

Fabian Loose

Zusammenfassung

In der vorliegenden Arbeit wird die paarweise Autorenschaftsverifikation von kurzen Texten untersucht. Dabei ist ein Textpaar (d_1, d_2) gegeben, und die Aufgabe besteht darin, zu entscheiden, ob beide Texte den selben Autor haben. Diese Aufgabe wird als zentrales Problem bei der Analyse fragwürdiger Autorenschaft identifiziert.

In [12] haben Koppel und Schler mit Unmasking ein Verfahren vorgestellt, das bei der paarweisen Autorenschaftsverifikation von langen Texten sehr gute Ergebnisse erzielt. In der vorliegenden Arbeit wird *ST-Unmasking*, ein neues Verfahren zur paarweisen Autorenschaftsverifikation von kurzen Texten vorgestellt, dass auf den Ideen von Koppel und Schler aufbaut. Dazu werden die Texte d_1 und d_2 anhand ihres Schreibstils repräsentiert. Für diese Repräsentation werden bekannte und neu entwickelte Stilmerkmale verwendet. Die Entscheidung, ob ein Textpaar (d_1, d_2) von einem einzigen Autor geschrieben wurde oder nicht, wird anhand verschiedener Ähnlichkeitsberechnungen zwischen den Stilmerkmalen von d_1 und d_2 entschieden. Dazu wird durch maschinelles Lernen ein Klassifikator trainiert, der Textpaare einer der Klassen *selber-Autor* oder *verschiedene-Autoren* zuordnet.

ST-Unmasking steigert bei einer Textlänge von 2 500 Wörtern die Anzahl richtig erkannter Textpaare um 14.5 Prozentpunkte auf 78% gegenüber Unmasking. Die besten neu entwickelten Stilmerkmale erreichen in der Evaluierung eine Trennschärfe, die im Bereich der besten bekannten Stilmerkmale liegt.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Es gibt zahlreiche Situationen, in denen die Autorenschaft eines Textes unklar ist oder in Frage gestellt wird. Literaturwissenschaftler und Historiker untersuchen Bücher oder andere Texte, bei denen die Kennzeichnung der Autorenschaft fehlt oder zweifelhaft ist [3, 9, 12]. Kriminologen führen forensische Untersuchungen durch, um die Authentizität von Bekennerschreiben oder Abschiedsbriefen sicherzustellen [17]. Lehrer, Dozenten und Wissenschaftler stehen häufig vor der Aufgabe, Plagiate erkennen zu müssen. In all diesen Fällen können Verfahren der automatischen, computergestützten Autorenschaftsverifikation wichtige Anhaltspunkte zur endgültigen Lösung des jeweiligen Problems liefern.

Formal lässt sich die Aufgabenstellung der paarweisen Autorenschaftsverifikation wie folgt darstellen: Gegeben sind zwei Texte d_1 und d_2, und die Aufgabe besteht darin, zu entscheiden, ob beide Texte den selben Autor haben. Dazu werden d_1 und d_2 als Computerrepräsentationen $\mathbf{d_1}$ und $\mathbf{d_2}$ dargestellt. Mit Hilfe einer Berechnungsvorschrift $c(\mathbf{d_1}, \mathbf{d_2})$ soll dann entschieden werden, ob d_1 und d_2 den selben Autor haben.

Die Abbildung eines Textes d in seine Repräsentation $\mathbf{d}$ geschieht mit Hilfe von statistischen Stilmerkmalen. Grundlage dafür ist die Annahme, dass jeder Autor einen eigenen Schreibstil hat, der sich messbar von denen anderer Autoren unterscheidet. Stilmerkmale quantifizieren den Schreibstil von Texten. Das wissenschaftliche Gebiet, welches sich mit der der Erforschung und Quantifizierung von Stilmerkmalen beschäftigt, wird als Stilometrie bezeichnet. In

Kapitel 3 werden die wichtigsten Stilmerkmale aus der Stilometrie vorgestellt. Ein Beitrag dieser Arbeit sind die ebenfalls in Kapitel 3 vorgeschlagenen, neu entwickelten Stilmerkmale.

Für die Realisierung einer geeigneten Berechnungsvorschrift $c(\mathbf{d_1}, \mathbf{d_2})$ haben sich in der aktuellen Forschung Verfahren des maschinellen Lernens durchgesetzt [5, 14, 17, 19]. In Kapitel 4 werden die hierfür relevanten Grundlagen des maschinellen Lernens erklärt. Darauf aufbauend wird in Kapitel 5 das von Koppel und Schler [12] vorgeschlagene Verfahren *Unmasking*, eines der derzeit leistungsfähigsten Verfahren zur Autorenschaftsverifikation vorgestellt. Mit Unmasking werden sehr gute Ergebnisse auf langen Texten (Länge eines Buches, rund 20000 Wörter) erzielt. Für kürzere Texte (eine A4-Seite, ca. 250-500 Wörter) existieren bislang noch keine zufriedenstellenden Verfahren. Aufbauend auf den Ideen des Unmaskings wird in dieser Arbeit die Autorenschaftsverifikation von kurzen Texten untersucht. Wichtigster Beitrag dieser Arbeit ist das in Kapitel 6 vorgestellte Verfahren *Short Text Unmasking (ST-Unmasking)* zur paarweisen Autorenschaftsverifikation von kurzen Texten. Die in Kapitel 3 vorgestellten Stilmerkmale sowie ST-Unmasking werden in Kapitel 7 evaluiert.

Kapitel 2

Verwandte Arbeiten

In diesem Kapitel wird die paarweise Autorenschaftsverifikation in den Kontext der aktuellen Forschung eingeordnet. Die Gliederung geht vom allgemeinen ins spezielle, beginnend mit der Autorenschaftsattribution als der Autorenschaftsverifikation übergeordneter Problemstellung. Die paarweise Autorenschaftsverifikation wird als zentrales Problem identifiziert, insbesondere im Zusammenhang mit kurzen Texten.

2.1 Autorenschaftsattribution

Die Mehrzahl der Forschungsarbeiten im Bereich der Analyse der Autorenschaft von Texten beschäftigt sich mit Autorenschaftsattribution. In der Autorenschaftsattribution wird ein anonymer Text einem von mehreren bekannten Autoren zugeordnet. Formal kann Autorenschaftsattribution wie folgt beschrieben werden: Gegeben ist eine Menge M von n Autoren $M = \{A_1, \ldots A_n\}$, für jeden Autor A_i ist ein hinreichend großes Textbeispiel[1] D_i vorhanden. Weiterhin ist ein anonymer Text d gegeben. Es ist die Aufgabenstellung, zu bestimmen, welcher der Autoren in M der Verfasser von d ist. Dabei besteht die

[1] Ein hinreichend großes Textbeispiel steht hier für eine Textmenge, die die stilistischen Gewohnheiten eines Autors reflektiert wie es Holmes in [8] für Autorenschaftsattribution allgemein als wesentlich beschrieben hat. In Graham et al. [6] wird von einer angemessen großen („reasonably large“) Textmenge gesprochen, die für eine statistische Signifikanz des berechneten Schreibstils erforderlich ist. Diese Textmenge kann aus einem großen oder mehreren kleineren Texten stammen.

Forderung, dass der wahre Autor von d in M enthalten ist. Autorenschaftsattribution wird als (Mehrklassen-)Klassifikationsproblem verstanden: Ordne d einer der bekannten Klassen $A_i \in M$ zu. Für diese Klassifikation haben sich Methoden des maschinellen Lernens durchgesetzt [7, 17]. Der Vorteil von maschinellem Lernen liegt insbesondere darin, dass entsprechende Verfahren mit einer sehr großen Anzahl an Stilmerkmalen umgehen können, selbst dann noch, wenn nur wenige davon tatsächlich für die Klassifikation von Nutzen sind. So werden zum Beispiel in [5] 70000 Stilmerkmale verwendet, um Autorenschaftsattribution mit deutschen Zeitungsartikeln durchzuführen. Als besonders geeigneter Klassifikator des maschinellen Lernens hat sich dabei die Support-Vector-Machine[2] herausgestellt. In [1] sind britische und amerikanische Romane Gegenstand der Untersuchungen. Auch hier wird eine Support-Vector-Machine verwendet, um die Autorenschaftsattribution durchzuführen. Es werden 200 bzw. 500 Stilmerkmale benutzt.

Problematisch ist die Anwendung der Autorenschaftsattribution in dem Fall, wenn der wahre Autor eines anonymen Textes d nicht in der Menge M der bekannten Autoren ist. Jedoch existieren Ansätze die dieses Problem angehen. Zum Beispiel wird in [14] im Falle, dass keine sichere Entscheidung getroffen werden kann, keine Zuordnung von d vorgenommen.

2.2 Autorenschaftsverifikation

Autorenschaftsverifikation kann im weiteren Sinne als eine spezielle Form der Autorenschaftsattribution mit nur einem Kandidaten A, und somit der Menge $M = \{A\}$ betrachtet werden. Im Unterschied zur Autorenschaftsattribution ist dabei immer zu beachten, dass auch eine Entscheidung gegen die Klasse A möglich sein muss. Ein anonymer Text d wird demnach entweder dem Autor A zugeordnet oder nicht, wenn keine sichere Entscheidung für A getroffen werden kann. Wie Autorenschaftsverifikation genau zu definieren ist, darüber besteht in der aktuellen Forschung kein klarer Konsens [10]. In [20] beispielsweise wird die Autorenschaftsverifikation in drei Problemklassen organisiert, die sich aus verschiedenen Anwendungsszenarien ergeben. Jedoch wird darauf hingewie-

[2]Die Support-Vector-Machine (SVM) wird in Kapitel 4 genauer erklärt.

sen, dass alle drei Probleme ineinander überführt werden können. Diese und die meisten in der Forschung verwendeten Definitionen der Autorenschaftsverifikation (nachfolgend AV) lassen sich formal wie folgt zusammenfassen: Gegeben ist ein hinreichend großes Textbeispiel[3] D eines Autors A und ein anonymer Text d. Die Aufgabe ist es, zu entscheiden ob A auch der Autor von d ist. Damit kann AV als Einklassen-Klassifikationsproblem verstanden werden: Ordne d bei ausreichender Konfidenz der Klasse A zu. Aufgrund der Struktur vieler Anwendungsfälle wird häufig davon ausgegangen, dass das Textbeispiel D deutlich größer ist als der anonyme Text d. Damit liegt die Herangehensweise, aus D ein Autorenprofil zu berechnen und mit Hilfe dieses Profils den Text d auf Klassenzugehörigkeit zu A zu prüfen, nahe, wie die folgenden Beispiele illustrieren: In [7] wird die AV mit studentischen Arbeiten von acht Autoren untersucht. Für jeden Autor sind neun Texte vorhanden, von denen je acht zum Textbeispiel D zusammengefasst werden. Der neunte Text wird, als d, der AV unterzogen. Es werden sehr gute Ergebnisse berichtet. An einer Universität ist es durchaus realistisch, von einem Studierenden A mehrere Arbeiten als Textbeispiel D zur Verfügung zu haben, um ihn als Autor einer fragwürdige Arbeit d zu verifizieren. Problematisch ist allerdings, wenn nicht sichergestellt ist, dass A tatsächlich der Autor aller Arbeiten in D ist. Ein weiteres Beispiel ist die intrinsische Plagiatanalyse, als Anwendungsfall der Autorenschaftsverifikation. Hier wird folgendes Problem untersucht: Gegeben ist ein angeblich von einem einzigen Autor A verfasster Text D. Untersucht wird die Frage, ob D Abschnitte $d_1, \ldots d_n \in D$ enthält, die nicht von A verfasst wurden [20]. Dazu wird der Schreibstil jedes Abschnitts d_i mit dem durchschnittlichen Schreibstil von D verglichen. Kann der Autor von D nicht als der Autor von d_i verifiziert werden, wird d_i als potentiell plagiiert betrachtet. Auch dieser Ansatz kann problematisch sein, wenn die Anzahl potentiell plagiierter Abschnitte sehr groß ist, wodurch der durchschnittliche Schreibstil in D nicht mehr repräsentativ für A ist.

[3]Ein hinreichend großes Textbeispiel ist hier ebenso definiert, wie für die Autorenschaftsattribution.

2.3 Paarweise Autorenschaftsverifikation

Mit der paarweisen Autorenschaftsverifikation (nachfolgend pAV) wird untersucht, ob zwei einzelne Texte d_1 und d_2 den selben Autor haben. Die pAV kann als Kernproblem der AV angesehen werden, aber auch als vereinfachende Umformulierung. Eine Überführung von AV nach pAV ist jederzeit durch Konkatenation aller Einzeltexte im Textbeispiel D möglich, die Umkehrung gilt speziell im Zusammenhang mit kurzen Texten nicht. Nicht immer kann davon ausgegangen werden, dass ein großes Textbeispiel D vorhanden ist. Speziell in diesen Fällen ist pAV sinnvoll. Das ist zum Beispiel bei Webkommentaren der Fall, wo viele unterschiedliche Autoren relativ kurze Texte abgeben. Aber auch dann, wenn nicht sicher gestellt ist, dass der gesamte Text in D tatsächlich (oder zumindest überwiegend) von einem einzigen Autor A verfasst wurde, wie es in den Beispielen des vorigen Abschnitts bereits angesprochen wurde.

Die pAV betrachtet beide Texte d_1 und d_2 als gleichwertig (und ggf. gleich lang) und legt damit eine Trennung in Autorenprofil und fragwürdigen Text nicht nahe. Tatsächlich ist das konkrete Profil (der konkrete Schreibstil) des Autors von d_1 oder d_2 für die Verifikation nicht von Belang, interessant ist nur die Frage nach gleicher oder verschiedener Autorenschaft.

Dadurch, dass prinzipbedingt an die beiden Texte d_1 und d_2 keine Anforderungen bezüglich der Länge gestellt werden, ist die pAV zentral bei der Autorenschaftsverifikation von kurzen Texten. Texte werden in der vorliegenden Arbeit dann als kurz angesehen, wenn die Forderung von Holmes [8] an einen Text, die stilistischen Gewohnheiten des Autors zu reflektieren, aufgrund zu geringer Datenmenge nicht oder zumindest nicht vollständig erfüllt sein kann.[4]

Es existieren nur einige wenige Arbeiten, zum Beispiel [6, 16], die sich dem Thema paarweiser Autorenschaftsverifikation angenommen haben, jedoch stellen diese aufgrund eingeschränkter Verallgemeinerbarkeit bzw. nicht exakt nachvollziehbarer Vorgehensweise keine zufriedenstellende Lösung des pAV-Problems dar.

Im Papier von Koppel und Schler [12] wird mit Unmasking ein Verfahren vorgestellt, dass für paarweise Autorenschaftsverifikation von langen Texten

[4] In der Forschung gibt es keinen klaren Konsens, in welchem Bereich diese Grenze liegt, in der vorliegenden Arbeit wird die notwendige Anzahl von Wörtern auf $5000-10000$ geschätzt.

verwendet werden kann. Dieses wird in Kapitel 5 detailliert besprochen. Besonders interessant an Unmasking ist, dass das Einklassen-Klassifikationsproblem AV als Zweiklassen-Meta-Klassifikationsproblem formuliert wird, indem ein Textpaar (d_1, d_2) einer der Klassen *selber-Autor* bzw. *verschiedene-Autoren* zugeordnet wird. Das ist insbesondere deshalb von Interesse, da Einklassen-Klassifikation ein ungleich härteres Problem als Mehrklassen-Klassifikation ist, wie auch die Experimente in [12] zeigen.

In der vorliegenden Arbeit wird pAV auf kurzen Texten, basierend auf den Ideen von Koppel und Schler, insbesondere der Meta-Klassifikation, untersucht.

Kapitel 3

Stilmerkmale

Jeder Autor hat seinen eigenen Schreibstil. Um diesen Schreibstil zu messen, wird im Forschungsgebiet der Stilometrie nach geeigneten Stilmerkmalen gesucht. Stilmerkmale haben gemeinsam, dass sie aus einem Text relativ einfach durch Zählen, Durchschnittsbildung etc. berechnet werden können. Ein einfaches Beispiel für ein Stilmerkmal ist die durchschnittliche Satzlänge eines Textes. In der aktuellen Forschung haben sich zwei Stilmerkmale als besonders geeignet erwiesen: die Häufigkeitsverteilung von Funktionswörtern auf der einen Seite und die Häufigkeitsverteilung von Buchstaben-N-Grammen auf der anderen Seite. Beide werden im folgenden Abschnitt näher erläutert. Zunächst wird jedoch der Frage nachgegangen, welche Anforderungen an Stilmerkmale gestellt werden. Zwei Eigenschaften sind dabei zentral:

- *Trennschärfe.* Stilmerkmale sollen den Schreibstil verschiedener Autoren möglichst gut differenzierbar machen.
- *Unabhängigkeit vom Inhalt.* Ähnlichkeiten der Stilmerkmale sollten unabhängig von Inhalt, Genre, Entstehungszeit (-epoche) und ähnlichen Variationen sein. Das ist aus zwei Gründen wichtig: Einerseits sollen inhaltsgleiche Texte zweier unterschiedlicher Autoren sicher getrennt werden, andererseits sollen Stilmerkmale robust gegen Veränderungen im Inhalt, Genre etc. über mehrere Texte ein und desselben Autors sein.

Je nach Anwendungsfall können weitere Anforderungen an Stilmerkmale gestellt werden. Beispielhaft seien an dieser Stelle Sprachunabhängigkeit und

schnelle Berechenbarkeit genannt.

Stilmerkmale werden üblicherweise als Vektoren repräsentiert, dabei steht jede Merkmalsausprägung für eine Dimension des Vektors, alle möglichen Merkmalsausprägungen zusammen spannen den Merkmalsraum auf. Die Repräsentation $\mathbf{d}$ eines Textes d durch Stilmerkmale wird auch als Stilvektor bezeichnet.

Alle in dieser Arbeit verwendeten Stilmerkmale haben gemeinsam, dass sie die Häufigkeitsverteilung von *Elementen* im Text darstellen und variieren nur in der Wahl und Anzahl dieser Elemente. Die folgenden beiden Abschnitte stellen bekannte und neue Elemente, respektive Stilmerkmale vor.

3.1 Bekannte Stilmerkmale

In diesem Abschnitt werden Stilmerkmale vorgestellt, die sich in der Forschung als gut geeignet erwiesen haben oder die die Grundlage der Neu- oder Weiterentwicklungen der im nächsten Abschnitt vorgestellten Stilmerkmale sind.

- *Funktionswörter.* Funktionswörter sind diejenigen Wörter in einem Text, die sehr häufig vorkommen und die gleichzeitig keine (oder kaum) inhaltliche Information tragen. Beispiele dafür sind Artikel, Konjunktionen oder Präpositionen. Funktionswörter nehmen einen großen Prozentsatz eines Textes ein (je nach genauer Definition und Sprache etwa $40-60\%$). Es wird davon ausgegangen, dass die Verwendung von Funktionswörtern nicht bewusst beeinflusst werden kann [1, 13, 17]. Die Bestimmung der Funktionswörter eines Textes kann auf zwei unterschiedliche Arten geschehen. Funktionswörter können aus einer im Voraus erstellten sprachspezifischen Liste (Stoppwörterbuch) stammen oder für jeden Text individuell berechnet werden. Bei der individuellen Berechnung werden die h häufigsten Wörter eines Textes als Funktionswörter benutzt. In der Forschung werden Größen für h im Bereich von $50-1000$ genannt [3, 17]. Insbesondere bei kurzen Texten, wenn h in der Größenordnung der Textlänge liegt, ist die Verwendung eines Stoppwörterbuchs vorzuziehen.

  Beispiele für Funktionswörter sind: *der, von, ist.*

- *Zeichen-N-Gramme.* Ein Zeichen-N-Gramm besteht aus N aufeinander folgenden Zeichen. Aufeinander folgende N-Gramme überlappen immer um $N-1$ Zeichen. Dabei können Satzzeichen, Zahlen, Leerzeichen, Zeilenumbrüche etc. in den N-Grammen enthalten sein oder auch nicht. In der aktuelle Forschung wird für N meist 3 oder 4 gewählt (Trigramme oder 4-Gramme). Dabei wird davon ausgegangen, dass insbesondere Trigramme relativ wenig Inhalt transportieren und somit die Häufigkeitsverteilung aller, also nicht nur der h häufigsten, Trigramme benutzt werden kann.

  Die Zeichen-Trigramme aus *„Ein Beispiel“* sind: *Ein, in_, n_B, _Be, Bei, eis, isp, spi, pie, iel.*

- *Zeichen-K-Skip-N-Gramme.* Zeichen-K-Skip-N-Gramme sind Zeichen-N-Gramme, bei denen zwischen jedem Zeichen K Zeichen übersprungen werden.

  Für die Zeichenkette *„abcdefg“* lauten die Zeichen-1-Skip-3-Gramme: *ace, bdf, ceg.*

- *Part-of-Speech.* Part-of-Speech (nachfolgend POS) sind die Wortklassen einer Sprache. Die Häufigkeitsverteilung der Wortklassen eines Textes wird ebenfalls als autorenspezifisch angesehen. Weiterhin kann davon ausgegangen werden, dass sich der Inhalt eines Textes nicht in der Verteilung der Wortklassen widerspiegelt.

  Die POS-Tags aus dem Satz *„Das ist ein Beispiel.“* lauten: *Artikel, Verb, Artikel, Substantiv.*

- *POS-N-Gramme.* Anders als zuvor wird nicht die Häufigkeitsverteilung der Wortklassen einzeln betrachtet, sondern aufeinander folgende POS-Tags werden zu N-Grammen zusammengefasst.

  Für obiges Beispiel lauten die POS-Trigramme:
  Artikel_Verb_Artikel, Verb_Artikel_Substantiv.

- *Wortlängen.* Die Häufigkeitsverteilung von Wortlängen ist ein Stilmerkmal, welches bereits im vorletzten Jahrhundert in Erwägung gezogen

wurde [8]. Für sich alleine betrachtet ist es anderen Stilmerkmalen deutlich unterlegen. Im nächsten Abschnitt wird es jedoch Grundlage für ein erweitertes Stilmerkmal sein.

Für obigen Beispielsatz sind die Wortlängen: *3, 3, 3, 8*.

3.2 Neue und erweiterte Stilmerkmale

- *Phoneme.* Phoneme sind die Laute einer Sprache. Ein Phonem kann dabei auf mehrere verschiedene Arten geschrieben werden, ebenso können gleiche Buchstaben oder Buchstabenkombinationen je nach Kontext verschiedene Phoneme ergeben. Dementsprechend können Phoneme nicht direkt aus der Zeichenfolge in einem Text berechnet werden, vielmehr wird ein Wörterbuch benötigt. Jedem Wort in diesem Wörterbuch ist seine Folge von Phonemen zugeordnet. Somit können nur Wörter betrachtet werden, die in dem Wörterbuch enthalten sind.

  Die Phoneme aus *„Ein Beispiel“* lauten: *a, ɪ, n, ’b, a, ɪ, ʃ, p, i:, l*.

- *Phonem-N-Gramme.* Phonem-N-Gramme sind N-Gramme, die aus N konsekutiven Phonemen gebildet werden.

  Für obiges Beispiel lauten die Phonem-Bigramme: *aɪ, ɪn, n’b, ’ba, aɪ, ɪʃ, ʃp, pi:, i:l*.

- *Vokal-N-Gramme.* Vokal-N-Gramme sind N-Gramme, die aus aufeinander folgenden Vokalen gebildet werden. Alle Nicht-Vokale (Konsonanten, Leerzeichen, Satzzeichen, Zahlen) werden übersprungen.

  Die Vokal-Trigramme aus *„Das ist ein Beispiel“* sind: *aie, iei, eie, iei, eii, iie*.

- *Vokal-Konsonant-N-Gramme.* Diese N-Gramme entsprechen prinzipiell den Zeichen-N-Grammen, jedoch wird jeder Vokal als v und jeder Konsonant als k dargestellt. Dadurch wird es möglich, größere Werte für N einzusetzen, ohne dass dadurch Inhalt transportiert wird.

Für *„ein Beispiel“* lauten die Vokal-Konsonant-6-Gramme: *vvkkvv, vkkvvk, kkvvkk, kvvkkv, vvkkvv, vkkvvk.*

- *Wortlängen-N-Gramme.* In Wortlängen-N-Grammen werden die Längen aufeinander folgender Wörter zu N-Grammen zusammengefasst. Damit steigt die Qualität dieses Merkmals gegenüber einfacher Wortlängen-Häufigkeitsverteilung an.

  Die Wortlängen-Bigramme aus *„Ein ganz neuer Satz“* sind: *3_4, 4_5, 5_4.*

- *K-Präfix-N-Gramme.* Die ersten K Zeichen eines Wortes bilden mit den ersten K Zeichen von N aufeinander folgenden Wörtern ein K-Präfix-N-Gramm. Dieses Merkmal kann zum Beispiel im Deutschen eine Präferenz für die Zeitformen Perfekt oder Plusquamperfekt messen.

  1-Präfix-Bigramme aus obigen Satz sind: *E_g, g_n, n_S.*

- *K-Suffix-N-Gramme.* Die letzten K Zeichen eines Wortes bilden mit den letzten K Zeichen von N aufeinander folgenden Wörtern ein K-Suffix-N-Gramm. Im Deutschen kann damit zum Beispiel die Präferenzen eines Autors bezüglich des Genitivs erfasst werden.

  Die 2-Suffix-Bigramme aus obigem Satz lauten: *in_nz, nz_er, er_tz.*

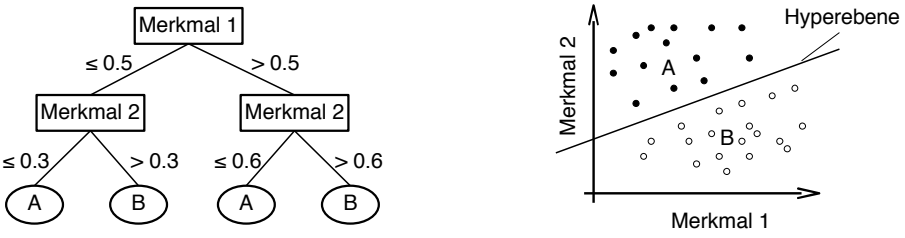
Kapitel 4

Maschinelles Lernen und Autorenschaft

In der aktuellen Forschung zur Autorenschaftsverifikation spielen Verfahren des maschinellen Lernens, speziell der automatischen Klassifikation eine zentrale Rolle. In diesem Kapitel werden die hierfür wichtigen Begriffe der automatischen Klassifikation erklärt und Besonderheiten bezüglich der Autorenschaftsverifikation aufgezeigt.

4.1 Automatische Klassifikation

Die automatische Klassifikation erfüllt die Aufgabe, ein Objekt o in eine von n bekannten Klassen $K_1, \ldots, K_n$ einzuordnen. Dazu wird o als Computerrepräsentation $\mathbf{o}$ in einem Vektorraum dargestellt. Der Vektorraum wird als Merkmalsraum und $\mathbf{o}$ als Merkmalsvektor bezeichnet. Die Repräsentation $\mathbf{o}$ trägt für die Klassifikation relevante Eigenschaften, jede Eigenschaft ist eine Dimension des Merkmalsraums. Im Kontext der Autorenschaftsanalyse ist o ein Text und $\mathbf{o}$ seine Repräsentation als Stilvektor. Die Zuordnung von $\mathbf{o}$ zu K_i geschieht durch einen Klassifikator $c(\mathbf{o}) \rightarrow \{K_1 \ldots K_n\}$. Der Klassifikator ist eine Funktion, die anhand einer Menge von vorklassifizierten Beispielobjekten (Trainingsmenge) gelernt wird. Beispiele für Klassifikatoren sind der Entscheidungsbaum, der Random Forest und die Support Vector Machine (SVM):



(a) Beispiel für einen Entscheidungsbaum.

(b) Veranschaulichung einer SVM.

Abbildung 4.1: Die Klassifikatoren Entscheidungsbaum und SVM.

- *Entscheidungsbaum.* In einem Entscheidungsbaum sind aufeinander folgende Entscheidungen hierarchisch strukturiert. An jedem Knoten wird über eine Dimension des Merkmalsraums entschieden und der Merkmalsraum bezüglich dieser Dimension in disjunkte Teilräume geteilt. Jedes Blatt ist mit einer der möglichen Klassen gekennzeichnet. Jedes zu klassifizierende Objekt wird anhand seiner Merkmalsausprägungen entlang eines Pfades von der Wurzel bis zu einem Blatt derjenigen Klasse zugeordnet, mit der das Blatt gekennzeichnet ist. Entsprechend liegt die Vektorrepräsentation dieses Objektes in dem Teilraum, der durch die Iterative Teilung des Merkmalsraums entlang des Pfades von der Wurzel zum Blatt beschrieben wird. Abbildung 4.1(a) zeigt exemplarisch einen Entscheidungsbaum, der anhand zweier Merkmale Objekte den Klassen A und B zuordnet.

- *Random Forest.* Der Random Forest klassifiziert Objekte anhand einer Mehrheitsentscheidung über viele unterschiedlich konstruierte Entscheidungsbäume [4].

- *SVM.* Die SVM ist ein Klassifikator, der in den durch die Merkmale aufgespannten Merkmalsraum Hyperebenen einfügt, die den Merkmalsraum in disjunkte Teilräume aufteilt. Jeder Teilraum ist mit einer der möglichen Klassen gekennzeichnet. Die Hyperebenen werden in der Trainingsphase dergestalt in den Merkmalsraum eingefügt, dass die nächsten

Merkmalsvektoren der angrenzenden Klassen den maximalen Abstand zur Hyperebene haben. Diese Merkmalsvektoren werden als Stützvektoren bezeichnet und sind ausreichend um die Hyperebene mathematisch zu beschreiben. Alle anderen Merkmalsvektoren der Trainingsmenge müssen nicht gespeichert werden, so dass die Größe der instanziierten SVM unabhängig von der Anzahl an Trainingsobjekten ist. In Abbildung 4.1(b) ist eine SVM für einen 2-dimensionalen Merkmalsraum grafisch veranschaulicht. Die Hyperebene liegt zwischen Instanzen der Klassen A und B.

Im Allgemeinen wird ein Klassifikator Fehler machen, also nicht jedes Objekt korrekt zuordnen. Soll ein Klassifikator für eine bestimmte Aufgabe eingesetzt werden, ist es wichtig, die Art und Menge der Fehler zu kennen, die der Klassifikator macht. Nur dann ist es möglich, die Klassifikationsergebnisse zu interpretieren. Um die Qualität eines trainierten Klassifikators zu testen, wird eine Testmenge benötigt. Für alle Objekte der Testmenge ist, analog zur Trainingsmenge, die Klassenzugehörigkeit bekannt. Nach Anwendung des Klassifikators kann die Übereinstimmung der berechneten mit der tatsächlichen Klassenzugehörigkeit bestimmt werden. Es existieren mehrere gebräuchliche Maße, um die Qualität von c zu beurteilen, in dieser Arbeit wird dazu die Korrektklassifikationsrate (Klassifikationsgüte) kkr verwendet. Diese berechnet sich wie folgt:

$$kkr = \frac{Anzahl\ richtig\ klassifizierter\ Objekte}{Anzahl\ aller\ Objekte}$$

Ist keine Testmenge vorhanden, so kann die Güte eines Klassifikators dennoch ermittelt werden. Dazu wird häufig die 10-fache Kreuzvalidierung verwendet. Dabei wird die Trainingsmenge in 10 gleichgroße Teilmengen geteilt. Jede der 10 Teilmengen wird einmal als Testmenge verwendet, wobei die übrigen 9 Teilmengen zusammen als Trainingsmenge benutzt werden. Daraus ergeben sich 10 Trainings-/Testdurchläufe. Als gesamte Klassifikationsgüte wird der Mittelwert der 10 einzelnen Klassifikationsgüte-Berechnungen ausgegeben [11].

4.2 Klassifikation von Autoren

Als Grundlage für die folgenden Kapitel wird in diesem Abschnitt die automatischen Klassifikation anhand der Autorenschaftsattribution und Autorenschaftsverifikation erklärt. Dabei werden die vormals eingeführten Begriffe der Mehrklassen-, Einklassen- und Meta-Klassifikation erklärt.

Mehrklassen-Klassifikation. Autorenschaftsattribution ist ein typischen Mehrklassen-Klassifikationsproblem. Bei einem Attributionsproblem sind n Autoren $A_1, \ldots, A_n$ jeweils mit einem Textbeispiel gegeben. Das Textbeispiel kann aus einem oder mehreren einzelnen Texten bestehen. Alle diese Texte werden als Stilvektoren repräsentiert. Alle Stilvektoren werden mit dem dazu gehörigen Autor markiert. Die Menge der gekennzeichneten Stilvektoren stellt die Trainingsmenge dar, mit der ein Klassifikator trainiert wird. Texte mit fragwürdiger Autorenschaft werden ebenfalls als Stilvektor im Merkmalsraum der Trainingsmenge repräsentiert. Der Klassifikator ordnet diese einem der Autoren $A_1, \ldots, A_n$ zu.

Einklassen-Klassifikation. Für die Autorenschaftsverifikation, als Einklassen-Klassifikationsproblem, ist eine Menge von Texten gegeben, die alle von einem einzigen Autoren A stammen. Ein anonymer Text d soll der Klasse A zugeordnet werden oder nicht. Als naive Herangehensweise, könnte versucht werden, eine Menge an Texten zusammenzutragen, die nicht von A verfasst wurden. Mit den Texten von A und denen, die nicht von A stammen, könnte dann ein Klassifikator trainiert werden, der d entsprechend in eine der Klassen A oder *nicht-A* klassifiziert. Auch wenn die Mehrzahl aller Texte weltweit nicht von A stammen, ist es praktisch unmöglich eine Repräsentative Menge an Texten anzugeben, die nicht von A verfasst wurden. Auf dieses Problem wurde ausführlich in [12] und [20] eingegangen.

Auch wenn es spezielle Einklassen-Klassifikatoren gibt (zum Beispiel die One-Class-SVM), bleibt das Problem schwer lösbar. In [12] wurden diesbezüglich Experimente durchgeführt, die die Unzulänglichkeit von Einklassen-Klassifikatoren bezüglich der Autorenschaftsverifikation zeigen. Aus diesem Grund ist es sinnvoll, andere Ansätze, wie zum Beispiel die Meta-Klassifikation zu verfolgen.

Meta-Klassifikation. Die Meta-Klassifikation wird für paarweise Autorenschaftsverifikation verwendet. Dabei wird ein Paar von Texten einer der Klassen *selber-Autor* oder *verschiedene-Autoren* zugeordnet. Das Vorgehen ist das folgende: Gegeben ist ein Textpaar (d_1, d_2). Aus (d_1, d_2) wird ein Vektor $\mathbf{s}$ von Eigenschaften berechnet, der stilistische Ähnlichkeiten bzw. Differenzen zwischen d_1 und d_2 enthält. Ein trainierter Klassifikator ordnet $\mathbf{s}$ einer der Klassen *selber-Autor* oder *verschiedene-Autoren* zu. Für das Training des Klassifikators wird eine vorklassifizierte Menge von Textpaaren angegeben, die mit ihrer tatsächlichen Klassenzugehörigkeit *selber-Autor* oder *verschiedene-Autoren* gekennzeichnet sind. Damit können bekannte und erprobte Verfahren der Mehrklassen-Klassifikation verwendet werden.

Zu beachten ist an dem Ansatz der Meta-Klassifikation allerdings, dass ein trainierter Meta-Klassifikator mehr oder weniger abhängig von der Domäne der Textpaare in der Trainingsmenge ist.

Kapitel 5

Unmasking

Unmasking ist ein Verfahren zur paarweisen Autorenschaftsverifikation von langen Texten [12]. Die zugrundeliegende Aufgabe lautet formal wie folgt: Gegeben sind zwei lange Texte d_1 und d_2, wobei sichergestellt ist, dass d_1 vollständig von Autor A_1 und d_2 vollständig von Autor A_2 verfasst wurde. Untersucht wird die Frage, ob $A_1 = A_2$ ist. Dazu wird aus dem Textpaar (d_1, d_2) zunächst ein Vektor $\mathbf{s}$ berechnet, der als Lernkurve bezeichnet wird, und die *Tiefe des Unterschieds*[1] zwischen d_1 und d_2 repräsentiert (Abschnitt 5.1). Anschließend wird $\mathbf{s}$ in eine der beiden Klassen *selber-Autor* und *verschiedene-Autoren* klassifiziert (Abschnitt 5.2).

5.1 Berechnung der Lernkurven

Die Texte d_1 und d_2 werden in Blöcke von 500 Wörtern geteilt. Für jeden der entstandenen Blöcke wird die dazugehörige Repräsentation als Stilvektor berechnet. Als Stilmerkmale dienen die 250 häufigsten Wörter aus der Menge der Wörter, die in beiden Dokumenten vorkommen. Die Stilvektoren aller Blöcke aus d_1 bilden die Repräsentation $\mathbf{D_1}$, analog dazu ist d_2 durch die Menge $\mathbf{D_2}$ der Stilvektoren aller Blöcke aus d_2 repräsentiert.

Zentrale Idee von Unmasking ist es nun, dass ein Klassifikator trainiert wird, der die zwei Mengen $\mathbf{D_1}$ und $\mathbf{D_2}$ trennen soll. Durch 10-fache Kreuzva-

[1] In [12] wird von „depth of difference“ gesprochen.

lidierung wird die Klassifikationsgüte des Klassifikators (hier SVM) bestimmt, die als Maß für die stilistische Unähnlichkeit zwischen d_1 und d_2 angesehen werden kann. Dieses Maß wird nachfolgend als Koppel-SVM-Trennschärfe bezeichnet. Wenn die Mengen $\mathbf{D_1}$ und $\mathbf{D_2}$ sehr gut voneinander getrennt werden können, liegt ein Indiz für unterschiedliche Autorenschaft von d_1 und d_2 vor. Nun ist eine wesentliche Beobachtung, dass zunächst auch Texte ein und desselben Autors auf diese Weise mit einer großen Klassifikationsgüte voneinander getrennt werden können. Jedoch kann weiterhin beobachtet werden, dass in diesen Fällen häufig nur sehr wenige Dimensionen der Stilvektoren für diese Trennung verantwortlich sind. Durch Eliminierung der diskriminativsten Dimensionen sinkt die Koppel-SVM-Trennschärfe im Falle gleicher Autorenschaft deutlich. Im Falle ungleicher Autorenschaft wird beobachtet, dass auch die übrigen Dimensionen ausreichen, um $\mathbf{D_1}$ von $\mathbf{D_2}$ relativ sicher trennen zu können. Diese Beobachtung wird für die Repräsentation des Textpaars (d_1, d_2) als Lernkurve (Vektor $\mathbf{s}$) benutzt, die die Tiefe des Unterschieds zwischen d_1 und d_2 repräsentiert. Die iterative Berechnung von $\mathbf{s}$ geschieht wie folgt:

1. Berechne die Koppel-SVM-Trennschärfe für $\mathbf{D_1}$ und $\mathbf{D_2}$ und speichere diese in $\mathbf{s}$.

2. Eliminiere aus $\mathbf{D_1}$ und $\mathbf{D_2}$ die 6 diskriminativsten Dimensionen.

3. Gehe zu Schritt 1.

Die Eliminierung von Dimensionen wird nach [12] in 9 Iterationsschritten durchgeführt. Nachfolgend wird die iterative Auswahl der Untermengen der Stilmerkmale, die nicht eliminiert werden, als Koppel-Feature-Removal bezeichnet.

In Abbildung 5.1 sind beispielhaft drei Lernkurven dargestellt. Es ist deutlich zu erkennen, dass bei einer der Kurven die Koppel-SVM-Trennschärfe (y-Achse) mit steigender Anzahl an Iterationen (x-Achse) deutlich schneller abfällt als bei den anderen. Diese Kurve wurde aus zwei Texten eines Autors berechnet, während den anderen Kurven Texte unterschiedlicher Autoren zugrunde lagen.

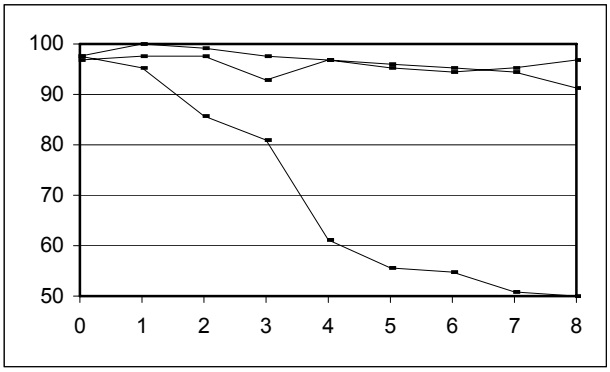


Abbildung 5.1: Dargestellt sind drei Lernkurven. Auf der y-Achse ist die Koppel-SVM-Trennschärfe aufgetragen, auf der x-Achse die Iterationen. Zur Berechnung wurde das Buch *House of Seven Gables* des Autors Nathaniel Hawthorne mit zwei Büchern anderer Autoren (obere Kurven) und einem weiteren Buch von Hawthorne (untere Kurve) verglichen. Quelle: [15].

5.2 Meta-Klassifikation

Die Meta-Klassifikation verläuft wie im vorigen Kapitel beschrieben. Aus einer Trainingsmenge mit Textpaaren von einem Autor und Textpaaren von verschiedenen Autoren werden alle Lernkurven berechnet und mit der entsprechenden Klassenzugehörigkeit gekennzeichnet. Auf dieser Menge von Vektoren wird ein Klassifikator trainiert, der die Lernkurve eines fragwürdigen Textpaares einer der Klassen *selber-Autor* oder *verschiedene-Autoren* zuordnet. In [12] und [15] wird als Klassifikator eine lineare SVM vorgeschlagen.

5.3 Eigenschaften von Unmasking

In [12] wird Unmasking auf einer Menge von Büchern evaluiert. Die Bücher haben eine durchschnittliche Länge von etwa 60 000 Wörtern. Dabei wird eine sehr hohe Klassifikationsgüte von 95.7% gemessen.

Offensichtliche Einschränkung ist jedoch, dass die Textlänge auf ein Minimum von etwa 2 500 Wörter begrenzt ist, und unterhalb von 5 000 Wörtern bereits keine zufriedenstellenden Ergebnisse mehr liefert [10]. Bei 2 500 Wörtern

ergeben sich pro Text durch das Aufteilen in Blöcke genau 5 Textblöcke, somit 10 für beide Texte zusammen, die als Minimum für die 10-fache Kreuzvalidierung notwendig sind.

Weiterhin ist auch bei Unmasking, wie generell bei Ansätzen mit Meta-Lernen im Bereich der paarweisen Autorenschaftsverifikation, eine Abhängigkeit des trainierten Klassifikators von der Domäne zu vermuten.

Kapitel 6

ST-Unmasking

In diesem Kapitel wird ST-Unmasking, ein neues, generisches Verfahren zur paarweisen Autorenschaftsverifikation vorgestellt. ST-Unmasking basiert auf den Ideen des Unmaskings von Koppel und Schler [12]: Gegeben ist ein Textpaar (d_1, d_2), aus diesem Textpaar wird unter Verwendung von Stilmerkmalen ein Vektor $\mathbf{s}$ berechnet, der die Ähnlichkeit, respektive Differenz[1], zwischen d_1 und d_2 repräsentiert. Anschließend wird $\mathbf{s}$ durch Meta-Klassifikation einer der Klassen *selber-Autor* oder *verschiedene-Autoren* zugeordnet.

ST-Unmasking kann als Generalisierung von Unmasking verstanden werden. Dabei wird besonderer Wert auf die Modularisierung des Verfahrens gelegt, um die Möglichkeit zu eröffnen, für bestimmte Anforderungen die geeignete Konfiguration zusammenstellen zu können. So kann in einer Konfiguration von ST-Unmasking das originale Unmasking von Koppel und Schler durchgeführt werden. Jedoch liegt der Schwerpunkt von ST-Unmasking darauf, kurze Texte verarbeiten zu können und damit den Einschränkungen von Unmasking zu begegnen.

[1]Der Begriff Ähnlichkeit wird in dieser Arbeit synonym für sowohl Ähnlichkeit als auch Differenz und Abstand verwendet, da eine Umrechnung von einer Differenz (Unähnlichkeit, Unterschied) bzw. eines Abstands in eine Ähnlichkeit jederzeit möglich ist.

6.1 Besonderheiten kurzer Texte

In diesem Abschnitt werden die für ST-Unmasking wichtigen Eigenschaften von kurzen Texten erörtert. Kurze Texte haben die Besonderheit, den Schreibstil des Autors nicht signifikant zu repräsentieren [6]. Dadurch wird es vorkommen, dass sich der Schreibstil in zwei kurzen Texten kaum messbar unterscheidet, obwohl beide Texte von zwei unterschiedlichen Autoren stammen, die, bezogen auf längere Texte, durchaus einen messbar unterschiedlichen Schreibstil haben. Auch der umgekehrte Fall ist möglich: Zwei kurze Texte können sich unter den verwendeten Stilmerkmalen drastisch unterscheiden, obwohl sie von einem einzigen Autor verfasst wurden. Um dem zu begegnen, wird in ST-Unmasking der Strategie nachgegangen, möglichst viele Informationen aus den kurzen Texte zu extrahieren. So werden in ST-Unmasking viele der in Kapitel 3 vorgestellten Stilmerkmale kombiniert verwendet. Weiterhin werden Strategien vorgeschlagen, den durch alle Stilmerkmalsausprägungen aufgespannten Merkmalsraum in viele unterschiedliche Teilräume zu projizieren. Die Ähnlichkeit zwischen d_1 und d_2 wird in jedem dieser Teilräume gemessen. Jeder so entstandene Ähnlichkeitswert wird als Aspekt der stilistischen Ähnlichkeit zwischen d_1 und d_2 verstanden. Eine der vorgeschlagenen Strategien wird das Koppel-Feature-Removal sein.

Weiterhin wird vorgeschlagen, externes Wissen in den Entscheidungsprozess der Autorenschaftsverifikation einfließen zu lassen. Externes Wissen im Falle der Autorenschaftsverifikation liegt sprach- bzw. domänenspezifisch in den Erwartungswerten jeder einzelnen Stilmerkmalsausprägung vor. Zum Beispiel repräsentieren die Erwartungswerte der Verteilung von Funktionswörtern externes Wissen. Die Idee ist es, die Texte d_1 und d_2 als Stilvektoren zu repräsentieren, die die Abweichungen vom erwarteten, durchschnittlichen Stil der Domäne enthalten. Ein ähnliches Vorgehen wird in [8] für die Autorenschaftsattribution berichtet, wo der Quotient aus tatsächlicher und erwarteter Häufigkeit einer Stilmerkmalsausprägung gebildet wird. Eine Verbindung zu kurzen Texten bestand in dieser Untersuchung jedoch nicht.

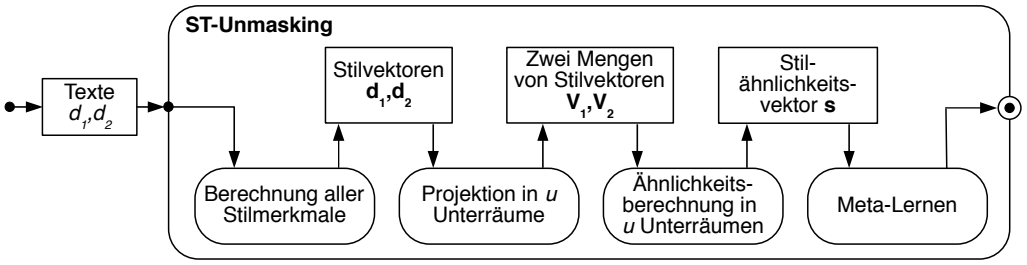


Abbildung 6.1: UML Aktivitätsdiagramm von ST-Unmasking.

6.2 Ablauf von ST-Unmasking

In diesem Abschnitt wird ein Überblick über den Ablauf von ST-Unmasking gegeben. Jeder einzelne Schritt wird in den nachfolgenden Abschnitten detailliert erklärt. ST-Unmasking kann in die folgenden 4 Schritte unterteilt werden:

1. Berechnung der Menge aller Stilmerkmale aus d_1 und d_2 und damit jeweils der Repräsentation von d_1 und d_2 als Stilvektor $\mathbf{d_1}$ und $\mathbf{d_2}$ aller Stilmerkmale. Alle Stilmerkmale zusammen bilden den Merkmalsraum.

2. Selektion mehrerer verschiedener Untermengen der Stilmerkmale (disjunkt oder nicht disjunkt). Das entspricht der Aufteilung des Merkmalsraums in viele Unterräume. Die Teilvektoren von $\mathbf{d_1}$ und $\mathbf{d_2}$ in allen Unterräumen bilden die Mengen von Vektoren $\mathbf{V_1}$ und $\mathbf{V_2}$.

3. Ähnlichkeitsberechnung in allen Unterräumen und Speicherung der Ähnlichkeitswerte im Stilähnlichkeitsvektor $\mathbf{s}$. In diesem Schritt findet optional die Anreicherung mit externem Wissen statt. Es können mehrere Ähnlichkeitsmaße verwendet werden.

4. Meta-Klassifikation von $\mathbf{s}$.

In Abbildung 6.1 ist der gesamte Prozess des ST-Unmaskings als UML Aktivitätsdiagramm dargestellt.

6.3 Berechnung der Stilmerkmale

Alle in Kapitel 3 vorgestellten Stilmerkmale werden jeweils für d_1 und d_2 berechnet. Je nach konkreter Aufgabe können weitere Stilmerkmale hinzugefügt werden, insbesondere Merkmale, die aus speziellen Meta-Informationen berechnet werden, wie sie zum Beispiel in [10] für E-Mails vorgeschlagen werden. Weiterhin wird sich in der Evaluierung in Kapitel 7 zeigen, dass auch das Hinzufügen eventuell schwacher Stilmerkmalen, also solcher mit geringer Trennschärfe, keine nennenswerte Verschlechterung der Klassifikationsgüte von ST-Unmasking nach sich zieht. Die Menge aller Stilmerkmale für d_1 und d_2 bilden die Repräsentationen $\mathbf{d_1}$ und $\mathbf{d_2}$ als Stilvektor.

6.4 Projektionsstrategien

Mit Hilfe der Projektionsstrategien werden, nach verschiedenen Kriterien, aus dem Merkmalsraum viele Unterräume berechnet. Zunächst wird diskutiert, welche Ziele damit verfolgt werden. Werden die Stilmerkmale aus Kapitel 3 betrachten, so kann festgestellt werden, dass bestimmte Stilmerkmale bestimmte Aspekte des Schreibstils widerspiegeln. Zum Beispiel kann davon ausgegangen werden, dass 1-Suffix-Bigramme in der deutschen Sprache die Präferenz eines Autors bezüglich des Genitivs transportieren können, da die Frequenz des Bigrams *s_s* bei intensiver Verwendung des Genitivs erhöht sein sollte. Funktionswörter als Stilmerkmale können auf eine Präferenz bezüglich des bestimmten oder unbestimmten Artikels hinweisen. Aber auch jede gemischte Menge von Stilmerkmalen wird als Träger eines abstrakten Aspekts des Schreibstils verstanden. Wird also der Merkmalsraum in viele Unterräume aufgeteilt, so repräsentiert jeder Unterraum einen gewissen abstrakten Aspekt des Stils. Auch das Koppel-Feature-Removal ist eine Aufteilung des Merkmalsraums in (nicht disjunkte) Unterräume, die in diesem Sinne jeweils einen bestimmten Aspekt des Schreibstils betrachten. Anders formuliert kann die Aufteilung des Merkmalsraums eine feingliedrigere Ähnlichkeitsstruktur zwischen d_1 und d_2 aufdecken. Weiterhin erhöht sich dadurch die Dimensionalität des Stilähnlichkeitsvektors $\mathbf{s}$, wodurch dem Meta-Klassifikator später eine feinere Auflösung

ermöglicht wird. Verschiedene Strategien zur Konstruktion der Unterräume sind möglich. Nachfolgend werden zunächst zwei naive Varianten vorgestellt, die den Merkmalsraum in disjunkte Unterräume teilen. Daran anschließend folgen Verfahren, die gezielt vorgehen, um auch aus einer eventuell kleinen Dimensionalität des Merkmalsraums viele nicht disjunkte Teilräume zu konstruieren. Insbesondere kann die nicht disjunkte Trennung an die diskunkte Trennung angeschlossen werden, so dass jeder Unterraum, der mit einem der naiven Verfahren erstellt wurde, mit einem der nicht naiven Verfahren weiter aufgeteilt wird.

6.4.1 Naive Projektionsstrategien

Die folgenden zwei Projektionsstrategien teilen dem Merkmalsraum in disjunkte Unterräume. Bei der *intuitiven Trennung* bildet jedes Stilmerkmal direkt einen Unterraum. Werden zum Beispiel Funktionswörter, Zeichen-Trigramme und Phonem-Bigramme als Stilmerkmale verwendet, so ergeben sich drei Unterräume: Alle Funktionswörter bilden den ersten Unterraum, alle Zeichen-Trigramme den Zweiten und alle Phonem-Bigramme den Dritten.

Eine weitere Projektionsstrategie ist die *zufällige Trennung*. Dabei wird der Merkmalsraum in u zufällige, disjunkte Unterräume geteilt. Die Dimensionalität der Unterräume ist durchschnittlich $\frac{1}{u}$ mal der Dimensionalität des gesamten Merkmalsraums. Diese Aufteilung geschieht durch Zufallsprojektion. Dabei ist die i-te Zufallsprojektion mit $i = 1, \ldots, u$ immer gleich, um die Vergleichbarkeit der Dimensionen in $\mathbf{s}$ zu gewährleisten.

6.4.2 Angepasstes Koppel-Feature-Removal

In [12] wird davon ausgegangen, dass nur wenige Stilmerkmale für eine eventuelle stilistische Unähnlichkeit zwischen zwei Texten eines Autors verantwortlich sind. Bei Texten verschiedener Autoren sind viele Stilmerkmalsausprägungen unähnlich. Im Unmasking wird mit dem Koppel-Feature-Removal das Ziel verfolgt, durch gezielte Elimination der trennschärfsten Stilmerkmale die Koppel-SVM-Trennschärfe zwischen zwei Texten zu reduzieren. Dabei wird festgestellt, dass die Größe der Koppel-SVM-Trennschärfe für Texte eines einzigen Autors

drastischer abnimmt als bei Texten verschiedener Autoren. Dass dieser Ansatz funktioniert, wird in mehreren Arbeiten experimentell gezeigt [12, 15, 19, 20]. Im Prinzip kann dieses Verfahren direkt auch im ST-Unmasking verwendet werden. Anders ist lediglich, dass nicht zwei Mengen von Vektoren $\mathbf{D_1}$ und $\mathbf{D_2}$ die Ausgangsbasis darstellen, sondern nur die zwei Vektoren $\mathbf{d_1}$ und $\mathbf{d_2}$. Die diskriminativsten Dimensionen von $\mathbf{d_1}$ und $\mathbf{d_2}$ lassen sich identisch berechnen. Die iterative Eliminierung der trennschärfsten Dimensionen erfolgt analog.

6.4.3 Geordnete Zufallsauswahl

Eine neu entwickelte Alternative zum Koppel-Feature-Removal ist die Geordnete Zufallsauswahl. Sei Q die Menge der Dimensionen die in $\mathbf{d_1}$ und $\mathbf{d_2}$ von 0 verschieden sind, also zum Beispiel die in d_1 und in d_2 vorhandenen Zeichen-Trigramme. Es werden u zufällige Stichproben $S_i, i = 1, \ldots, u$ aus Q gezogen. S_i sind nicht disjunkte, echte Teilmengen von Q mit $S_i \neq S_j$ für $i \neq j$. Die in den Stichproben S_i enthaltenen Dimensionen bilden Unterräume des Merkmalsraums der Stilmerkmale. Es wird die Ähnlichkeit von $\mathbf{d_1}$ und $\mathbf{d_2}$ in allen sich ergebenden Unterräumen gemessen, damit ergeben sich u Ähnlichkeitswerte. Die Hypothese ist, dass sich die Verteilung dieser Ähnlichkeitswerte für ein Textpaar (d_1, d_2) eines einzigen Autors anders verhält, als für ein Textpaar zweier unterschiedlicher Autoren. Als Ähnlichkeitsmaß kann auch hier die Koppel-SVM-Trennschärfe, aber auch jedes andere Ähnlichkeits- oder Distanzmaß verwendet werden. Der Argumentation von Koppel und Schler in [12] folgend, dass, im Falle von Texten eines einzigen Autors, die Trennung der Texte nur durch wenige Dimensionen im Merkmalsraum gewährleistet wird, kann auch hier davon ausgegangen werden, dass nur wenige Stichproben S_i diese Dimensionen enthalten. Somit werden im Falle gleicher Autorenschaft viele der berechneten Ähnlichkeitswerte groß sein (bzw. der Koppel-SVM-Trennschärfe klein).

Umgekehrt kann argumentiert werden, dass die Ähnlichkeit zwischen den Stilvektoren zweier Texte unterschiedlicher Autoren unter einer bestimmten Auswahl von Dimensionen zufällig hoch sein kann. Ist die Ähnlichkeit aber in vielen verschiedenen Unterräumen hoch, dann liegt ein starkes Indiz für gleiche

Autorenschaft vor. Diese Eigenschaft wird ausführlich in [14] im Zusammenhang mit Autorenschaftsattribution diskutiert.

Da die Auswahl der Dimensionen abhängig von den tatsächlichen vorhandenen Elementen in Q ist, ist die Reihenfolge der Stichproben arbiträr. Damit ist auch die Reihenfolge der in den Unterräumen berechneten Ähnlichkeitswerte arbiträr, lediglich die Verteilung ist interessant. Um später eine Vergleichbarkeit der Dimensionen im Stilähnlichkeitsvektor $\mathbf{s}$ zu gewährleisten, muss die Verteilung beschrieben werden. Dazu werden die Ähnlichkeitswerte ihrer Größe nach geordnet. Im weiteren Sinne entspricht die geordnete Liste der Ähnlichkeitswerte den $\frac{1}{u}$-Quantilen der Verteilung, und ist demnach eine geeignete Beschreibung der Verteilung. Für $u = 10$ würde die sortierte Liste der Ähnlichkeitswerte den Dezilen (0.1-Quantile) der Verteilung der Ähnlichkeitswerte entsprechen.

6.5 Ähnlichkeitsberechnung

In diesem Schritt wird für jeden Unterraum die Ähnlichkeit der Teilvektoren von $\mathbf{d_1}$ und $\mathbf{d_2}$ im jeweiligen Raum berechnet. Dafür können mehrere Ähnlichkeitsmaße verwendet werden. Die Ähnlichkeitswerte werden nacheinander in den Stilähnlichkeitsvektor $\mathbf{s}$ geschrieben.

Sei u die Anzahl der Unterräume und $\pi_i : \mathbf{R}^n \rightarrow \mathbf{R}^m$ mit $n > m$ und $i = 1, \ldots, u$ die Projektion aus dem Merkmalsraum in den i-ten Unterraum. Sei weiterhin v die Anzahl der verwendeten Ähnlichkeitsmaße φ_j mit $j = 1, \ldots, v$. Dann wird die w-te Dimension $\mathrm{s_w}$ mit $w = (j-1)u+i$ in $\mathbf{s}$ wie folgt berechnet:

$$\mathrm{s_w} = \varphi_j(\pi_i(\mathbf{d_1}), \pi_i(\mathbf{d_2}))$$

Ähnlichkeitswerte, die in den gleichen Räumen aber mit verschiedenen Ähnlichkeitsmaßen berechnet werden, können wiederum als Aspekte der stilistischen Ähnlichkeit verstanden werden. Die folgenden Maße werden als Ähnlichkeits- bzw. Distanzmaße für das ST-Unmasking vorgeschlagen:

- *Kosinus-Ähnlichkeit.* Dabei wird der Kosinus des Winkels zwischen den Stilvektoren berechnet. Das entspricht dem Skalarprodukt der normali-

sierten Stilvektoren.

- *Kendalls Tau.* Mit Kendalls Tau wird die Rangkorrelation zwischen zwei Listen gemessen. Die Anwendung als Ähnlichkeitsmaß für zwei Stilvektoren geschieht wie folgt: Für jeden der beiden Stilvektoren wird eine Liste erstellt, die alle Dimensionen des Stilvektors in der Reihenfolge des Wertes ihrer Ausprägung enthält. Kendalls Tau wird mit diesen Listen berechnet.

- *Stamatatos-ND1.* Dieses Maß misst die Distanz zwischen den Stilvektoren nach Stamatatos nd_1 [18]: Aus $\mathbf{d_1}$ und $\mathbf{d_2}$ werden alle Dimensionen entfernt, die in einem der beiden Vektoren 0 sind. Die so entstandenen Vektoren werden mit $\mathbf{d_1^*}$ und $\mathbf{d_2^*}$ bezeichnet und haben die Dimensionalität z. $\mathrm{d}_{1\mathrm{i}}^*$ sei die i-te Dimension in $\mathbf{d_1^*}$, $\mathrm{d}_{2\mathrm{i}}^*$ die i-te Dimension in $\mathbf{d_2^*}$. Dann berechnet sich nd_1 wie folgt:

$$nd_1(\mathbf{d_1^*}, \mathbf{d_2^*}) = \frac{1}{4z} \sum_{i=1}^{z} \left(\frac{2\,(\mathrm{d}_{1\mathrm{i}}^* - \mathrm{d}_{2\mathrm{i}}^*)}{\mathrm{d}_{1\mathrm{i}}^* + \mathrm{d}_{2\mathrm{i}}^*} \right)^2$$

- *Koppel-SVM-Trennschärfe.* Die Koppel-SVM-Trennschärfe kann für eine Textlänge ab 2 500 Wörter berechnet werden.

Vor der Ähnlichkeitsberechnung findet gegebenenfalls die Wissensanreicherung durch die Verrechnung der Erwartungswerte der einzelnen Stilmerkmalsausprägungen statt: Sei $\mathbf{e}$ der Vektor der Erwartungswerte aller Stilmerkmale im Merkmalsraum, dann berechnet sich die w-te Dimension in $\mathbf{s}$ durch:

$$\mathrm{s_w} = \varphi_j(\pi_i(\mathbf{d_1} - \mathbf{e}), \pi_i(\mathbf{d_2} - \mathbf{e}))$$

6.6 Meta-Klassifikation.

Die Meta-Klassifikation erfolgt analog zum Unmasking. Als Klassifikator wird Random Forest vorgeschlagen, der sich in den Experimenten (Kapitel 7) als besonders geeignet erweisen wird.

6.7 Unmasking mit ST-Unmasking

Werden als Stilmerkmale die 250 häufigsten Wörter aus $d_1 \cap d_2$, als Strategie für die Merkmalsaufteilung das Koppel-Feature-Removal und als Ähnlichkeitsmaß die Koppel-SVM-Trennschärfe verwendet, so entspricht ST-Unmasking dem Unmasking, wie es von Koppel und Schler vorgestellt wurde.

Kapitel 7

Evaluierung

In diesem Kapitel werden die vorgestellten Stilmerkmale sowie Unmasking und ST-Unmasking experimentell untersucht. Dazu werden zunächst Korpora vorgestellt, die zur Evaluierung erstellt wurden. Anschließend wird der grundsätzliche Aufbau und Ablauf der Experimente erklärt, sowie eine Übersicht über die variierten Parameter gegeben. In Abschnitt 7.3 werden die in Kapitel 3 vorgestellten Stilmerkmale evaluiert, Abschnitt 7.4 beschreibt experimentell den Weg von Unmasking zu ST-Unmasking und Abschnitt 7.5 untersucht den Einfluss der in Kapitel 6 vorgeschlagenen Parameter von ST-Unmasking.

7.1 Korpora

Für die Evaluierung der vorgestellten Stilmerkmale und der Verfahren für die Autorenschaftsverifikation wurden zwei Textkorpora erstellt. Bevor diese vorgestellt werden, wird der Frage nachgegangen, welche Anforderungen an einen Korpus für die paarweise Autorenschaftsverifikation gestellt werden. Wichtige Anforderungen sind die folgenden:

- Die Autorenschaft der Texte muss bekannt sein.
- Es müssen zu einem großen Teil mehrere Texte pro Autor vorhanden sein, so dass viele Textpaare mit gleicher Autorenschaft gebildet werden können. Textpaare mit verschiedener Autorenschaft können kombinatorisch meist wesentlich mehr gebildet werden.

- Jeder Text im Korpus darf nur von einem einzigen Autor verfasst worden sein.

Weiterhin sind im Folgenden Eigenschaften aufgelistet, die für einen Autorenschaftskorpus problematisch sein könnten:

- Übersetzungen, zum Beispiel von deutschen Büchern ins Englische sollten vermieden werden, insbesondere dann, wenn Texte eines fremdsprachlichen Autors von verschiedenen Übersetzern in die Zielsprache überführt wurden.

- Texte, die inhaltlich oder bezüglich des Genres über die Grenzen eines Autors hinweg sehr inhomogen sind, könnten problematisch sein, da die Klassifikation nicht zwangsläufig auf den Schreibstil zurückzuführen ist.

Ein Korpus, der alle genannten Anforderungen erfüllt, existiert bislang nicht. Einen solchen Korpus zu erstellen, der ausreichend viele Texte enthält, ist eine sehr aufwendige Aufgabe, da die Texte größtenteils von Hand selektiert werden müssten. Andererseits kann ein nach den oben genannte Kriterien, künstlich konstruierter Korpus wiederum eine Verzerrung realer Gegebenheiten darstellen.

Wichtigste Grundlage für die Konstruktion eines Korpus ist die Verfügbarkeit von geeigneten Daten. Eine große Datenbasis bietet das Projekt Gutenberg[1], wo eine Vielzahl von Büchern, deren Copyright abgelaufen ist, frei als Text heruntergeladen werden kann. Auf Grundlage der gesamten Zusammenstellung von Texten aus dem Jahr 2008, die als DVD zum Download angeboten wird, wurde ein Korpus erstellt. Dieser Korpus wird im folgenden als *G2008* bezeichnet. Der gesamte Korpus besteht aus knapp 20 000 englischen Texten von 7 400 verschiedenen Autoren. Daraus wurde durch zufälliges Ziehen von Autoren eine Trainingsmenge und eine Testmenge erstellt. Diese Mengen sind die bezogen auf Texte und Autoren disjunkt. Die Trainingsmenge besteht aus 200 Textpaaren, von denen 100 paarweise den selben Autor haben und 100 paarweise zwei unterschiedliche Autoren. Damit sind die Klassen *selber-Autor*

[1] `http://www.gutenberg.org`

und *verschiedene-Autoren* gleichstark vertreten. Die Testmenge ist prinzipiell identisch aufgebaut, mit dem Unterschied, dass je Klasse 300 Textpaare enthält.

Einschränkungen erfährt dieser Korpus aufgrund der folgenden Eigenschaften: Die Texte sind vollständige Bücher, und als solche mit durchschnittlich 60 000 Wörtern relativ lang. Um diesen Korpus dennoch für die Evaluierung von kürzeren Texten zu verwenden, werden künstlich Teile aus den Büchern extrahiert. Allerdings ist nicht sicher gestellt, dass die so erstellte Menge an kurzen Texten reale Bedingungen simuliert. Jedoch ist es von Vorteil, dass der G2008-Korpus lange Texte enthält, da somit auf einer einheitlichen Datenbasis Experimente mit langen und kurzen Texten durchgeführt werden können, wodurch eine gute Vergleichbarkeit der Ergebnisse auf langen und kurzen Texten gewährleistet ist.

Ein weiterer Korpus wurde auf Grundlage von Kommentartexten der Internetplattform IMDB[2] erstellt. IMDB ist eine Online Datenbank, die Informationen zu vielen Filmen und Serien enthält. Zu jedem Film können Benutzer Kommentare bzw. Kritiken schreiben. Auf der Basis aller Kommentare, die im Oktober 2010 veröffentlicht waren, wurden alle diejenigen ausgewählt, die mindestens 250 Wörter enthielten. Daraus wurde der *IMDB*-Korpus konstruiert. Der Korpus enthält knapp 360 000 Texte von über 80 000 Autoren mit durchschnittlich 440 Wörtern pro Text. Auch daraus wurde für die Experimente eine Trainingsmenge und eine Testmenge konstruiert. Beide Mengen sind ebenso aufgebaut, wie die des G2008-Korpus. Jedoch ist die mögliche Länge der Texte für die Experimente auf maximal 750 Wörter begrenzt.

7.2 Experimentieraufbau, Grundgerüst

In diesem Abschnitt wird der grundsätzliche Aufbau der Experimente erklärt. Jedes Experiment wird mit der Trainings- bzw. Testmenge der oben beschriebenen Korpora durchgeführt. Es werden in jedem Experiment eine oder mehrere festgelegte Textlängen untersucht. Dazu wird jeder Text entsprechend der

[2] `http://www.imdb.com`

geforderten Textlänge in Blöcke geteilt und anschließend zufällig einer der entstandenen Blöcke als Textbeispiel verwendet. Sollen zum Beispiel Experimente mit der Textlänge von 1 000 Wörtern durchgeführt werden, und ein Text hat die Länge von 83 000 Wörtern, so wird dieser in 83 Blöcke von je 1 000 Wörtern geteilt. Anschließend wird einer dieser Blöcke als Repräsentant dieses Textes zufällig aus allen 83 Blöcken gezogen. Somit ergeben die Blöcke nicht zwingend Sinneinheiten, da gegebenenfalls sowohl Kapitel als auch Absätze und Sätze aufgebrochen werden. Weiterhin werden alle Buchstaben in Kleinbuchstaben gewandelt und sämtliche Formatierungen und Satzzeichen entfernt. Dieser Schritt dient der Vermeidung des Auswendiglernens eventueller Korpusspezifika, die sich zum Beispiel durch eine spezielle Formatierung von Texten bestimmter Autoren im G2008-Korpus ergeben könnten. Für jedes Experiment werden die folgenden Parameter eingestellt:

1. Liste der zu verwendende Stilmerkmale.
2. Liste der Projektionsstrategien.
3. Liste der Ähnlichkeitsmaße.
4. Auflistung der zu untersuchenden Textlängen.
5. Zu verwendender Korpus.

In einigen Experimenten wird weiterhin der Meta-Klassifikator variiert, oder die Verwendung von externem Wissen in Form von Erwartungswerten der Stilmerkmalsausprägungen verwendet. In Tabelle A.1 werden alle Parameter und ihre Abkürzungen vollständig aufgelistet. Weiterhin sind in Tabelle A.2 die Parameter aller Experimente aufgeführt.

Nach der Festlegung aller Parameter durchläuft jedes Experiment eine Trainings- und eine Testphase. In der Trainingsphase wird für jedes Textpaar der Trainingsmenge der Stilähnlichkeitsvektor berechnet. Mit allen Stilähnlichkeitsvektoren wird anschließend der Klassifikator trainiert. Analog dazu werden auch für die Testmenge alle Stilähnlichkeitsvektoren berechnet und damit der trainierte Klassifikator getestet. Als Maß für die Bewertung der Klassifikation wird die Klassifikationsgüte verwendet. Diese ist insbesondere in Hinblick auf

die gleiche Größe der beiden Klassen (*selber-Autor* und *verschiedene-Autoren*) in der Trainings- sowie der Testmenge, und der daraus resultierenden 50% a-priori Wahrscheinlichkeit für beide, sinnvoll.

7.3 Evaluierung der Stilmerkmale

Die Evaluierung der Stilmerkmale ist in drei Schritte aufgeteilt. Untersucht werden die folgenden Fragen:

- Wie stark ist die Trennschärfe einzelner Stilmerkmale bezüglich der Meta-Klassifikation?
- Wie groß ist die paarweise Korrelation zwischen allen Stilmerkmalen?
- Wie unabhängig sind die Stilmerkmale von inhaltlicher Ähnlichkeit der untersuchten Texte?

Viele der in Kapitel 3 vorgestellten Stilmerkmale haben Parameter, wie zum Beispiel N in N-Grammen. Für die möglichen Ausprägungen wurden Experimente durchgeführt, die aufgrund der großen Anzahl an dieser Stelle nicht aufgelistet werden. Die Stilmerkmale werden im Folgenden in der jeweils geeignetsten Einstellung ihrer Parameter verwendet, welche aus der entsprechenden Bezeichnung ersichtlich ist (zum Beispiel werden Zeichen-N-Gramme mit $N = 3$ verwendet und somit als Zeichen-Trigramme bezeichnet).

Experiment 1 (*Ex1*): Trennschärfe.
Dieses Experiment ist in viele Einzelexperimente aufgeteilt: Für jedes Untersuchte Stilmerkmal einzeln wurde ein Experiment wie oben beschrieben durchgeführt. Das Klassifikationsergebnis für jedes Stilmerkmal dient als Maß für die Trennschärfe dieses Stilmerkmals. Zusätzlich wurde auf der Trainingsmenge jeweils 10-fache Kreuzvalidierung durchgeführt, und die daraus berechnete Klassifikationsgüte als Vergleichswert für die Trennschärfe betrachtet. Als Datenbasis wurde der G2008-Korpus verwendet. Die untersuchte Textlänge lag bei 20 000 Wörtern. Die Ergebnisse des Experiments sind in Tabelle 7.1 aufgeführt.

Stilmerkmal	**Trennschärfe**	
	Kreuzvalidierung	*Testmenge*
Funktionswörter	0.86	**0.84**
Zeichen-Trigramme	0.85	0.81
Zeichen-1-Skip-Trigramme	0.84	0.78
POS-Triramme	**0.87**	0.83
Phonem-Bigramme	0.83	0.82
Vokal-Trigramme	0.78	0.73
Vokal-Konsonant-6-Gramme	0.72	0.68
Wortlängen-Bigramme	0.80	0.69
1-Präfix-Bigramme	0.79	0.77
2-Suffx-Bigramme	0.85	0.81

Tabelle 7.1: Trennschärfe von Stilmerkmalen. Einerseits wurde die Trennschärfe anhand 10-facher Kreuzvalidierung auf der Trainingsmenge ermittelt, andererseits durch Validierung eines auf der Trainingsmenge trainierten Klassifikators auf der Testmenge.

Als Baseline für die Trennschärfe wird der Zufall angenommen. Damit liegt die Baseline bei 0.5, da zwei Klassen zugeordnet werden, die die gleiche Größe und somit die selbe a-priori Wahrscheinlichkeit haben. Um diesen Wert empirisch zu überprüfen, wurde ein Pseudo-Stilmerkmal entwickelt. Für dieses Stilmerkmal wird ausgehend vom Hashwert des zugrunde liegenden Textes ein 100-dimensionaler Vektor von (Pseudo-)Zufallswerten berechnet. Der Vektor repräsentiert die Verteilung von 100 Stilmerkmalsausprägungen. Durch die Verwendung des Hashwertes als Ausgang für die Berechnung der Zufallswerte ist sicher gestellt, dass für ein und den selben Text immer der selbe Zufallsvektor berechnet wird. Die Zufallswerte sein gleichverteilt. Die Klassifikationsgüte für dieses Pseudo-Stilmerkmal beträgt 0.51 für die Kreuzvalidierung und 0.53 für die Testmenge.

Die Trennschärfe aller Stilmerkmale liegt deutlich über der Baseline, wobei Funktionswörter und POS-Trigramme am stärksten trennen. Zwei der neu entwickelten Stilmerkmale, Phonem-Bigramme und 2-Suffix-Bigramme, liegen ebenfalls im oberen Bereich, wohingegen insbesondere die Vokal-Konsonant-6-Gramme in beiden Testvarianten am schlechtesten abschneiden.

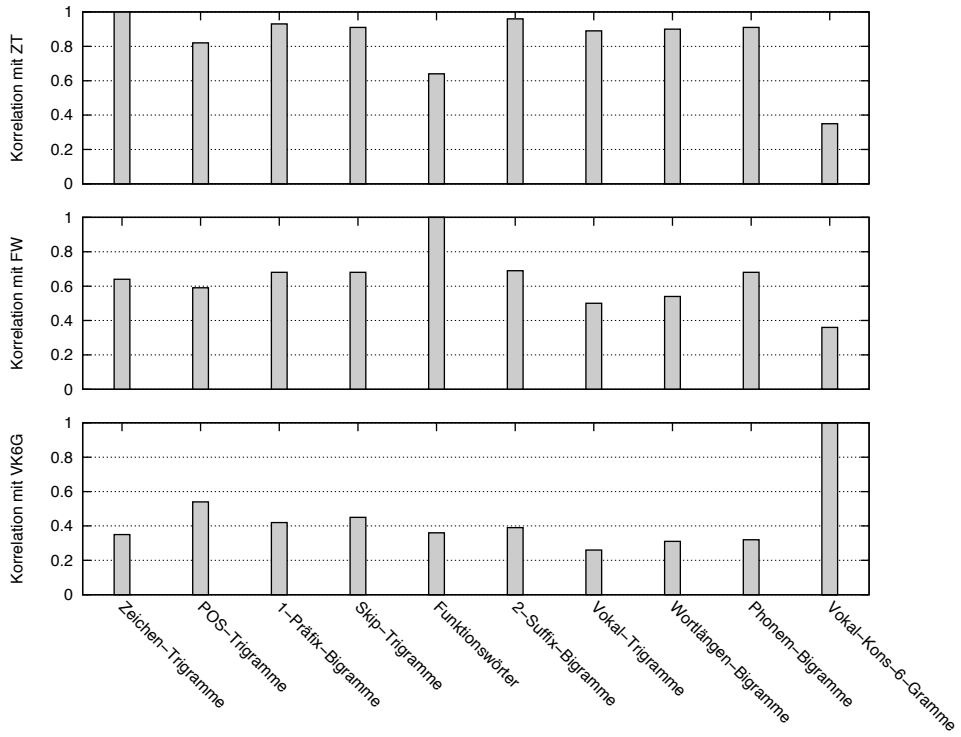


Abbildung 7.1: Es sind jeweils die Korrelationskoeffizienten (y-Achsen) eines Stilmerkmals zu allen anderen Stilmerkmalen (x-Achse) dargestellt.

Weiterhin wurde getestet wie sich die Kombination eines trennscharfen Stilmerkmals, den Funktionswörtern, mit dem Pseudo-Stilmerkmal, als Repräsentant eines schlechten Stilmerkmals, verhält. Bei der Kreuzvalidierung lag die Kombination mit einer Klassifikationsgüte von 0.85 etwas unter dem Ergebnis der Funktionswörter als alleiniges Stilmerkmal. Dagegen war bei der Validierung auf der Testmenge mit 0.86 ein leichter Anstieg der Klassifikationsgüte messbar.

Experiment 2 (*Ex2*): Korrelation von Stilmerkmalen.
Sollen mehrere Stilmerkmale miteinander kombiniert, bzw. gemischt werden, so ist nur dann ein Mehrwert an Information zu erwarten, wenn diese Stilmerkmale möglichst wenig korrelieren. Im folgenden Experiment wird paarweise die Korrelation der vorgestellten Stilmerkmale gemessen. Dazu wird wiederum die G2008 Trainingsmenge verwendet. Aus den 10 Stilmerkmalen ergeben sich 45

paarweise Kombinationen und damit 45 Einzelexperimente. Für die Einzelexperimente wird die Projektionsstrategie *intuitive Trennung* und die Kosinus-Ähnlichkeit verwendet. Dadurch hat der Stilähnlichkeitsvektor nur zwei Dimensionen, eine pro Kosinus-Ähnlichkeit unter den zwei untersuchten Stilmerkmalen. Die zwei Dimensionen können als Wertepaare betrachtet werden, für die über die gesamte Trainingsmenge der Korrelationskoeffizient berechnet wird. In Abbildung 7.1 sind beispielhaft die Korrelationskoeffizienten von drei Stilmerkmalen zu allen anderen Stilmerkmalen dargestellt. Es ist deutlich zu erkennen, dass die Zeichen-Trigramme am stärksten mit allen anderen Stilmerkmalen korrelieren. Die Korrelationskoeffizienten der hier nicht dargestellten Stilmerkmale verhalten sich ähnlich denen der Zeichen-Trigramme.

Es konnte in weiteren Experimenten jedoch nicht nachgewiesen werden, dass bei der Kombination von Stilmerkmalen eine hohe oder niedrige Korrelation eine Aussage über die zu erwartende Steigerung der Klassifikationsgüte ermöglicht. So können unter Umständen sogar hochgradig korrelierende Stilmerkmale in Kombination die Klassifikationsgüte steigern, während schwach korrelierende Kombinationen bisweilen kaum eine messbare Verbesserung bewirken.

Experiment 3 (*Ex3*): Unabhängigkeit vom Inhalt.
Eine vollständige Unabhängigkeit der Stilmerkmale vom Inhalt eines Textes kann kaum möglich sein. Der Grund dafür ist, dass alle vorgestellten Stilmerkmale Verteilungen messen, die anhand verschiedener Berechnungsvorschriften aus den Zeichen eines Textes extrahiert werden. Der Inhalt des Textes hat natürlich einen großen Einfluss auf die Verteilung der Zeichen. Somit hat der Inhalt einen Einfluss auf die Stilmerkmale, jedoch sollte dieser möglichst gering sein.

Das folgende Experiment soll die Abhängigkeiten von Stil und Inhalt messen. Dazu muss zunächst ein Maß gefunden werden, dass die inhaltliche Ähnlichkeit zweier Texte messen kann. Auf dem Gebiet des textbasiertes Informationretrievals spielt die Repräsentation des Inhalts eines Textes d eine zentrale Rolle. Eine verbreitete Variante aus diesem Forschungsbereich ist die Repräsentation von d als Vektor von Termfrequenzen (tf-Gewichte). Dazu werden aus d alle Funktionswörter entfernt und alle übrig gebliebenen Wörter werden auf

ihre Stammform reduziert. Diese Wörter werden als Terme bezeichnet, die relative Häufigkeiten der Terme in d bilden die tf-Gewichte. Alle Terme bilden einen Vektorraum, jeder Text ist in diesem Raum als Vektor der tf-Gewichte seiner Terme repräsentiert. Diese Vektoren werden als Dokumentvektoren bezeichnet. Die Ähnlichkeit zweier Texte wird üblicherweise durch den Kosinus zwischen den Dokumentvektoren der Texte berechnet.

Analog zum vorangegangenen Experiments wurden die Korrelationskoeffizienten der Ähnlichkeiten unter dem tf-Modell mit allen Stilähnlichkeiten berechnet. Der durchschnittliche Korrelationskoeffizient liegt bei 0.68. Daraus lässt sich zunächst ablesen, dass, unter den beschriebenen Modellen, eine nicht unwesentliche Korrelation zwischen Stil und Inhalt vorliegt. Allerdings lässt sich daraus keine eindeutige Schlussfolgerung ziehen. Die Korrelation kann zustande kommen, weil die Stilmerkmale viel Inhalt transportieren oder weil das Inhaltsmaß viel Stil transportiert.

Ohne der Möglichkeit, inhaltliche Ähnlichkeit unabhängig von stilistischer Ähnlichkeit zu berechnen, ist es nicht möglich, eine Unabhängigkeit der Stilmaße vom Inhalt rechnerisch nachzuweisen. Eine Möglichkeit bestünde in der Konstruktion eines Korpus, für dessen Texte die inhaltliche Ähnlichkeit manuell von Menschen eingeschätzt würden.

7.4 Von Unmaksing zu ST-Unmasking

In diesem Abschnitt wird aus Unmasking, wie es von Koppel und Schler in [12] beschrieben wurde, die Verallgemeinerung ST-Unmasking entwickelt. Dazu werden Schritt für Schritt einzelne Komponenten ausgetauscht und die Veränderungen experimentell untersucht. Dabei ist insbesondere das Verhalten der Klassifikationsgüte bei abnehmender Textlänge von Interesse.

Experiment 4 (*Ex4*): Unmasking.
Abbildung 7.2 zeigt die Klassifikationsgüte von Unmasking in der originalen Konfiguration bezüglich verschiedener Textlängen. Es ist deutlich zu erkennen, dass die Klassifikationsgüte ab einer Textlänge von 20 000 Wörtern aufwärts konstant hoch ist. Unterhalb von 20 000, insbesondere unterhalb von 10 000

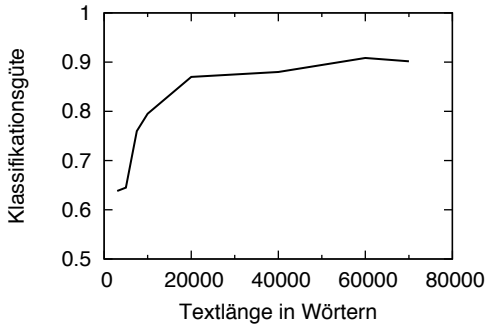
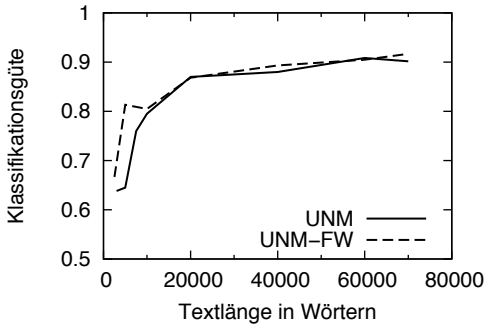


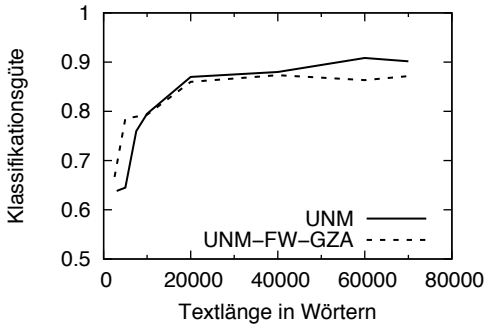
Abbildung 7.2: Diese Grafik zeigt die Klassifikationsgüte von Unmaskings bezogen auf verschiedene Textlängen.

Wörtern sinkt die Klassifikatonsgüte drastisch. Neben der generell zu erwartenden sinkenden Klassifikationsgüte bei kürzer werdenden Texten kann davon ausgegangen werden, dass speziell die sinkende Anzahl an Textblöcken innerhalb der Berechnung der Koppel-SVM-Klassifikationsgüte problematisch ist. Bei 2 500 Wörtern pro Text sind nur noch 5 Blöcke pro Text vorhanden. Dadurch sinkt sowohl die Güte des internen Klassifikators, als auch die Genauigkeit der 10-fachen Kreuzvalidierung und damit der Koppel-SVM-Klassifikationsgüte.

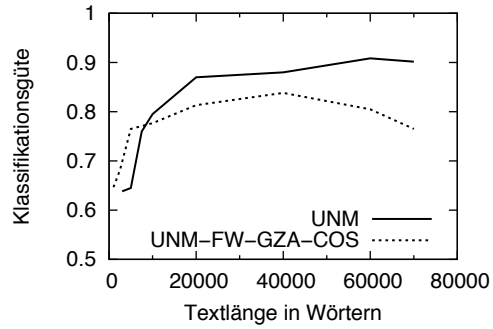
Experiment 5 (*Ex5*): Funktionswörter aus einem Wörterbuch.
Im Unmasking wird als Stilmerkmal die Häufigkeitsverteilung der 250 häufigsten Wörter verwendet, die in beiden Texten vorkommen. Bei kurzen Texten sind das im extremen Fall alle Wörter und damit eine stark inhaltsbezogene Repräsentation. Aus diesem Grund wird für die Autorenschaftsverifikation von kurzen Texten die Verwendung eines Wörterbuches zur Bestimmung der Funktionswörter vorgeschlagen. Zunächst wird getestet, wie sich die Verwendung eines Wörterbuches anstelle der 250 häufigsten Wörter zur Bestimmung der Funktionswörter als Stilmerkmal auswirkt. In Abbildung 7.3(a) sind beide Varianten gegenübergestellt. Interessant ist die verbesserte Klassifikationsgüte im Bereich kürzerer Texte, bei längeren Texten ergeben sich keine nennenswerten Unterschiede.



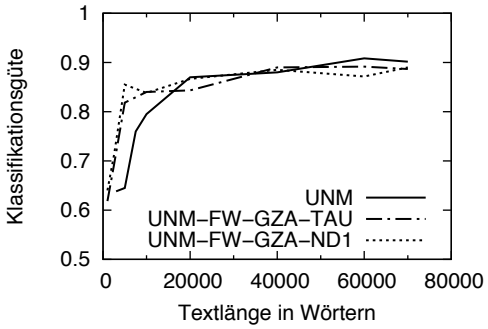
(a) Unmasking (UNM) im Vergleich mit einer Variation, bei der 250FW durch FW ersetzt wurde.



(b) UNM und eine Variation mit FW und GZA anstelle von KFR.



(c) UNM und eine Variation mit FW, GZA und COS.



(d) UNM und eine Variation mit FW, GZA und TAU bzw. ND1.

Abbildung 7.3: In diese Grafik ist die Klassifikationsgüte von Unmasking mit der Klassifikationsgüte verschiedener Varianten davon, mit geänderten Komponenten gegenübergestellt.

Experiment 6 (*Ex6*): Projektionsstrategie.
Weiterhin wird auf dem Weg zu ST-Unmasking zusätzlich das Koppel-Feature-Removal durch die geordnete Zufallsauswahl ersetzt, um beide Verfahren zu vergleichen. Dieses Experiment ist in Abbildung 7.3(b) dargestellt. Im Bereich längerer Texte ist die Klassifikationsgüte für das Koppel-Feature-Removal konstant etwas größer.

Experiment 7 (*Ex7*): Variation der Ähnlichkeitsmaße.
Im nächsten Schritt wird zusätzlich die Koppel-SVM-Klassifikationsgüte durch jeweils eines der Maße Kosinus, Kendalls Tau und Stamatatos-ND1 ersetzen.

Dadurch ist insbesondere die Autorenschaftsverifikation von kurzen Texten möglich, da die Textlänge nur durch die Koppel-SVM-Klassifikationsgüte nach unten begrenzt war. In Abbildung 7.3(c) ist die Gegenüberstellung von Unmasking und der Variation mit der Kosinus-Ähnlichkeit dargestellt. Die Klassifikationsgüte der Variation ist bei kurzen Texten etwas größer, bei längeren Texten aber deutlich schlechter als beim originalen Unmasking. Abbildung 7.3(d) zeigt die Variation mit Kendalls Tau und der Stamatatos-ND1. Beide Variationen haben bei kurzen Texten eine deutlich größere Klassifikationsgüte und verhalten sich bei längeren Texten vergleichbar zum unmodifizierten Unmasking. Ein weiterer positiver Nebeneffekt ist eine deutliche Beschleunigung der Laufzeit durch Verwendung der algorithmisch einfacheren Ähnlichkeitsmaße. Die ursprüngliche Variante mit der Koppel-SVM-Klassifikationsgüte ist etwa um Faktor 10 langsamer.

7.5 Evaluierung von ST-Unmaksing

ST-Unmasking wird als Framework verstanden, das je nach individueller Anwendung zusammengestellt werden kann. Für verschiedene Anwendungen sind gegebenenfalls die Parameter zu optimieren. Da davon ausgegangen werden kann, dass die meisten Parameter nicht unabhängig voneinander sind, ist die Suche nach der optimalen Konfiguration schwer, insbesondere durch die große Parameteranzahl. Es ergeben sich hunderte von Kombinationsmöglichkeiten, von denen jede einzeln getestet werden müsste, um aus dieser gesamten Liste die beste Variante für die individuelle Aufgabe zu ermitteln. Manuell ist das nur mit sehr großem Aufwand möglich. Deshalb wurde für die Evaluierung von einer Unabhängigkeit der Parameter ausgegangen, und jeder Parameter einzeln untersucht.

Experiment 8 (*Ex8*): Vergleich der Ähnlichkeitsmaße.
Die gleichzeitige Verwendung mehrerer Ähnlichkeitsmaße ist nur dann von Vorteil, wenn sich diese unterschiedlich verhalten. Dann werden, in der vorgeschlagenen Terminologie, die Ähnlichkeitswerte unter verschiedenen Ähnlichkeitsmaßen als verschiedene Aspekte der stilistischen Ähnlichkeit betrachtet. In Tabelle 7.2 ist die Klassifikationsgüte unter den drei Ähnlichkeitsmaßen Ko-

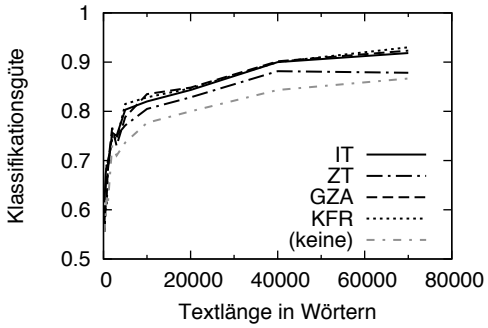
Stilmerkmal	**Kosinus**	**Kendall**	**Stamatatos**	*zusammen*
Zeichen-Trigramme	0.77	**0.81**	0.78	*0.85*
POS-Triramme	0.75	0.78	**0.82**	*0.87*
Vokal-Konsonant-6-Gramme	**0.67**	0.63	0.63	*0.72*

Tabelle 7.2: Trennschärfe der Ähnlichkeitsmaße COS, TAU und ND1 für drei Stilmerkmale.

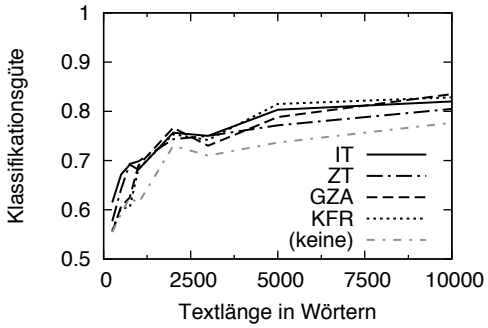
sinus, Kendalls Tau und Stamatatos-ND1 exemplarisch für drei Stilmerkmale aufgetragen. Je nach Stilmerkmal ergibt eines dieser Maße eine größere Klassifikationsgüte. Daran zeigt sich, dass sich die Ähnlichkeitsmaße unterschiedlich verhalten, und je nach verwendeten Stilmerkmalen und Projektionsstrategien eines der Maße geeigneter sein kann.

Experiment 9 (*Ex9*): Vergleich der Projektionsstrategien
In diesem Experiment werden die vier vorgestellten Projektionsstrategien, intuitive Trennung, zufällige Trennung, Koppel-Feature-Removal und geordnete Zufallsauswahl zur Projektion des Merkmalsraums in mehrere Unterräume verglichen. Zu vermuten wäre, dass die beiden ersten Verfahren, die als naiv gelten können, schlechtere Resultate erzielen. Abbildung 7.4(a) zeigt die Ergebnisse des Experiments. Zu erkennen ist, dass die zufällige Trennung im Bereich längerer Texte etwas schlechtere Klassifikationsergebnisse erzielt. Die Kurven der drei anderen Varianten liegen so dicht beieinander, dass von keinem nennenswerten Unterschied gesprochen werden kann. Als Baseline dient eine ST-Unmasking Variante ohne jeglicher Projektionsstrategie, dort wird jeweils nur eine Ähnlichkeit im Stilmerkmalsraum berechnet. Da drei Ähnlichkeitsmaße verwendet werden, haben die Stilähnlichkeitsvektoren lediglich drei Dimensionen. Erwartungsgemäß ist diese Variante den anderen unterlegen. In Abbildung 7.4(b) sind aus dem selben Experiment alle Kurven im Bereich kürzerer Texte dargestellt. Im vordersten Bereich liegt die intuitive Trennung am höchsten.

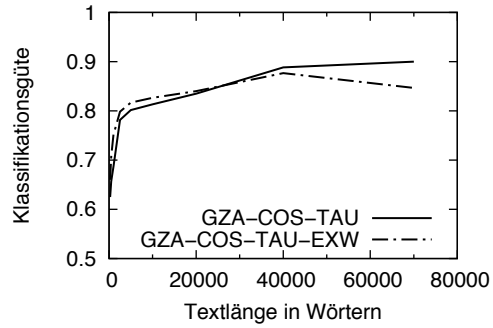
Für die einzelnen Projektionsstrategien wurden Parameter gewählt, die sich als günstig erwiesen haben. So wurde das Koppel-Feature-Removal mit 100 Iterationen durchgeführt, bei jeder Iteration wurden die trennschärfsten 0.2% der Dimensionen im Merkmalsraum (etwa 3000 Dimensionen) entfernt. In der ge-



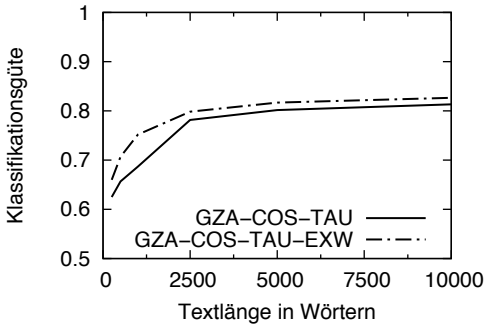
(a) Alle vier Projektionsstrategien IT, ZT, GZA und KFR im Vergleich, als Baseline: keine Projektionsstrategie.



(b) Alle vier Projektionsstrategien im Bereich kürzerer Texte, als Baseline: keine Projektionsstrategie.



(c) ST-Unmasking mit und ohne Verrechnung externen Wissens (EXW).



(d) ST-Unmasking mit/ohne EXW im Bereich kürzerer Texte.

Abbildung 7.4: Die Grafik zeigt die Klassifikationsgüte von ST-Unmasking in der Variation der Projektionsstrategien und der Verwendung von externem Wissen.

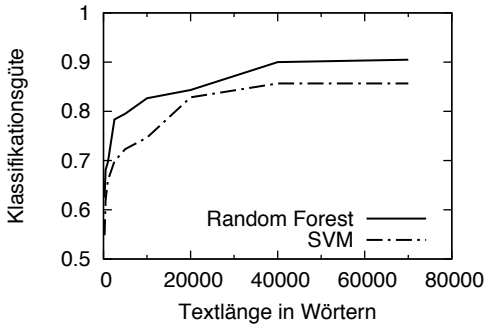
ordneten Zufallsauswahl wurden 100 Stichproben gezogen, jede einzelne hatte dabei eine Größe von 5% der Dimensionen im Merkmalsraum. Für die zufällige Trennung wurden 10 Unterräume eingestellt, um eine Vergleichbarkeit zur intuitiven Trennung herzustellen, die durch die Verwendung aller 10 Stilmerkmale ebenfalls 10 Unterräume ergab.

Experiment 10 (*Ex10*): Externes Wissen – Erwartungswerte
Als eine Möglichkeit, die Qualität der paarweisen Autorenschaftsverifikation von kurzen Texten zu verbessern, wurde in Kapitel 6 vorgeschlagen, externes Wissen in den Entscheidungsprozess einfließen zu lassen. Konkret wird in die-

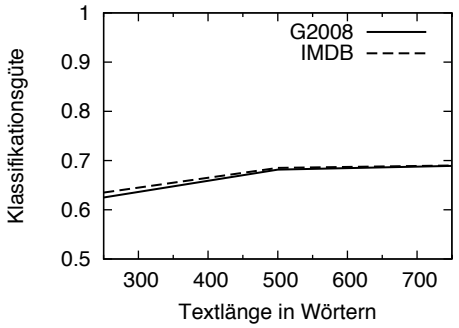
sem Experiment der Erwartungswert jeder einzelnen Stilmerkmalsausprägung von der tatsächlichen Ausprägung subtrahiert. Die Erwartungswerte werden durch die relative Häufigkeit jeder Stilmerkmalsausprägung über 10 000 000 Wörter im G2008-Korpus abgeschätzt. Damit besteht der Stilvektor jedes untersuchten Textes aus der Abweichung vom durchschnittlichen Stil im Korpus. Abbildung 7.4(c) zeigt die Ergebnisse über die Textlängen von $250 - 70\,000$ Wörtern einmal für die Variante mit externem Wissen, einmal ohne. Ab etwa unterhalb von 20 000 Wörtern liegt die Klassifikationsgüte der Variante mit externem Wissen etwas höher, insbesondere unterhalb von 10 000 Wörtern, wie in Abbildung 7.4(d) deutlicher zu erkennen ist. Interessant, wenn auch nicht wünschenswert, ist die Beobachtung oberhalb von 40 000 Wörtern. Dort sinkt die Klassifikationsgüte wieder deutlich ab, verglichen mit der Variante ohne externem Wissen. Dieses Verhalten ist konstant auch bei geänderten Einstellungen der anderen Parameter zu beobachten. Zu vermuten wäre, dass mit steigender Textlänge der Schreibstil der Autoren nur noch wenig vom erwarteten Stil abweicht, und viele Dimensionen in den Stilvektoren Null oder fast Null werden.

Experiment 11 (*Ex11*): Verschiedene Meta-Klassifikatoren
In [12] wird als Meta-Klassifikator eine SVM verwendet. In der vorliegenden Arbeit wird Random Forest als Meta-Klassifikator vorgeschlagen, da dieser Klassifikator in den Experimenten mit Unmasking und ST-Unmasking konstant bessere Ergebnisse lieferte. Abbildung 7.5(a) zeigt exemplarisch für eine Konfiguration den Unterschied zwischen SVM und Random Forest. Über den gesamten Bereich aller untersuchter Textlängen hinweg ist die Klassifikationsgüte von Random Forest deutlich über der der SVM.

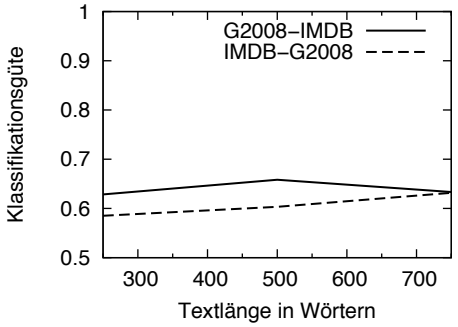
Experiment 12 (*Ex12*): Verschiedene Korpora
Als weitere Stufe der Evaluierung von ST-Unmasking werden nun Experimente über mehrere Korpora hinweg durchgeführt. Zunächst wird die Klassifikationsgüte auf dem IMDB-Korpus getestet. Bezogen auf kurze Texte kann diese Datenquelle eventuell als realistischer angesehen werden, da die Texte tatsächlich kaum länger sind als die extrahierten Textblöcke. Aus dieser Eigenschaft begründet sich jedoch auch, dass nur Textlängen bis 750 Wörter untersucht werden können. Abbildung 7.5(b) zeigt die entsprechenden Klas-



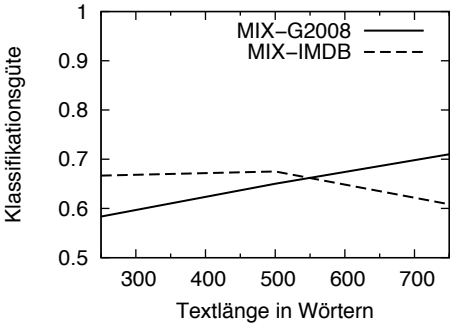
(a) Variation des Meta-Klassifikators.



(b) Klassifikationsgüte auf IMDB im Vergleich zu G2008.



(c) Training auf IMDB und Test auf G2008 (IMDB-G2008) im Vergleich zur umgekehrten Variante (G2008-IMDB).



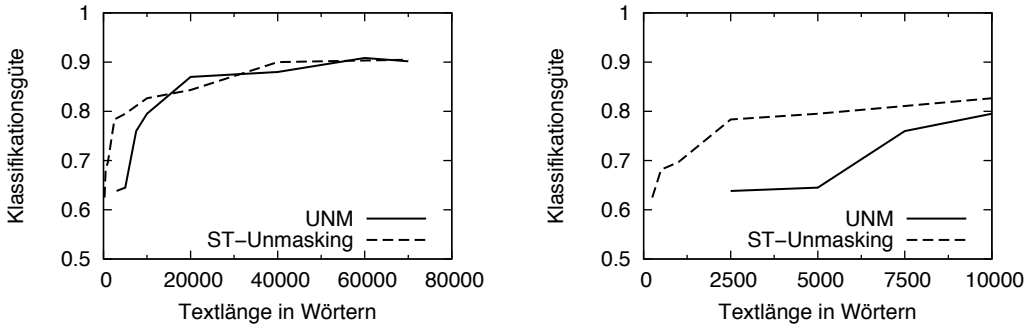
(d) Training auf gemischter Trainingsmene (MIX) und Test auf IMDB bzw. G2008.

Abbildung 7.5: Es ist die Klassifikationsgüte von ST-Unmasking mit verschiedenen Meta-Klassifikatoren bzw. auf verschiedenen Korpora abgebildet.

sifikationsgüten, als Vergleich ist die Kurve für den G2008-Korpus berechnet worden. Die Unterschiede sind marginal.

Von besonderem Interesse sind die folgenden Kreuz-Tests. Dabei wird mit der Trainingsmenge eines Korpus trainiert und mit der Testmenge des anderen Korpus getestet. Auf diese Weise lässt sich die Abhängigkeit des trainierten Klassifikators von der Domäne feststellen. Abbildung 7.5(c) zeigt die Klassifikationsergebnisse mit beiden Kreuz-Test-Varianten. Die Ergebnisse sind erwartungsgemäß schlechter als in dem Fall, dass Trainings- und Testmenge aus dem selben Korpus stammen.

Weiterhin wird aus beiden Korpora durch Zufallsauswahl eine gemischte



(a) Unmasking vs. ST-Unmasking.

(b) Unmasking vs ST-Unmasking im Bereich kürzerer Texte.

Abbildung 7.6: ST-Unmasking im Vergleich mit Unmasking.

Trainingsmenge erstellt um anschließend jeweils auf der Testmenge von IMDB und G2008 die Klassifikationsgüte zu berechnen. In beiden Fällen verhält sich die Klassifikationsgüte unvorhersehbar (Abbildung 7.5(d)).

ST-Unmasking vs. Unmasking

Abschließend wird eine gute Konfiguration von ST-Unmasking mit dem originalen Unmasking verglichen. Dazu sind in Abbildung 7.6 die Ergebnisse aus den Experimenten Ex1 und Ex11 aufgetragen. Unmasking kann nur bis zu einer Textlänge von 2 500 Wörtern durchgeführt werden, jedoch ist selbst im Bereich von $2\,500 - 7\,500$ Wörtern ein deutlicher Unterschied in der Klassifikationsgüte zu erkennen. Zwischen 2 500 und 5 000 Wörtern ist die Klassifikationsgüte von ST-Unmasking um 14 Prozentpunkte höher.

Kapitel 8

Diskussion und Ausblick

In dieser Arbeit konnte gezeigt werden, dass mit ST-Unmasking sicherlich ein Ansatz zur paarweisen Autorenschaftsverifikation von kurzen Texten geschaffen wurde, der Potential enthält. Das bestehende Verfahren Unmasking, zur paarweisen Autorenschaftsverifikation von langen Texten, wurde dazu analysiert und durch Abstraktion der einzelnen Schritte ein generalisiertes Framework für die paarweise Autorenschaftsverifikation von langen und kurzen Texten entwickelt. Für jeden der Schritte im Unmasking wurden Alternativen entwickelt und untersucht. Dabei lag der Schwerpunkt auf der Möglichkeit, kurze Texte verarbeiten zu können, was mit Unmasking prinzipbedingt nicht möglich ist. Es wurde gezeigt, dass die paarweise Autorenschaftsverifikation mit Texten der Länge von weniger als 10 000 Wörtern mit den vorgeschlagenen Instanziierungen von ST-Unmasking deutlich größere Klassifikationsgüten erreicht als mit Unmasking. Bei der Textlänge von 2 500 Wörtern liegt die Klassifikationsgüte von ST-Unmasking mit 78% um 14.5 Prozentpunkte über der von Unmasking. Weiterhin können prinzipiell beliebig kurze Texte verarbeitet werden. Jedoch sinkt auch die Klassifikationsgüte von ST-Unmasking deutlich unterhalb einer Schwelle von etwa 2 500 Wörtern. Auf Texten der Länge von 250 und 500 Wörtern wurden 66% und 72% richtig klassifizierte Textpaare gemessen. Unter den neu entwickelten Stilmerkmalen erweisen sich speziell die Phonem-Bigramme und die 2-Suffix-Bigramme in Bezug auf die Trennschärfe als qualitativ gleichwertig zu den besten bekannten Stilmerkmalen.

In [12] messen Koppel und Schler mit Unmasking eine Klassifikationsgüte

von 95.7% auf ganzen Büchern mit einer durchschnittlichen Länge von etwa 60 000 Wörtern. In der vorliegenden Arbeit lag der beste Wert für das originale Unmasking bei 90.1% und 70 000 Wörtern. Dieser Unterschied ist nicht unwesentlich und wird deshalb an dieser Stelle diskutiert. Es können mehrere Unterschiede in der Vorgehensweise identifiziert werden: Der erste wesentliche Unterschied liegt in der Experimentieranordnung von Koppel. Für 21 Bücher von 10 Autoren wird jeweils eine Leave-One-Out-Kreuzvalidierung durchgeführt. Damit sind immer auch Texte des zu testenden Autors in der Trainingsmenge. In der vorliegenden Arbeit wurde es grundsätzlich vermieden, dass Texte eines Autors sowohl in der Trainings- als auch in der Testmenge vorhanden sind. Weiterhin ist es möglich, dass die hier durchgeführte Eliminierung sämtlicher Formatierungen die Klassifikationsgüte negativ beeinflusst. Als letzter Punkt lässt sich anführen, dass der von Koppel verwendete Datensatz mit 21 Texten relativ klein ist, so dass die Ergebnisse statistisch nicht repräsentativ sein müssen. Im Allgemeinen ist es schwierig, Vergleiche zu anderen Arbeiten zur paarweisen Autorenschaftsverifikation zu ziehen. Dafür gibt es mehrere Gründe: Oft sind die verwendeten Korpora nicht verfügbar oder die konkrete Auswahl der Texte ist unklar. Weiterhin sind teilweise die Algorithmen und Parameter nicht im Detail rekonstruierbar. Häufig sind die Ansätze verschiedener Arbeiten zwar ähnlich, jedoch nicht unmittelbar vergleichbar.

Eine Frage, die in dieser Arbeit nicht in ausreichendem Maße beantwortet werden konnte, ist die, nach der Unabhängigkeit der Stilmerkmale vom Inhalt. Zentrales Problem ist es, eine geeignete Repräsentation für den Inhalt zu finden, die keine stilistischen Besonderheiten des Textes misst. Um dieses Problem zu umgehen, könnte ein anderer Ansatz verfolgt werden: Es könnte ein Korpus manuell konstruiert werden, der einerseits aus inhaltlich verschiedenen Textpaaren mit gleicher Autorenschaft, und andererseits aus inhaltlich gleichen Textpaaren mit verschiedener Autorenschaft besteht. Stilmerkmale, die auf einem solchen Korpus funktionieren, können als weitgehend unabhängig von Inhalt der Texte gelten. Es lässt sich jedoch festhalten, dass die vorgestellten, bekannten Stilmerkmale in der Forschung etabliert sind. Auch wird teilweise die Meinung vertreten, dass Stilmerkmale auch dann verwendet werden können, wenn sie Informationen über den Inhalt transportieren, sofern

das in der untersuchten Domäne unproblematisch ist. So werden zum Beispiel in [5] alle Wörter eines Textes als Stilmerkmal verwendet. In [14] werden alle Zeichen-4-Gramme verwendet, wobei der Grad an Inhalt, den Zeichen-4-Gramme transportieren, als relativ hoch eingeschätzt wird.

Ein anderer, die Stilmerkmale betreffender Aspekt, ist die Benutzung von Meta-Informationen als Stil- oder eher Autorenmerkmal. Um die Autorenschaftsverifikation von kurzen Texten weiter zu verbessern, wird es notwendig sein, individuelle Meta-Informationen, sowie kontext- und domänenspezifische Merkmale zu verwenden. Für eine konkrete Aufgabe liegt die Herausforderung sicherlich eher in der Entwicklung geeigneter Merkmale.

Aufgrund der großen Anzahl abhängiger Parameter, die ST-Unmasking enthält, ist es sehr aufwendig besonders gute Einstellungen für eine konkrete Aufgabe zu finden. Als zukünftige Aufgabe könnte untersucht werden, inwiefern eine Optimierung der Parameter automatisch in einem Lernprozess geschehen kann, zum Beispiel die automatische Selektion von Stilmerkmalen, die bezüglich einer Trainingsmenge besonders geeignet sind.

Ein bislang wenig beachtetes, angrenzendes Forschungsgebiet, in dem die paarweise Autorenschaftsverifikation unmittelbar verwendet wird, ist das Clustering von Autoren. Auf einer Menge von Texten wird die paarweise Autorenschaftsverifikation angewendet. Diese Menge kann als Graph dargestellt werden: Jeder Text ist ein Knoten und zwischen zwei Texten ist genau dann eine Kante, wenn sie als *selber-Autor* klassifiziert wurden. Im Idealfall bilden sich mehrere nicht zusammenhängende Teilgraphen, wobei jeder Teilgraph einen Autor darstellt. Geeignete Clusteralgorithmen können gegebenenfalls Fehler ausgleichen, die bei der Autorenschaftsverifikation entstanden sind. Textpaare, die sich in einem Teilgraphen befinden, aber nicht durch eine Kante direkt verbunden sind, können dann dennoch als vom selben Autor stammend erkannt werden. Besonders interessant kann dieser Ansatz für die intrinsische Plagiatanalyse sein, wenn alle Abschnitte auf diese Weise geclustert werden. Dann spielt es keine Rolle wie groß der Anteil an plagiiertem Text innerhalb eines Dokuments ist, was derzeit ein wesentlicher Parameter für das Gelingen der intrinsischen Plagiatanalyse ist.

Weiterführend könnte ST-Unmasking folgendermaßen untersucht werden: Unter der Argumentation der geordneten Zufallsauswahl, dass es keine Rolle spielt in welchem Unterraum die stilistische Ähnlichkeit zweier Texte hoch ist, sondern vielmehr in wie vielen Unterräumen eine hohe oder niedrige Stilähnlichkeit gemessen wurde, könnte es sinnvoll sein den gesamten Stilähnlichkeitsvektor $\mathbf{s}$ zu sortieren. Dann wäre es egal, unter welchem Ähnlichkeitsmaß oder in welchem Unterraum des Stilmerkmalsraums eine Stilähnlichkeit gemessen wurde. Dazu müssten natürlich die verwendeten Ähnlichkeitsmaße tatsächlich die Ähnlichkeit messen (nicht die Distanz) und auf das selbe Intervall abbilden. Möglicherweise sinkt dadurch die Domänenabhängigkeit der Meta-Klassifikation. Erste Versuche diesbezüglich waren erfolgversprechend.

Anhang A

Tabellen zur Evaluierung

Nachfolgend wird in Tabelle A.1 eine Übersicht über alle Parameter von ST-Unmasking, sowie deren Ausprägung und Abkürzung gegeben. Weiterhin sind in Tabelle A.2 die Parameter alle Experimente aus Kapitel 7 aufgelistet.

Typ	**mögliche Ausprägung**	**Abkürzung**
Korpus	Gutenberg 2008	G2008
	IMDB	IMDB
Stilmerkmale	Funktionswörter (Wörterbuch)	FW
	250 häufigste Wörter	250FW
	Zeichen-Trigramme	ZT
	Zeichen-1-Skip-Trigramme	ZST
	POS-Trigramme	POS
	Phonem-Bigramme	PHB
	Vokal-Trigramme	VT
	Vokal-Konsonant-6-Gramme	VK6G
	Wortlängen-Bigramme	WLB
	1-Präfix-Bigramme	PRB
	2-Suffix-Bigramme	SUB
	Alle Stilmerkmale	ALL
Ähnlichkeitsmaße	Kosinus-Ähnlichkeit	COS
	Kendalls Tau	TAU
	Stamatatos-ND1	ND1
	Koppel-SVM-Trennschärfe	SVMT
Projektionsstrategien	Intuitive Trennung	IT
	Zufällige Trennung	ZT
	Koppel-Feature-Removal	KFR
	Geordnete Zufallsauswahl	GZA
Externes Wissen	Verwendung externen Wissens	EXW
	Keine Verwendung	-
Meta-Klassifikator	Random Forest	RF
	SVM	SVM
Textlänge	$250 - 70\,000$	-

Tabelle A.1: Alle Parameter von ST-Unmasking mit den möglichen Ausprägungen, dazu alle zugewiesenen Abkürzungen.

Typ	**Ex1**	**Ex2**	**Ex3**	**Ex4**	**Ex5**	**Ex6**	**Ex7**	**Ex8**	**Ex9**	**Ex10**	**Ex11**	**Ex12**
Korpus	G2008	G2008	G2008	G2008	G2008	G2008	G2008	G2008	G2008	G2008	G2008	G2008 IMDB
Stilmerkmale	ALL	ALL	ALL	250FW	FW	FW	FW	ZT POS VK6G	ALL	ALL	ALL	ALL
Ähnlichkeits-maße	COS TAU ND1	COS	COS	SVMT	SVMT	SVMT	COS TAU ND1	COS TAU ND1	COS TAU ND1	COS TAU	COS TAU ND1	COS TAU ND1
Projektions-strategien	IT	IT	IT	KFR	KFR	GZA	GZA	IT	IT ZT KFR GZA	GZA	GZA	KFR GZA
Externes Wissen	-	-	-	-	-	-	-	-	-	- EXW	-	-
Meta-Klassifi-kator	RF	RF	RF	RF	RF	RF	RF	RF	RF	RF	RF SVM	RF
Textlänge	20 000	20 000	20 000	2 500– 70 000	2 500– 70 000	2 500– 70 000	1 000– 70 000	20 000	250– 70 000	250– 70 000	250– 70 000	250– 750

Tabelle A.2: Auflistung der eingestellten Parameter in allen Experimenten.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Argamon, S. und S. Levitan: *Measuring the usefulness of function words for authorship attribution.* In: *Proceedings of ACH/ALLC*, Band 5. Citeseer, 2005.

[2] Axelsson, M.W.: *USE-the Uppsala Student English corpus: an instrument for needs analysis.* ICAME journal, 24:155–157, 2000.

[3] Binongo, J.N.G.: *Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution.* CHANCE-BERLIN THEN NEW YORK-, 16(2):9–17, 2003.

[4] Breiman, L.: *Random forests.* Machine learning, 45(1):5–32, 2001.

[5] Diederich, J., J. Kindermann, E. Leopold und G. Paass: *Authorship attribution with support vector machines.* Applied Intelligence, 19(1):109–123, 2003.

[6] Graham, N., G. Hirst und B. Marthi: *Segmenting documents by stylistic character.* Natural Language Engineering, 11(04):397–415, 2005.

[7] Halteren, H.V.: *Author verification by linguistic profiling: An exploration of the parameter space.* ACM Transactions on Speech and Language Processing (TSLP), 4(1):1–17, 2007.

[8] Holmes, D.I.: *Authorship attribution.* Computers and the Humanities, 28(2):87–106, 1994.

[9] Holmes, D.I.: *The evolution of stylometry in humanities scholarship.* Literary and Linguistic Computing, 13(3):111, 1998.

[10] Iqbal, F., L.A. Khan, B. Fung und M. Debbabi: *e-mail authorship verification for forensic investigation.* In: *Proceedings of the 2010 ACM Symposium on Applied Computing*, Seiten 1591–1598. ACM, 2010.

[11] Kohavi, R.: *A study of cross-validation and bootstrap for accuracy estimation and model selection.* In: *International joint Conference on artificial intelligence*, Band 14, Seiten 1137–1145. Citeseer, 1995.

[12] Koppel, M. und J. Schler: *Authorship verification as a one-class classification problem.* In: *Proceedings of the twenty-first international conference on Machine learning*, Seite 62. ACM, 2004.

[13] Koppel, M., J. Schler und S. Argamon: *Computational methods in authorship attribution.* Journal of the American Society for information Science and Technology, 60(1):9–26, 2009.

[14] Koppel, M., J. Schler und S. Argamon: *Authorship attribution in the wild.* Language Resources and Evaluation, Seiten 1–12, 2010.

[15] Koppel, M., J. Schler und E. Bonchek-Dokow: *Measuring differentiability: Unmasking pseudonymous authors.* Journal of Machine Learning Research, 8:1261–1276, 2007.

[16] Sharples, M. und T. van der Geest: *Detecting stylistic inconsistencies in collaborative writing.* To appear in The new writing environment: Writers at work in a world of technology, 1995.

[17] Stamatatos, E.: *A survey of modern authorship attribution methods.* Journal of the American Society for information Science and Technology, 60(3):538–556, 2009.

[18] Stamatatos, E.: *Intrinsic Plagiarism Detection Using Character n-gram Profiles.* In: *3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE*, Band 2, Seite 38, 2009.

[19] Stein, B. und S.M. zu Eissen: *Intrinsic plagiarism analysis with meta learning.* In: *Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection*, Seiten 45–50. Citeseer, 2007.

[20] Stein, B., N. Lipka und S.M. zu Eissen: *Meta analysis within authorship verification.* In: *19th International Conference on Database and Expert Systems Application*, Seiten 34–39. IEEE, 2008.