Hochschule Wismar
University of Technology, Business and Design
Fakultät für Ingenieurwissenschaften

Master-Thesis

Automatisierte Autorschaftsanalyse in der deutschen Sprache mittels forensischer Linguistik

Abschlussarbeit zur Erlangung des Grades eines

Master of Engineering

der Hochschule Wismar

Eingereicht von:	Manuel Dorobek
	Studiengang IT-Sicherheit und Forensik
Erstgutachter:	Frau Prof. Dr.-Ing. Antje Raab-Düsterhöft
Zweitgutachter:	Herr Prof. Dr.-Ing habil. Andreas Ahrens

Östringen, den 7. Januar 2021

Kurzreferat

Im kriminalistischen Kontext ergibt sich oftmals die Frage nach dem Autor eines anonymen Dokuments. Beispielsweise muss der Verfasser eines Erpresserschreibens innerhalb einer Gruppe von Verdächtigen identifiziert werden. Neben den manuellen Methoden existieren in der englischen Sprache bereits erfolgreiche Verfahren, die Autorschaftsanalyse zu automatisieren.

In der deutschen Sprache hingegen sind kaum Veröffentlichungen in diesem Bereich zu finden. Aufgrund von sprachlichen Eigenheiten ergibt sich die Forschungsfrage, ob eine automatisierte Autorschaftsanalyse auch in der deutschen Sprache möglich ist. Zu diesem Zweck soll ein Klassifikationsmodell für die deutsche Sprache entwickelt werden und eine Bewertung der existierenden Merkmale an diesem erfolgen.

Nach einem Vergleich von drei Algorithmen fiel die Wahl des Klassifikators auf Support Vector Machines. Anhand einer Datengrundlage deutscher Musikreviews konnten Klassifikationsergebnisse von 96,4% bei der Unterscheidung von 25 Autoren unter der Verwendung von 60 Texten als Trainingsdaten erzielt werden. In der Bewertung der Merkmale zeigten sich „Zeichen-Affix-4-Gramme“ als effektivstes Merkmal. Themenabhängige Merkmale konnten aufgrund einer starken Korrelation unter gleichbleibenden Ergebnissen ausgelassen werden.

Abstract

Automated Authorship Analysis in German Language Using Forensic Linguistics

In a criminal context, the question of the author of an anonymous document often arises. For example, the author of an extortion letter must be identified within a group of suspects. In addition to manual methods, there already exist successful methods for automating authorship analysis in the English language.

In the German language, however, there are hardly any publications in this area. Due to linguistic peculiarities, the research question arises whether an automated authorship analysis is also possible in the German language. For this purpose, a classification model for the German language will be developed and an evaluation of the existing features will be performed using it.

After a comparison of three algorithms, the choice of the classifier fell on Support Vector Machines. Using a database of German music reviews, classification results of 96.4% could be achieved when distinguishing 25 authors using 60 texts as training data. In the feature evaluation, "Char-Affix-4-Grams" showed to be the most effective feature. Content-dependent features could be omitted due to a sustained strong correlation and consistent results.

Inhalt

1. Einleitung 8
2. Vorbetrachtungen 9
2.1. Stand der Forschung 9
2.2. Definition der Forschungsfrage Ziele der Arbeit 11
2.3. Vorgehensweise 12
3. Autorschaftsanalyse 16
3.1. Kernbereiche 17
3.2. Paradigmen 18
3.3. Herausforderungen und Anforderungen an die Datengrundlage 19
4. Natural Language Processing 22
4.1. Python Bibliotheken 22
4.2. Tokenization 23
4.3. Textnormalisierung 24
4.3.1. Stammformreduktion und Lemmatisierung 25
4.3.2. Inhalts- und Funktionswörter 25
4.3.3. Stoppwörter 26
5. Datenverständnis 27
5.1. Datengrundlage 27
5.2. Merkmalskategorien 29
5.3. Bekannte Merkmale in der Forschung 31
5.4. Ausgeschlossene Merkmale 33
5.5. Themenabhängige Merkmale 35
5.6. Ausgewählte Merkmale 36
5.6.1. Bag of Words 37
5.6.2. Wort-N-Gramme 39
5.6.3. Anzahl der Wörter 41
5.6.4. Wortlänge 43
5.6.5. Wortschatzreichtum 45

5.6.6. Sonderzeichen ... 49
5.6.7. Zeichen-N-Gramme ... 51
5.6.8. Ziffern ... 55
5.6.9. Funktionswörter ... 56
5.6.10. Wortarten ... 58
5.6.11. Wortarten-N-Gramme ... 60
5.6.12. Satz-Anfänge/-Endungen ... 62
5.6.13. Verständlichkeitsindizes ... 64
6. Grundlagen maschinelles Lernen ... 67
6.1. Verwendete Bibliotheken ... 67
6.2. Gruppierung von Funktionen ... 68
6.3. Mathematische Grundlagen des überwachten Lernens ... 70
6.4. Multi-Label-Klassifikation ... 72
6.5. Klassifikationsalgorithmen ... 73
6.5.1. K-Nearest-Neighbor ... 73
6.5.2. Naïve Bayes ... 76
6.5.3. Support Vector Machines ... 79
6.6. Bias und Varianz Dilemma ... 84
6.7. Klassifikationsmetriken ... 85
6.8. Kreuzvalidierungsverfahren ... 89
6.8.1. Holdout Methode ... 90
6.8.2. K-fache Kreuzvalidierung ... 90
6.8.3. Hyperparameteroptimierung ... 92
6.9. Label-Kodierung ... 92
6.10. Skalierung der Daten ... 92
6.11. Verringern der Dimensionalität ... 94
6.11.1. Filter ... 94
6.11.2. Wrapper ... 96
6.11.3. Eingebettete Methoden ... 97
6.11.4. Hauptkomponentenanalyse ... 97

7. Datenvorbereitung 104
7.1. Vorverarbeitung 104
7.1.1. Merkmalsausprägungen mit geringer Häufigkeit 105
7.1.2. Merkmalsausprägungen mit hoher Häufigkeit 107
7.1.3. Aufteilung in Test- und Trainingsdatensätze 109
7.1.4. Individuelle relative Häufigkeit 109
7.2. Merkmalsauswahl 112
7.2.1. Merkmalsauswahl durch einen iterativen Filter 113
7.2.2. Bewertung der Ergebnisse des iterativen Filters 118
7.3. Merkmalsvarianten 120
7.3.1. Extraktionsmethode Wortlänge 121
7.3.2. Extraktionsmethode Ziffern 124
7.3.3. Bestimmen der Länge von Wort-N-Grammen 125
7.3.4. Bestimmen der Länge von Zeichen-N-Grammen 128
7.3.5. Wortarten-N-Gramme 133
8. Modellierung 135
8.1. Wahl des Klassifikators 135
8.2. Zusammenfassung der verwendeten Merkmale 136
8.3. Optimieren der Parameter 137
8.4. Wahl der Skalierungsmethode 138
8.5. Bewertung der Merkmale 140
8.5.1. Bewertung der individuellen Merkmale 140
8.5.2. Differenzfunktion 142
8.5.3. Bewertung themenabhängiger Merkmale 144
8.5.4. Zusammenfassende Bewertung der Merkmale 146
8.6. Klassifikationswerte des Modells 146
9. Evaluierung 148
9.1. Benötigte Rechenleistung 148
9.2. Bewertung der Ergebnisse 149
9.3. Ausblick 153

Literaturverzeichnis 155

Bildverzeichnis 171

Tabellenverzeichnis 172

Anlagen 173

A1 Mailverkehr: Re Masterarbeit Autorenschaftserkennung.msg 173

A2 Python Skript: Scraper.py 173

A3 Python Skript: Extractor.py 173

A4 Python Skript: Processor.py 174

A5 CSV-Datei: musikreviews.csv 174

A6 CSV-Datei: musikreviews_balanced_authors.csv 174

A7 Daten 175

Verzeichnis der Abkürzungen 177

Thesen 178

Selbstständigkeitserklärung 179

1. Einleitung

Schon 1996 schrieb die deutsche Philologin[1] und Politologin Heidrun Deborah Kämper über den Nachweis der Autorenschaft in einem Fall über Ausgleichsansprüche. Unter anderem wurde der nicht nachweisbare Besitz eines Aktienpaketes behauptet. Die Antragstellerin fand während des langanhaltenden Verfahrens einen angeblichen Brief ihres Schwiegervaters aus dem Jahre 1948. Eine Textanalyse des Briefs und Vergleichsmaterials konnte beweisen, dass es sich bei diesem Brief um eine Fälschung handeln musste. (Kämper, 1996) Bis heute wird der Rat von Experten auf dem Gebiet der forensischen Linguistik herangezogen, um in Fällen zu ermitteln und eine Aussage darüber zu treffen, ob ein Text von einem bestimmten Autor geschrieben wurde oder um ein Profil des Autors zu erstellen (Bauer, 2018).

Eine standardisierte Methode existiert dabei in der Autorenerkennung nicht (Dern, 2003). Experten verwenden hierzu eine Reihe von verschiedenen Ansätzen, welche auf der Untersuchung der Orthographie, Morphologie, der Syntax oder der Lexik basieren und diese qualitativ bewerten (ebd.). Neben diesen Verfahren existieren auch quantitative Ansätze, die durch statistische Methoden und dem Definieren sprachlicher Merkmale versuchen, die Charakteristik eines Autors zu erfassen und dadurch dessen individuellen Stil zu ermitteln (Bönninghoff, et al., 2019, S. 695) In der englischen Sprache existieren bereits einige automatisierte Verfahren unter diesem Ansatz.

Die Vorteile einer Automatisierung sind dabei zum einen eine Entlastung der Experten und zum anderen eine Klassifikation aufgrund einer unbeeinflussten Objektivität der Algorithmen (Liu Y. H., 2019, S. 9). In der deutschen Literatur existieren kaum Arbeiten der automatisierten Autorschaftsanalyse. Aus diesem Grund ist das Forschungsziel dieser Arbeit ein solches Modell für die deutsche Sprache zu entwickeln und eine Bewertung der Merkmale durchzuführen.

[1] Philologie ist die zusammenfassende Bezeichnung für die Sprach- und Literaturwissenschaft einer Sprache

2. Vorbetrachtungen

Zur Vorbereitung der Arbeit erfolgt eine Analyse der Forschung und die Definition der genauen Forschungsfrage. Im Rahmen dieser Vorbetrachtungen werden die methodische Vorgehensweise sowie die sich daraus ergebende Struktur der Arbeit dargestellt.

2.1. Stand der Forschung

Ein guter Überblick über bisherige Forschungsarbeiten im Bereich der Autorschaftsanalyse bietet Stamatatos (2009, S. 1ff.). Nach ihm gehen die ersten Versuche einer Quantifizierung des Schreibstils auf Mendenhall (1887) mit einer Studie über Stücke von Shakespeare zurück. Gefolgt wurde diese Studie von statistischen Arbeiten durch Zipf (1932) und Yule (1939; 1944).

Mehrere Arbeiten bezeichnen Wallace & Mosteller (1964) als Ursprung stilometrischer Verfahren zur Autorschaftserkennung und als einflussreichste Arbeit in diesem Umfeld (Abbasi & Chen, 2008, S. 4; Stamatatos, 2009, S. 1). Ihre Methode basierte auf einer Bayes'schen statistischen Analyse der Häufigkeiten einer kleinen Anzahl gemeinsamer Wörter (z. B. "und", "zu" usw.) und lieferte signifikante Unterscheidungsergebnisse zwischen den Autorenkandidaten (ebd.).

Stilometrische Merkmale sind Attribute, welche den Schreibstil eines Autors beschreiben. Nach dem Erfolg von Wallace & Mosteller (1964) befasste sich die Forschung auf diesem Gebiet damit, solche Merkmale zu finden und den Schreibstil eines Autors zu quantifizieren. So beschrieb Rudman (1998), dass in der Analyse der Autorschaft über 1.000 verschiedene Merkmale vorgeschlagen wurden.

Stamatatos (2009, S. 2) führt weiter an, dass das Hauptproblem der damaligen Verfahren das Fehlen einer objektiven Bewertung der vorgeschlagenen Methoden war. Er beschreibt die wichtigsten methodischen Einschränkungen der Bewertungsverfahren mit den folgenden fünf Punkten:

- Zu lange (meist ganze Bücher), wahrscheinlich stilistisch nicht homogene Textdaten.
- Eine zu kleine Anzahl der möglichen Autoren (meist 2-3).
- Die Evaluationskorpora wurden nicht auf das Thema kontrolliert.
- Die Bewertung der Methoden war hauptsächlich intuitiv.
- Ein Vergleich der Methoden ist durch mangelnde Benchmark Daten schwierig.

Eine Lösung dieser Probleme führt Stamatatos (2009, S. 2) durch den technischen Fortschritt in den späten 1990er Jahren an. Die steigende Menge an elektronischen Texten hatte im Wesentlichen zwei Auswirkungen. Zum einen bestand die Notwendigkeit die Informationen effizient zu verarbeiten, was in leistungsstarken Algorithmen für maschinelles Lernen, Evaluierungsmethoden dieser und in der Entwicklung von Werkzeugen zur Verarbeitung und Analyse von Texten (Natural Language Processing) führte. Zum anderen weckte die Vielzahl an Texten neue Interessensgebiete für die Anwendung der Autorschaftsanalyse (Madigan, et al., 2005). Daher bezeichnet er die darauffolgende Zeit als neue Ära der Technologie der Autorschaftsanalyse, welche durch das Entwickeln praktischer Anwendungen geprägt wird. (Stamatatos, 2009, S. 2)

Mittlerweile existiert eine Vielzahl an Verfahren zur automatisierten Autorschaftsanalyse. Beispielsweise konnten Abbasi & Chen (2008) in einem Verfahren mit kleineren Gruppen mit bis zu 100 möglichen Autoren Erkennungsraten von bis zu 94% erreichen. Für den Nachweis, ob ein Dokument tatsächlich von einem Autor stammt, wurden von Potha & Stamatatos (2014) Ergebnisse mit Erkennungsraten von bis zu 84,5% erzielt.

Neue Ergebnisse im Bereich der Autorschaftsanalyse werden oftmals in den von der CLEF-Initiative veranstalteten „PAN Evauation Labs“ erzielt. Die CLEF-Initiative (Conference and Labs of the Evaluation Forum) ist nach eigener Aussage ein selbstorganisiertes Gremium, dessen Hauptaufgabe die Förderung von Forschung, Innovation und Entwicklung von Informationszugangssystemen mit dem Schwerpunkt auf mehrsprachiger und multimodaler Information mit

verschiedenen Strukturebenen ist (Webis Group, 2020a). Im Rahmen der Veröffentlichungen ihrer geteilten Aufgaben (Webis group, 2020b) und Konferenzen werden eine Vielzahl von guten bis sehr guten Verfahren und verschiedenen Aspekten der Autorschaftsanalyse vorgestellt (Webis group, 2020c).

Auch für einzelne Merkmale innerhalb der Autorschaftsanalyse existiert eine Vielzahl an Studien. Beispielsweise führen Sapkota, Bethard, Montes, & Solorio, (2015) eine genauere Analyse von Zeichen-N-Grammen durch. Ein guter Überblick über den Stand der Forschungen zur Profilierung von Autoren und diesbezüglicher Merkmale zeigt Nini (2015).

Insgesamt handelt es sich bei der Autorschaftsanalyse um ein großes Forschungsgebiet mit einer Vielzahl an Arbeiten. Jedoch stellt der Untersuchungsgegenstand dieser Arbeiten meist die englische Sprache dar. Im Rahmen der Recherche dieser Arbeit konnten keine Veröffentlichung gefunden werden, welche sich mit der Autorschaftsanalyse in der deutschen Sprache befasst. Eine Ausnahme ist die Master-Thesis von Oren (2012). Ebenfalls konnte keine Arbeit über die Bewertung von Merkmalen der Autorschaftsanalyse in der deutschen Sprache gefunden werden.

2.2. Definition der Forschungsfrage Ziele der Arbeit

Wie bereits im vorherigen Kapitel dargelegt, existieren nahezu keine Arbeiten über die Autorschaftsanalyse in der deutschen Sprache. Es besteht die Annahme, dass die deutsche Sprache durch Eigenheiten gekennzeichnet ist, woraus sich wiederum Eigenheiten der verwendeten Merkmale ergeben. Unter dieser Annahme ist es nicht gegeben, dass eine Klassifikation mit bekannten Mitteln in der deutschen Sprache möglich ist.

Daraus resultiert das Ziel der Arbeit ein Modell der Autorschaftsattribution in der deutschen Sprache aufzustellen, welches ermöglicht den Autor eines anonymen Textes aufgrund von einer Datengrundlage zu identifizieren. Weiterhin soll eine

Bewertung der hierfür verwendeten Merkmale erfolgen und gegebenenfalls existierende Eigenheiten dieser aufgezeigt werden.

Als Grundlage des Modells und der verwendeten Merkmale sollen Arbeiten der englischen Sprache dienen. Hierdurch kann eine Auswertung darüber erfolgen, ob Merkmale und Verfahren, welche sich in der englischen Sprache als erfolgreich erwiesen haben, auch in der deutschen Sprache anwendbar sind.

2.3. Vorgehensweise

Der CRISP-DM (Cross Industry Standard Process for Data Mining) ist ein für Datamining entwickeltes Standardmodell, welches nach Vorhies (2016) auch erfolgreich in Data Science Projekten eingesetzt werden kann. Oftmals wird der Prozess für seinen hohen Dokumentationsaufwand und die langsamen Iterationszyklen kritisiert. Dennoch zeigt sich CRISP-DM selbst für komplexe Projekte als eine sehr gute Anleitung (Saltz & Hotz, 2020). Aus diesem Grund werden im Rahmen dieser Arbeit nicht die vollumfänglichen Details des Standardmodells verfolgt, sondern vielmehr werden die Phasen des Prozesses als grundsätzliche Vorgehensweise verstanden. Im Folgenden werden der grundsätzliche Aufbau des CRISP-DM nach Shearer (2000) und die daraus folgende methodische Vorgehensweise dieser Arbeit erläutert.

CRISP-DM gliedert den Lebenszyklus eines Data Mining-Projekts in sechs Phasen: Geschäftsverständnis, Datenverständnis, Datenvorbereitung, Modellierung, Evaluierung und Bereitstellung (vgl. Bild 1). Dabei werden die Phasen nicht ausnahmslos sequenziell abgearbeitet, sondern Rückschlüsse aus Folgephasen in vorherige Schritte eingearbeitet. Beispielsweise können in der Modellierung notwendige Anforderungen an die Datenaufbereitung erkannt werden. Einen Abschluss findet der Prozess immer in der Bereitstellung des erarbeiteten Modells.

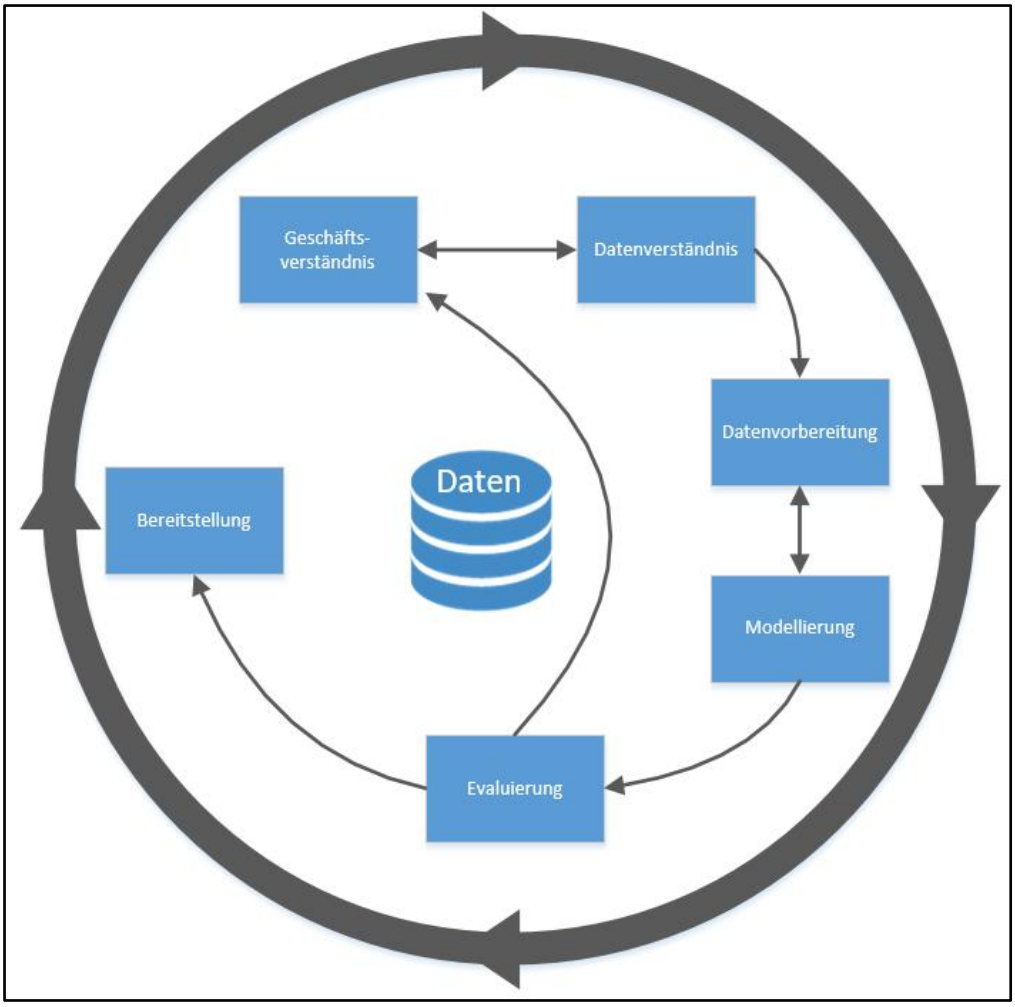


Bild 1: Lebenszyklus nach CRISP-DM - Eigene Darstellung - Anlehnung an (Saltz & Hotz, 2020)

Das **Geschäftsverständnis** konzentriert sich darauf die Projektziele aus der Sicht des Unternehmens zu verstehen und die konkreten Ziele und Anforderungen festzulegen. Für spätere Analysen ist dieses Verständnis unabdingbar. Als Resultat dieser Phase folgt eine Formulierung der Aufgabenstellung und der geplanten Vorgehensweise.

Als Grundlage dieser Arbeit wird in Kapitel 2.2 die Forschungsfrage dieser Arbeit definiert. Für ein tieferes Verständnis werden in Kapitel 3 die grundlegenden Begriffe und Methoden der Autorschaftsanalyse aufgezeigt.

Im **Datenverständnis** soll ein erster Überblick über die zur Verfügung stehenden Daten und deren Qualität erlangt werden. Üblicherweise werden hier Hypothesen über verborgene Informationen aufgestellt. Die Phase des Datenverständnisses

umfasst vier Schritte: Die Sammlung, Beschreibung und Untersuchung der Daten sowie die Überprüfung der Datenqualität.

Da als Grundlage der Arbeit keine Daten vorliegen, wird in dieser Arbeit die Datengrundlage in dieser Phase vorgestellt. Weiterhin werden die Merkmale ausgewählt und aus dieser Grundlage extrahiert. Diese Merkmale werden als Ausgangslage für die eigentliche Phase des Datenverständnisses verwendet.

Die **Datenvorbereitung** umfasst alle Aktivitäten zur Aufbereitung der Daten, welche im Modell verwendet werden. Die fünf Schritte der Datenvorbereitung sind das Selektieren, die Bereinigung, die Konstruktion, die Integration und die Formatierung der Daten.

Der Fokus dieser Phase der Arbeit liegt auf der Auswahl der Daten, beziehungsweise der Reduktion der Dimensionalität. Zu diesem Zweck werden zum einen irrelevante Merkmalsausprägungen gefiltert und zum anderen die Leistung einiger Merkmalsvarianten bewertet und anhand dieser Ergebnisse ausgewählt.

Im Rahmen der **Modellierung** werden verschiedene Techniken ausgewählt und optimale Parameter festgelegt. Einige Techniken haben spezifische Anforderungen an die Form der Daten. Daher kann ein Rückschritt in die vorherige Phase erforderlich sein. In dieser Arbeit wird die Modellierung um die Bewertung der verwendeten Merkmale erweitert.

Die **Evaluierung** prüft das Modell hinsichtlich der Konstruktion und dem Erfüllungsgrad der Ziele. Es erfolgt ein Abgleich der erstellten Datenmodelle. Es soll sichergestellt werden, dass das Modell die Geschäftsziele ordnungsgemäß erreicht und alle Anforderungen berücksichtigt wurden. Die wichtigsten Schritte sind hier die Auswertung der Ergebnisse, die Überprüfung des Prozesses und die Festlegung der nächsten Schritte.

Im Rahmen dieser Arbeit umfasst diese Phase das Bewerten der erarbeiteten Ergebnisse und das Aufzeigen eines darauf basierenden Ausblicks.

Eine **Bereitstellung** des Modells ist in dieser Arbeit nicht vorgesehen. Gleichermaßen wird darauf verzichtet, den entwickelten Programmcode

aufzuzeigen. Eine Beschreibung des Programmcodes erfolgt im Anhang. Geschrieben wurde der Code in Python, spezielle verwendete Bibliotheken werden in den jeweiligen Kapiteln zu den Grundlagen beschrieben.

Ergänzend an diese Phasen werden elementare technische Grundlagen für diese Arbeit erläutert. Dies umfasst Grundlagen des maschinellen Lernens und des Natural Language Processings.

3. Autorschaftsanalyse

Die Autorschaftsanalyse oder Autorenerkennung „ist die linguistische Bewertung fraglicher schriftsprachlicher Texte in forensischen, kriminalistischen oder in sonst einer Form sicherheitsrelevanten Kontexten [sic]“ (Dern, 2009, S. 19). Dies beruht auf dem Ansatz, dass jedes manuell erstellte Material unweigerlich einige Merkmale des Verfassers widerspiegelt (Narayanan, et al., 2012, S. 300). Die Merkmale verschiedener Texte werden einem Vergleich unterzogen, um eine Aussage über den Autoren zu treffen.

Demnach handelt es sich bei dem für die Autorenerkennung relevanten Fachwissen um die Linguistik oder Sprachwissenschaft. Diese setzt sich mit der Sprache, ihrer Struktur und Anwendung auseinander. Da die Sprache biologisch, neurologisch und kulturell bedingt ist, besitzt der Untersuchungsgegenstand dieser Wissenschaft eine entsprechende Komplexität. Wird der Autor eines anonymen Textes bestimmt, geschieht dies meistens im Kontext des Rechtes oder einer Straftat und somit mittels forensischer Linguistik, einer Unterdisziplin der Linguistik und der Schnittstelle von Sprache, Gesetz und Verbrechen (Dern, 2009, S. 17 f.).

Eine standardisierte Methode im Bereich der Autorenerkennung existiert nicht (Dern, 2003). Vielmehr existiert eine Reihe von verschiedenen Ansätzen, welche auf die Untersuchung der Orthographie, Morphologie, Syntax oder der Lexik basieren und diese qualitativ bewerten (ebd.). Ein Beispiel für eine solche Untersuchung ist (Bredthauer, 2013, S. 73). Drommel (2016) bezeichnet diese Verfahren als qualitative Verfahren. Durchgeführt werden sie von Experten der modernen Linguistik. Eine äquivalente digitale Realisierung einer solchen Untersuchung in verschiedenen sprachlichen Ebenen besitzt eine solche Komplexität, dass sie bis heute in dieser Form nicht realisiert werden konnte. Stattdessen werden Textmerkmale definiert, welche versuchen, die Charakteristik eines Autors zu erfassen und dessen individuellen Stil zu ermitteln (Bönninghoff, et al., 2019, S. 695). Diese werden von Drommel (2016) als

quantitative Verfahren bezeichnet. Sie basieren auf der Zählung, Messung und Auswertung der Textmerkmale mittels statistischer Methoden.

Im Rahmen dieser Arbeit werden ausschließlich quantitative Verfahren behandelt, obgleich qualitative Verfahren Grundlage für die Begründung einiger Textmerkmale darstellen können.[2]

Im Folgenden werden die Kernbereiche der Autorschaftsanalyse und die Unterschiede zwischen profil- und textbasierten Verfahren aufgezeigt. Des Weiteren werden Herausforderungen der Autorschaftsanalyse erläutert und die daraus abgeleiteten Anforderungen an die Daten definiert.

3.1. Kernbereiche

Die Autorschaftsanalyse kann in drei Kernbereiche unterteilt werden: Autorschaftsattribution, Autorschaftsverifikation und Sprachprofiling (Halvani & Steinebach, 2014, S. 36). Diese werden im Folgenden näher erläutert.

Das Ziel der Autorschaftsattribution ist es, anhand von Merkmalen gegebener Texte mit bekannten Autoren einen anonymen Text einem Autor zuzuweisen (Stamatatos, 2009, S. 1). Dabei sollen die Merkmale den Schreibstil des Autors wiedergeben. Grundsätzlich wird hier zwischen zwei Problemen unterschieden, dem „Closed-Set-Problem“ und dem „Open-Set-Problem“. Bei einem Closed-Set-Problem ist der Autor des Dokumentes in der Trainingsmenge enthalten, bei einem Open-Set-Problem nicht. Demnach ist ein Closed-Set-Problem in der Regel einfacher zu lösen, da hier eine Prüfung gegen eine bekannte Anzahl von Autoren und nicht gegen theoretisch unendlich viele Autoren erfolgt (Halvani & Steinebach, 2014, S. 36).

Die Autorschaftsverifikation hingegen prüft zwei oder mehrere Dokumente auf denselben Autor. Das Ergebnis ist wahrscheinlichkeitsbasiert. Durch einen

[2] Demzufolge sind unter Begriffen wie Autorschaftsanalyse oder Autorenerkennung immer qualitative Verfahren zu verstehen.

Schwellenwert erfolgt eine Antwort auf die Frage mit „ja“, „nein“ oder „unbeantwortet“ (Bönninghoff, et al., 2019, S. 694).

Das Ziel des Sprachprofilings ist es, die demografischen Merkmale des Autors wie Alter, Geschlecht und Bildungshintergrund durch Analyse seines Textes vorherzusagen (Vijayan & Govilkar, 2019, S. 1065).

Dabei überschneiden und ergänzen sich die verwendeten Merkmale und Techniken dieser drei Kerngebiete oftmals. Beispielsweise kann eine vorgelagerte Profilierung eines Textes den Autorenkreis einschränken und wesentlich dazu beitragen, die Autorschaft herauszufinden oder zu prüfen.

Ein Ziel dieser Arbeit ist das Identifizieren des anonymen Autors eines Textes. Dies entspricht einem Verfahren der Autorschaftsattribution. Es entspricht einem gängigen Muster in der forensischen Linguistik, dass aufgrund externer Randbedingungen wie Fallkonstellationen oder Anknüpfungstatsachen davon auszugehen ist, dass der Autor unter einer eingrenzbaren Gruppe von Personen ist (Drommel, 2016, S. 190). Ein Beispiel ist eine Gruppe von Personen, welche Zugriff auf die in einem inkriminierten[3] Schreiben verwendeten Informationen haben. Dies entspricht einem „Closed-Set-Problem“. Im Rahmen dieser Arbeit wird mit dem Aufbau des Modells ebenfalls ein „Closed-Set-Problem“ behandelt.

3.2. Paradigmen

Die Modelle in der Autorschaftsanalyse werden weiterhin nach zwei verschiedenen Paradigmen unterschieden: Das instanzbasierte und das profilbasierte Paradigma (Potha & Stamatatos, 2014, S. 316). Diese unterscheiden sich darin, wie die Eingabetexte eines Autors behandelt werden.

Profilbasierte Methoden verketten alle bekannten Texte eines Autors miteinander. Das daraus entstehende Dokument wird verwendet, um die Eigenschaften dessen Stils zu erfassen. Intanzbasierte Ansätze hingegen

[3] Darunter zählen Texte welche selbst eine Straftat darstellen oder im Kontext einer Straftat relevant sind.

behandeln jeden Text mit bekanntem Autor einzeln. Hierbei wird jeder Text als ein Vektor seiner Attribute dargestellt und dient als Eingabe für einen Klassifikationsalgorithmus. (Stamatatos, 2009, S. 13)

Instanzbasierte Ansätze sind effektiver, wenn mehrere Texte für die Autoren existieren (Potha & Stamatatos, 2014, S. 316). Profilbasierte Ansätze werden verwendet, wenn nur eine limitierte Anzahl an Dokumenten zur Verfügung stehen. Des Weiteren merkt Stamatatos (2009, S. 13) an, dass durch das Zusammenfassen der Texte eines profilbasierten Ansatzes Unterschiede zwischen den einzelnen Texten eines Autors vernachlässigt werden und extrahierte stilometrische Merkmale im Vergleich zu den ursprünglichen Texten abweichen. Ein Ziel dieser Arbeit ist es, diese Merkmale aufzuzeigen und zu bewerten, weshalb für das Modell ein instanzbasiertes Paradigma gewählt wird.

3.3. Herausforderungen und Anforderungen an die Datengrundlage

Die Autorschaftsanalyse birgt zahlreiche Herausforderungen, welche zur Komplexität und dem hohen Forschungspotential dieser Disziplin beitragen. In diesem Kapitel werden die wichtigsten Herausforderungen und die daraus folgenden Anforderungen an den Datensatz aufgezeigt.

Insbesondere bei inkriminierten Texten kommt es häufig zu einer bewussten Verfälschung des Schreibstils, um die eigene Identität zu verschleiern oder den Verdacht von sich zu lenken (Bredthauer, 2013, S. 1). Das Erkennen solcher Täuschungen ist ein eigenes Forschungsgebiet und wird beispielsweise in Arbeiten von Bredthauer (2013) und Halvani, Steinebach, & Neitzel (2014) behandelt.

Zur Extraktion einiger Merkmale werden Werkzeuge des NLP (Natural Language Processing) verwendet. Die Grundlage dieser Werkzeuge ist immer ein Textkorpus[4] einer Sprache, auf welchen sie trainiert werden. Handelt es sich um

[4] Carstensen, et al (2010) definieren einen Textkorpus als „Sammlung gesprochener oder schriftlicher Äußerungen, die digital erfasst, also auf Rechnern gespeichert und maschinenlesbar sind, und für eine linguistische oder computerlinguistische Aufgabe aufbereitet wurden“.

unterschiedliche Sprachen innerhalb eines Textes, können diese Werkzeuge fehlerhafte Ergebnisse liefern. Zusätzlich besitzt jeder Autor aufgrund sprachlicher Eigenheiten einen anderen Stil in der jeweiligen Sprache. Beispielsweise tendieren deutsche Wörter dazu, deutlich länger zu sein als Wörter in der englischen Sprache (vgl. Kapitel 5.6.4). Aus diesem Grund muss bei der Autorschaftsanalyse immer die jeweilige Sprache beachtet werden.

Eine weitere Abhängigkeit des Schreibstils besteht durch das Genre oder den Inhalt des Textes. In einem Bewerbungsschreiben wird ein anderes sprachliches Register verwendet als in einer WhatsApp Nachricht. Ein Artikel über Politik enthält ein anderes themenspezifisches Vokabular als ein Artikel über Sport. Aus diesem Grund besteht ein Konflikt zwischen der Abhängigkeit einiger Merkmale vom Autor und dem Inhalt oder Genre (Bönninghoff, et al., 2019, S. 695).

Bei Texten mit einer geringen Länge kann es sein, dass die Textmenge nicht ausreicht, um den Stil eines Autors wiederzugeben. Dies ist darauf zurückzuführen, dass nicht alle Merkmale in den kurzen Texten auftreten und somit nicht verglichen werden. Es ist dabei zum Beispiel unwahrscheinlicher, dass sich bestimmte Worte ausreichend oft wiederholen, die ein Autor bevorzugt verwendet (Forman, 2002, S. 2). In der Literatur werden verschiedene Längen vorgeschlagen. Sanderson & Guenter (2006) verwendeten Blöcke mit einer Länge von 500 Zeichen. Feiguina & Hirst (2007) führten Experimente mit 200, 500 und 1.000 Wörtern durch und vermerkten eine abnehmende Genauigkeit des Modells mit abnehmenden Längen (Stamatatos, 2009, S. 17).

Eine weitere Herausforderung sind gemeinschaftlich erstellte Dokumente, beispielsweise von Redaktionen. Bei diesen kann der Text nicht eindeutig einem Autor zugewiesen werden. Hierdurch kann Textmerkmal nicht direkt einem Autor zugewiesen werden.

Zuletzt gilt es zu erwähnen, dass sich der Schreibstil einer Person über die Zeit ändern kann und einige Merkmale in Abhängigkeit zum Alter des Autors stehen. So gibt es in der Profilierung einige Merkmale, die verwendet werden, um das Alter des Autors zu bestimmen (Nini, 2015, S. 173). Liegen geschriebene

Dokumente desselben Autors zeitlich auseinander, können dadurch verschiedene Schreibstile in diesen existieren.

Das Prüfen der Merkmale ist nicht Teil dieser Arbeit. Aus diesem Grund bestehen folgende Anforderungen an den vom Modell verwendeten Datensatz:

1. Die Autoren dürfen nicht bewusst versuchen ihren Schreibstil zu verfälschen
2. Die Texte müssen hauptsächlich in deutscher Sprache geschrieben sein. Einzelne kleinere Passagen in englischer Sprache, wie englische Namen oder Zitate, können nicht ausgeschlossen werden.
3. Die Texte müssen grundsätzlich dasselbe Genre und Thema besitzen.
4. Die Texte müssen über eine ausreichende Länge verfügen, welche es ermöglicht, den Stil des Autors zu erfassen.
5. Ein Text darf nur von einem Autor verfasst sein.
6. Die Texte dürfen zeitlich nicht zu weit voneinander entfernt erstellt worden sein.
7. Es muss sichergestellt sein, dass eine ausreichende und etwa gleiche Anzahl an Texten für jeden Autor zur Verfügung steht.

4. Natural Language Processing

Raimund H. Drommel, einer der führenden deutschen Sprachprofiler, beschreibt die Computerlinguistik als „Untersuchung natürlicher Sprachen mithilfe von Computeralgorithmen“ (Drommel, 2016). Darunter fällt das Beschreiben von Phänomenen, Strukturen und Prozessen natürlicher Sprache, um diese zu analysieren oder nachbilden zu können, aber auch computerbasierte Modellvorstellungen für linguistische Untersuchungen. Unter natürlicher Sprache werden Sprachen verstanden, welche nicht geplant wurden, sondern sich über lange Zeit durch ihre Verwendung entwickelt haben (Kipp, 2018). Beispiele hierfür sind Deutsch, Englisch oder Spanisch. Natural Language Processing bezeichnet die Entwicklung von Algorithmen, die Computer dazu befähigen, natürliche Sprache zu verstehen (Henschel, 2019). Eine der größten Herausforderungen stellt hierbei die Vielzahl an Variationen der natürlichen Sprache, abhängig vom spezifischen Kontext, dar. Der Mensch lernt diese schon in frühen Jahren und erweitert sein Wissen sein ganzes Leben lang, um den Kontext und damit die sprechende Person zu verstehen. Im Bereich des NLP wurden für diese Herausforderung verschiedene auf maschinellem Lernen basierende Modelle entwickelt (Prateek, 2017, S. 249).

4.1. Python Bibliotheken

Nach (Liu Y. H., 2019, S. 47) sind die populärsten Python Bibliotheken für NLP das NLTK (Natural Language Toolkit), SpaCy, Genism und TextBlob. Das NLTK ist die führende Plattform im Bereich des NLP. Es vereinigt Bibliotheken und Programme der Programmiersprache Python für Anwendungen der Computerlinguistik und ermöglicht den Zugriff auf über 50 Korpora und lexikalische Ressourcen (NLTK.org, 2020b). Entwickelt wurde es ursprünglich für die Forschung, findet aber mittlerweile auch große Anwendung in der Industrie (Stadler, 2020). NLTK unterstützt von allen Bibliotheken die meisten natürlichen Sprachen (NLTK.org, 2020a). Auch wenn die Tokenization zu Sätzen

vergleichsweise schnell ist, sind die anderen Algorithmen des NLTK eher langsam (Explosion, 2020b).

SpaCy hingegen ist auf eine effiziente Nutzung von CPU und Arbeitsspeicher optimiert und verwendet neuronale Netze zum Trainieren einiger Modelle. Wegen dieser Effizienz ist es vor allem in der Industrie verbreitet. Geschrieben ist die Bibliothek in Python und Cyton (Explosion, 2020c). Im Vergleich zu NLTK besitzt SpaCy weniger Flexibilität und dafür stark auf Präzision und Geschwindigkeit optimierte Verfahren. Es werden ebenfalls weniger Sprachen als bei NLTK unterstützt (Explosion, 2020b).

Genism und TextBlob besitzen einige Zusatzfunktionen, die ergänzend zu den vorherigen Bibliotheken eingesetzt werden können. Beispielsweise besitzt TextBlob eine Rechtschreibeüberprüfung und Spracherkennung. Die beiden Bibliotheken sind jedoch nicht für die deutsche Sprache verfügbar und werden deshalb in dieser Arbeit nicht weiter betrachtet.

4.2. Tokenization

Der englische Begriff Tokenization bezeichnet das Aufteilen eines Textes in definierte Teile, wie Wörter, Sätze oder Buchstaben. Diese Teile werden als Token bezeichnet. Ein N-Gramm besteht aus N aufeinanderfolgenden Teilen, wobei N der Anzahl der Teile entspricht. Während des Vorganges können Zeichen wie Punkte oder Smileys entfernt werden. Die Komplexität ist dadurch deutlich höher als beim Trennen nach Leer- und Satzzeichen (Liu Y. H., 2019, S. 49). Dies wird durch den folgenden Beispielsatz verdeutlicht:

„Das englische Akronym S.W.A.T. steht für Special Weapons And Tactics.“

Hier muss S.W.A.T. als eigenes Wort und nicht als vier durch Punkte getrennte Buchstaben erkannt werden.

Weiterhin kann einem Token je nach seiner Position, Funktion oder Kontext eine Wortart zugeordnet werden. Die Wortart kann sich dabei nach der grammatikalischen oder lexikalischen Funktion des Tokens im Satz richten. Beispielsweise werden Adjektive, Verben und Konjunktionen erkannt, aber auch

Interpunktionszeichen oder Eigennamen. Dieser Prozess nennt sich Part-of-speech-tagging (kurz: PoS-Tagging) (Liu Y. H., 2019, S. 53). Im Rahmen dieser Arbeit wird die Bibliothek SpaCy für das PoS-Tagging verwendet. Die möglichen Wortarten der Bibliothek befinden sich in der offiziellen Dokumentation (Explosion, 2020a).

4.3. Textnormalisierung

Natürliche Sprache neigt als menschliches Erzeugnis dazu, einer gewissen Zufälligkeit zu folgen. Das Normalisieren des Textes hat das Ziel, diesen Zufall und dadurch die vom Computer zu verarbeitenden Dimensionen an Information zu minimieren (Duque, 2020).

Die reduzierte Information steht im direkten Konflikt mit der Extraktion der Merkmale eines Autors. Der produzierte „Zufall“ eines Textes beschreibt die Information, welche verwendet werden soll, um den Autor zu identifizieren. Dennoch gibt es einige Merkmale, welche auf die Normalisierung des Textes angewiesen sind. Beispielsweise darf beim Zählen der Häufigkeit einzelner Worte nicht jede morphologische[5] oder orthographische[6] Variante eines Wortes als eigenes Wort behandelt werden. Andererseits dürfen bei Suffixen, welche die Endsilben von Wörtern extrahieren, diese nicht durch Stammformreduktion oder Lemmatisierung entfernt worden sein. Aus diesem Grund ist die Verwendung und der Normalisierung abhängig vom jeweiligen Merkmal.

Es existiert kein Standardverfahren, welche Schritte für eine Normalisierung eines Textes durchzuführen sind. Diese sind immer abhängig vom gegebenen Text und dem Verwendungszweck. In dieser Arbeit kann die Normalisierung abhängig vom jeweiligen Merkmal folgende Schritte enthalten:

- Umwandeln aller Majuskel in Minuskel

[5] Morphologie bezeichnet in der Sprachwissenschaft die Formenlehre, denen die Wörter durch Deklination und Konjugation unterliegen (DWDS, 2020a).
[6] Orthographie oder Rechtschreibung beschreibt die allgemein übliche Schreibweise von Wörtern (DWDS, 2020b).

- Stammformreduktion oder Lemmatisierung
- Eliminieren von Stoppwörtern

4.3.1. Stammformreduktion und Lemmatisierung

Die Stammformreduktion (engl. Stemming) und die Lemmatisierung (engl. Lemmatization) haben das Ziel, morphologische oder orthographische Varianten eines Wortes auf eine kanonische Grundform zurückzuführen. Sie unterscheiden sich dabei in der entstehenden kanonischen Form. Bei der Stammformreduktion handelt es sich dabei um einen Wortstamm, welcher durch das regelbasierte Abschneiden von Suffixen entsteht. Das resultierende Wort muss dabei keinem Wörterbuch entstammen. In der Lemmatisierung ist die kanonische Form ein Wort, welches in einem Wörterbuch nachgeschlagen wird. Ebenfalls wird die Wortart eines Wortes berücksichtigt. Im NLP wird dies benötigt, um bedeutungsgleiche Wörter in verschiedenen Formen zu finden (Prateek, 2017, S. 252 f.).

4.3.2. Inhalts- und Funktionswörter

Inhaltswörter zeichnen sich durch ihre lexikalische Bedeutung aus. Sie alleine reichen aus, um das Thema eines Textes vollständig zu erfassen (DWDS, 2020). Funktionswörter hingegen besitzen als Antonym der Inhaltswörter keine lexikalische Bedeutung. Ihre Aufgabe ist die Verbindung der Inhaltswörter in einem Satz (Kaster, Siersdorfer, & Weikum, 2005, S. 2). Sie beschreiben demnach die Beziehung zwischen Inhaltswörtern.

Basierend auf dieser Unterscheidung sind die Wortarten des PoS-Taggings der Bibliothek SpaCy (Explosion, 2020a) folgendermaßen zuzuordnen:

- Inhaltswörter: Substantive, Verben, Adverbien, Adjektive, Numerale, und Eigennamen
- Funktionswörter: Adpositionen, Hilfsverben, Konjunktionen, Artikel, Pronomen und Partikel

4.3.3. Stoppwörter

Bei Stoppwörtern handelt es sich um Wörter, welche sehr häufig in allen Texten vorkommen, jedoch keine Informationen beinhalten, um zwischen verschiedenen Klassen von Texten zu unterscheiden. Demnach existiert für jeden Korpus und jede Textsammlung eine eigene Liste von Stoppwörtern.

SpaCy besitzt eine Liste der allgemeinen deutschen Stoppwörter. Diese umfasst 543 Wörter, beispielsweise „ganzer“, „dazwischen“, „gerade“, „den“ und „gehabt“. Das Entfernen von Stoppwörtern kann bei der Arbeit mit Funktionen wie einem Wortzähler hilfreich sein, welche häufig vorkommende Wörter nicht gewichten (Raschka, 2016, S. 243).

5. Datenverständnis

Als Datengrundlage des maschinellen Lernens dienen Merkmale. Unter einem Merkmal ist eine individuell messbare, meist numerische Eigenschaft oder Charakteristik eines beobachteten Phänomens zu verstehen. Vergleichbar sind diese mit erklärenden Variablen, die in statistischen Techniken wie der linearen Regression verwendet werden (Bishop, 2006, S. 262). Im Rahmen dieser Arbeit handelt es sich dabei um Merkmale eines Textes, welche potentiell eine Aussage über einen unbekannten Autor treffen können.

Die Grundlage aussagekräftiger Merkmale ist das spezifische Fachwissen (Liu Y. H., 2019, S. 29). Dieses wird zum einen benötigt, um neue Merkmale zu erstellen und zum anderen, um für das spezifische Problem passende Merkmale aus der Literatur auszuwählen. Die spätere Leistung des Modells ist stark abhängig von der Qualität dieser Merkmale (Coelho & Richert, 2015, S. 42).

Bei dem für die Autorerkennung relevanten Fachwissen handelt es sich um die Linguistik oder Sprachwissenschaft (vgl. Kapitel 3). Im Rahmen des Datenverständnisses wird deshalb zuerst die ausgewählte Datengrundlage erläutert und die zu untersuchenden Merkmale aufgezeigt. Dabei wird die Wahl der Merkmale durch bereits existierende Arbeiten und Erkenntnisse in der relevanten Forschung begründet.

5.1. Datengrundlage

Die Datengrundlage sind Artikel der Seite „Musikreviews.de“. Musikreviews.de ist ein Magazin für Album-Kritiken, Nachrichten und Interviews aus den Bereichen Metal, Progressive und Rock. Laut Thoralf Koß, dem Chefredakteur des Magazins, unterliegt die Veröffentlichung der Artikel alleinig dem entsprechenden Redakteur. Es werden keinerlei Kürzungen oder andere Änderungen vorgenommen. Dies ist eine notwendige Grundlage für die Autorschaftsanalyse, da demnach ein Text alleinig einem Autor zugewiesen werden kann. Die einzige Ausnahme besteht im Falle von Flüchtigkeitsfehlern,

wie beispielsweise bei einem Buchstabendreher (vgl. Anhang A1). Auch der Versuch der Autoren, ihre Identität bewusst zu verschleiern kann ausgeschlossen werden, da es sich bei den Artikeln um keine inkriminierten Texte handelt und somit keine entsprechende Motivation zu erwarten ist. Die Texte werden ausschließlich in deutscher Sprache verfasst, Ausnahmen sind Zitate, Bandnamen sowie Songnamen.

Das Erstellen der Textsammlung erfolgt mithilfe eines Webscrapers, welcher auf der der Bibliothek „BeautifulSoup“ basiert. Der zugehörige Code befindet sich im Anhang A2. Neben dem Text des Artikels und dem zugehörigen Autor werden das Genre, das Erscheinungsdatum, der Titel und die URL in einer CSV-Datei gespeichert.

Die Datei umfasst 19.557 Artikel von 52 verschiedenen Autoren, zuzüglich der verschiedenen Gastautoren, welche unter dem Namen „Gast-Rezensent“ veröffentlicht werden. Dabei schwankt die Anzahl der veröffentlichten Artikel pro Autor stark. Die Spanne der Veröffentlichungen der Autoren bewegt sich dabei zwischen 3 und 7.428 Artikeln. Dieses starke Ungleichgewicht besitzt deutliche Auswirkungen auf die Güte der Analysen und Modelle. Aus der Gesamtheit der Artikel werden deshalb nur die ersten 100 Artikel von Autoren ausgewählt, die mehr als 100 Artikel verfasst haben. Bei 25 Autoren mit mindestens 100 Texten entspricht dies 2.500 Artikeln. In der Praxis lassen sich eher selten 100 Beispieltexte für einen Autor finden. Somit ist die Anforderung einer ausreichenden Anzahl an Texten für jeden Autor ebenfalls gegeben. Die durchschnittliche Länge der Artikel beträgt 378,95 Wörter. Zwar stellten (Feiguina & Hirst (2007) bei ihren Tests mit je 1.000, 500 und 200 Wörtern langen Texten eine abnehmende Genauigkeit fest, konnten jedoch noch gute Ergebnisse mit 500 Wörtern erzielen. Aus diesem Grund wird in dieser Arbeit angenommen, dass die durchschnittliche Länge von 378,95 Wörtern ausreicht, um eine Bewertung der Merkmale vorzunehmen.

Potentielle Fehler können durch die zeitliche Distanz der Artikel und die verschiedenen Genres entstehen. Der Zeitraum, in welchem die Artikel geschrieben wurden, differiert für die einzelnen Autoren stark. Beispielsweise

schrieb Autor 2[7] alle Artikel in einem Monat und 16 Tagen. Autor 6 hingegen in 6 Jahren und 18 Tagen. Im Durchschnitt liegt die zeitliche Distanz des ersten und letzten Artikels bei 2 Jahren, 7 Monaten und 25 Tagen. In diesem Zeitraum kann sich der Schreibstil eines Autors durchaus ändern, weshalb dieser Aspekt als mögliche Fehlerquelle angesehen wird.

Einige Merkmale sind durch ihre Abhängigkeit zum Inhalt beziehungsweise Thema des Textes in der Autorschaftsattribution sehr umstritten. Eine genauere Erläuterung der Merkmale erfolgt in Kapitel 5.5. Ob innerhalb der verwendeten Datengrundlage eine thematische Abhängigkeit der Merkmale zum Genre der rezensierten Musik existiert, lässt sich nicht bestimmen. Die Herausforderung hierbei ist, dass keine einheitlichen Genres existieren. Daraus resultiert mit 8.579 Genres für die ursprünglichen 19.557 Artikel eine sehr hohe Anzahl. Dabei existiert keine genaue Definition über die einzelnen Kategorien wie zum Beispiel „Black Metal“ und „Death Metal“. Vielmehr besteht die Annahme, dass die Einordnung in Genres aufgrund einer subjektiven Meinung der Autoren und nicht nach einem festen Schema erfolgt. Eine Untersuchung dessen ist nicht Teil dieser Arbeit, es werden lediglich zwei weitere Fehlerquellen in der Evaluierung beachtet:

1. Die Vorhersage themenabhängiger Merkmale zeigt sehr schlechte Ergebnisse, da sich der Inhalt der einzelnen Artikel zu sehr unterscheidet.
2. Bestimmte Autoren schreiben nur über bestimmte Genres. Dies hätte zur Folge, dass themenabhängiger Merkmale den Autoren aufgrund des Genres und nicht des Schreibstils identifizieren.

5.2. Merkmalskategorien

Die Gruppierung der Merkmale in Kategorien richtet sich dabei stark nach Stamatatos (2009). Sie unterscheidet sich dahingehend, dass zum einen eine Gruppierung in anwendungsspezifischen Merkmale nicht stattfindet und zum

[7] Um die informationelle Selbstbestimmung der einzelnen Personen zu wahren werden die Namen der Autoren pseudonymisiert. Weitere Informationen in Kapitel 6.9.

anderen eine weitere Gruppierung der lexikalischen Merkmale in zeichenbasierte Merkmale anlehnend an Abbasi & Chen (2008) erfolgt. Des Weiteren wird davon abgesehen komplexe semantische Beziehungen aufgrund der bisher mangelnden technischen Möglichkeiten weiter zu betrachten (Stamatatos, 2009). Stattdessen werden einfachere semantische Beziehungen wie beispielsweise Funktionswörter den anderen Kategorien untergeordnet. Anwendungsspezifische Merkmale beziehen sich auf die Textorganisation und das Layout. Dies ist ausschließlich bei verschiedenen Online-Texten relevant, bei welchen beispielsweise die verwendete Schriftfarbe oder Paragraphenlänge untersucht werden. Diese Kategorie wird im weiteren Verlauf nicht weiter betrachtet, da die Datengrundlage nur unformatierte Textdaten enthält.

In der Literatur wird der Konflikt zwischen der Abhängigkeit zum Autor und dem Kontext bzw. Inhalt des Textes einiger Merkmale bemängelt. (Abbasi & Chen, 2008 (2008) führen deshalb eine eigene Kategorie für themenabhängige Merkmale ein. Durch die Zugehörigkeit der Merkmale zu anderen Gruppen wird in dieser Arbeit stattdessen ein Indikator hinzugefügt, welcher beschreibt, ob eine vermutete Abhängigkeit eines Merkmals zum Inhalt des Textes besteht.

Die verwendeten Kategorien stellen sich wie folgt dar:

- **Lexikalisch**
  Befasst sich mit der Untersuchung der isolierten Wörter ohne die Berücksichtigung des Textzusammenhangs, wie beispielsweise die Anzahl der unterschiedlichen Wörter oder die durchschnittliche Wortlänge.
- **Symbolisch**
  Der Text wird als eine Folge von Zeichen betrachtet. Dazu gehören zum Beispiel die verwendeten Zahlen oder Buchstabenfolgen in Form von N-Grammen.
- **Syntaktisch**
  Untersucht die Verknüpfung sprachlich verwandter Einheiten in einem Satz, wie Wortgruppen oder übliche Verbindungen von Wörtern. Beispiele hierfür sind Wortarten-N-Gramme.

5.3. Bekannte Merkmale in der Forschung

Als Grundlage für die verwendeten Merkmale in dieser Arbeit werden fünf verschiedene Arbeiten mit guten Ergebnissen, vorrangig in der englischen Sprache, verwendet. Von Stamatatos (2009) und Halvani, Steinebach, & Neitzel (2014) wird ein Überblick über die jüngsten Fortschritte der automatisierten Ansätze zur Zuschreibung der Autorschaft gegeben, wobei deren Merkmale sowohl für die Textrepräsentation als auch für die Textklassifikation untersucht werden. Halvani, Steinebach, & Neitzel (2014) fokussieren sich hierbei vor allem auf mögliche Maßnahmen gegen eine Erkennung. Potha & Stamatatos (2014) beschreiben ein sehr effektives profilbasiertes Modell. Profilbasierte Modelle behandeln alle Textproben eines Autors kumulativ. Das von Abbasi & Chen (2008) vorgestellte Modell erreicht bei einer Unterscheidung von 100 Autoren eine Genauigkeit von bis zu 94%. Narayanan, et al. (2012) beschäftigen sich hingegen mit einem Modell zur Autorschaftserkennung bei bis zu 100.000 möglichen Autoren. Hier ist der richtige Autor in etwa 35% der Fälle unter den 20 besten Treffern. Diese Arbeiten verwenden für die Ermittlung des Autors trotz der verschiedenen Ansätze eine Reihe von teilweise übereinstimmenden Merkmalen.

Neben ihren guten Ergebnissen wurden die Arbeiten aufgrund der verschiedenen Ansätze ausgewählt. Dies hat zur Folge, dass einige Schnittmengen und Differenzen in den verwendeten Merkmalen und Vorgehensweisen besteht. Insbesondere existieren Merkmale mit unterschiedlichen Namen, welche sich auf dieselbe Information beziehen und dadurch Redundanzen erzeugen.

Beispielsweise sprechen Narayanan, et al. (2012, S. 305) von Frequenzen bei den verwendeten Merkmalen. Unter diesen Frequenzen ist hier kein Wiederholungsmaß zu verstehen, sondern die relative Häufigkeit der Vorkommnisse. Die Autoren schreiben hier von denselben Merkmalen, wodurch folgende Zusammenfassung erfolgt:

- Bag-of-Words, Wort-Frequenzen → Bag-of-Words
- Wortlänge, Wortfrequenz 1-20 Buchstaben → Wortlänge

- Anzahl Sonderzeichen, Sonderzeichenfrequenz → Sonderzeichen
- Zeichentypen, Zeichenfrequenz → Sonderzeichen
- Ziffern, Zahlenfrequenz → Ziffern
- Funktionswörter, Funktionswörterfrequenz → Funktionswörter

Tabelle 1 gibt einen Überblick, über die Merkmale aus der beschriebenen Literatur. Die Gruppierung erfolgt dabei nach den in Kapitel 5.2 5.2genannten Kategorien. Die Farbe Grün entspricht einem befürworten der Merkmale, Rot einer klaren Ablehnung des Merkmals und Weiß dafür, dass ein Merkmal in der Arbeit nicht erwähnt wurde.

- 1: Halvani, Steinebach, & Neitzel (2014)
- 2: Abbasi & Chen (2008)
- 3: Narayanan, et al. (2012)
- 4: Potha & Stamatatos (2014)
- 5: Stamatatos (2009)
- T: Es ist von einer deutlichen Themenabhängigkeit des Merkmals auszugehen

Tabelle 1: Bekannte Merkmale der Literatur – Eigene Darstellung

Merkmal	Beschreibung	Literatur					T
		1	2	3	4	5	
Lexikalisch							
Bag-of-Words	Anzahl der vorkommenden Wörter Beispiel (“Maus”, “Haus”)						x
Wort-N-Gramme	Beispiel: 2-Gramm "großes Haus", 3-Gramm "mein großes Haus"						x
Anzahl der Wörter	Anzahl der Wörter						x
% Buchstaben. pro Wort	Anteil der Gesamtzeichen pro Wort						
Wortlänge	Frequenz von Wörtern mit 1–20 Buchstaben						
Wortschatzreichtum	Wortschatz, z.B., hapax legomena, dis legonema, Yule’s K						
Präfixe	[Text]beispiel (Vorsilbe)						
Infixe	Text[bei]spiel (innerer Wortbestandteil)						
Suffixe	Textbei[spiel] (Nachsilbe)						
Anglizismen	Wortentlehnungen (z. B. Mail, Newsletter, Chat)						
Neologismen	Kunstwörter (z. B. Abmahnwelle, Nerd, googeln)						
Kollokationen	Häufig vorkommende Wortverbindungen: „starker Tobak“						
Wort-Komplexität	Wörter mit x Vokalen						

Wortgestaltung	Frequenz von Wörtern mit verschiedenen Kombinationen von Groß- und Kleinbuchstaben.						
Synonyme	Synonyme						
Rechtschreibfehler	Rechtschreibfehler						
Symbolisch							
Anzahl der Buchstaben	Anzahl der Buchstaben(A-Z, Ä,Ö,Ü, a-z, ä, ö, ü, ß)						x
Sonderzeichen	Anzahl Sonderzeichen (z.B., @#$%ˆ)						
Zeichentypen	Buchstaben, Zahlen, etc.						
Zeichen-N-Gramme	Beispiel: 2-Gramm "ab", 3-Gramm "abc"						~*
Ziffern	Ziffern (0-9)						
Syntaktisch							
Funktionswörter	Artikel (der, das, einer, …), Konjunktionen (und, oder, …)						
Wortarten	Frequenz der Wortarten (e.g., Nomen, Adjektiv)						
Wortarten-N-Gramme	Beispiel: 2-Gramm "Nomen, Adjektiv"						
Phrasen/Redewendungen	Redensarten wie z. B. „aus dem Nähkästchen plaudern“, …						
Satzlänge	Länge der Sätze						
Satz-Komplexität	Vorfeld/Mittelfeld/Nachfeld-Komplexitäten, …						
Syntaktische Paare	Frequenz deines jeden paares (A, B), bei welchem B das Kind von B im parse tree ist						
Satz- und Phrasenstruktur	Satz- und Phrasenstruktur						
Satz-Anfänge/-Endungen	Satzanfang(Nomen), Satzende(finites Verb), …						
Semantische Beziehungen	Beschreibung der Beziehungen, die Bedeutungen zueinander einnehmen können.						
Grammatikalische Fehler	Falsche Verwendung von Genus, Kasus, Kommata, …						
Neuschreiberegeln	Messungen bezüglich der Nomenphasen, Verbphrasen, längen der Verben, etc.						
Verständlichkeits-Indizes	Gunning Fog Index, Flesch-Kincaid Reading Ease, …						

*Abhängig von der verwendeten Superklasse der Zeichen-N-Gramme

5.4. Ausgeschlossene Merkmale

Einige der in Tabelle 1 aufgezeigten Merkmale können bereits vor einer genauen Evaluation ausgeschlossen werden. Dies kann an aufgrund einer Redundanz der Informationen oder an einem hohen Fehlerpotential der Merkmale erfolgen.

Der prozentuale Anteil der „Gesamtbuchstaben pro Wort“[8] wird als redundant angesehen, da bereits die Merkmale „Wortlänge“ und „Anzahl der Wörter“ verwendet werden. Das Merkmal kann anhand dieser berechnet werden und wird

[8] Für eine bessere Lesbarkeit werden alle Merkmale in diesem Kapitel in Anführungszeichen gesetzt.

aus diesem Grund nicht als Mehrwehrt angesehen. Auch die „Anzahl der Buchstaben“ besitzt demnach keine weiteren Informationen. „Prä-, In- und Suffixe“ werden im Rahmen der „Zeichen-N-Gramme“ extrahiert. Auch das Merkmal „Zeichentypen“ erfasst keine Information, die nicht bereits in den Merkmalen „Sonderzeichen“ oder „Ziffern“ enthalten wären. Das Ziel der Merkmale „Satz- und Phasenstruktur“, „Neuschreiberegeln“ und „syntaktischen Paare“ ist, eine individuelle syntaktische Struktur eines Autors zu erkennen. Dies wird durch den weniger komplexen Ansatz der „Wortarten-N-Gramme“ behandelt. Die „Wortkomplexität“ soll anhand der Vokale in einem Wort gemessen werden, was in Korrelation zur „Länge des Wortes“ steht. Die „Wortgestaltung“ ist abhängig von der Art des Textes. In der Datengrundlage dieser Arbeit kommen nur Worte mit einem führenden Großbuchstaben oder aus ausschließlich Kleinbuchstaben bestehende Worte vor, was eine sehr hohe Gemeinsamkeit mit der „Wortartbestimmung“ aufweist. Das Verwenden von „Synonymen“ resultiert in einem höheren Wert für den „Wortschatzreichtum“, weshalb dieses Merkmal ebenfalls nicht gesondert betrachtet wird. „Kollokationen“ beschreiben häufige Wortverbindungen, diese werden durch das Bestimmen von „Wort-N-Grammen“ aufgezeigt. „Satzkomplexitäten“ werden in verschiedenen Merkmalen behandelt. Die Struktur wird durch „Wortarten-N-Gramme“ bestimmt und die „Komplexität des Textes“ durch „Verständlichkeitsindizes“.

Eine zuverlässige Rechtschreibprüfung ist sehr komplex. Dies liegt zum einen daran, dass die deutsche Sprache Abweichungen besitzen kann (beispielsweise durch Komposita) und zum anderen, dass oftmals Fremdsprachen (beispielsweise Neologismen oder Eigennamen wie Bandnamen) in den Texten aufzufinden sind. Bibliotheken wie Pyspellchecker (Barrus, 2020) wiesen in den ersten Prüfungen sehr viele falsche positive Ergebnisse auf, was zu einer Verfälschung der Ergebnisse führen kann. Ein weiteres Argument gegen „Rechtschreibfehler“ als Merkmal ist, dass die meisten elektronisch geschriebenen Texte ohnehin einer Rechtschreibprüfung unterzogen werden. In Folge dessen wird das Merkmal bestimmt, ob die verwendete Rechtschreibprüfung des Schreibenden oder der verwendeten Bibliothek in der

Analyse bessere Ergebnisse liefert. Eine Überprüfung der Grammatik besitzt ähnliche Herausforderungen, wobei diese vorrangig durch ihre Komplexität bestimmt wird. Ein vielversprechenderer Ansatz für ein Merkmal ist es, statt Grammatikfehlern die geschriebene Grammatik in Form des Satzbaus zu analysieren. Dies geschieht in Form von „Zeichen-“ und „Wortarten-N-Grammen“. Hohe Fehlerraten sind auch beim Bestimmen von „Anglizismen“ zu erwarten. Nicht jedes englische Wort entspricht einem Anglizismus oder wird von einer Spracherkennung korrekt erkannt. Auch „Neologismen“ lassen sich nur schwer bestimmen, und es ist noch kein Algorithmus bekannt, welcher diese zuverlässig erkennen kann.

Die Erkennung von „Phrasen und Redewendungen“ ist nur durch einen Abgleich mit einer Liste möglich. Diese kann zum einen nicht alle Redewendungen inklusive aller Abwandlungen dieser erfassen und würde in einer großen Menge von Nullwerten resultieren. Erkannt werden könnten sie in Form von langen „Wort-N-Grammen“. Hier werden häufige Kombinationen mehrerer Wörter gezählt.

Der Aufwand für die Ermittlung semantischer Beziehungen innerhalb eines Satzes oder Textes stellt einen hohen Aufwand dar und wird im Rahmen dieser Arbeit nicht betrachtet. Das Merkmal könnte aber ein informatives Merkmal darstellen. Eine gesonderte Untersuchung kann hier wertvolle Ergebnisse liefern.

5.5. Themenabhängige Merkmale

Einige Merkmale sind wegen ihrer Abhängigkeit vom Inhalt beziehungsweise Thema des Textes in der Autorschaftsattribution sehr umstritten. Der Unterscheidungswert dieser Merkmale steht in direktem Zusammenhang mit dem Grad der inhaltlichen Divergenz zwischen den Autoren und den Testsätzen. Das bedeutet, dass ein Ergebnis verfälscht werden kann, wenn ein Merkmal zu stark durch den Inhalt des Textes beeinträchtigt wird und dadurch der Schreibstil eines Autors nicht erkannt werden kann. Ein gutes Beispiel ist BoW (Bag-of-Words). Hier stehen genretypische Wörter im Konflikt mit solchen, die einem Autor zuzuordnen sind. Die inhaltliche Abhängigkeit konnte bereits in einigen

Studien, nachgewiesen werden (Kaster, Siersdorfer, & Weikum, 2005). Narayanan, et al. (2012) sehen deshalb von der Verwendung des Merkmals ab. Im Gegensatz dazu existieren auch Arbeiten mit sehr guten Ergebnissen unter der Verwendung des Modells wie Abbasi & Chen (2008). Auch Wort-N-Gramme unterliegen dieser Kritik, da sich dieses Merkmal sehr ähnlich zu BoW verhält. Der einzige Unterschied liegt darin, dass hier Wort-Gruppen und keine einzelnen Wörter gezählt werden.

Auch die „Textlänge“ und damit die „Anzahl der verwendeten Buchstaben“ zeigt eine Abweichung je nach Thema des Textes. So konnte David Kriesel in seinem Vortrag auf dem 33c3-Kongress nachweisen, dass Artikel des Ressorts Politik, Sport und Panorama kürzer ausfallen als Artikel, welche dem Ressort Kultur zuzuordnen sind. Als Grundlage dienten dabei über 100.000 Artikel (Kriesel, 2016).

Einige Zeichen-N-Gramme besitzen ebenfalls eine starke Abhängigkeit zum Inhalt des Textes. Die einzelnen Klassen und deren Abhängigkeit werden im späteren Verlauf der Arbeit näher erläutert.

5.6. Ausgewählte Merkmale

In vorherigen Kapiteln wurde eine Vielzahl von verschiedenen möglichen Merkmalen in der Autorschaftsanalyse aufgezeigt. Einige dieser Merkmale konnten konsolidiert oder vorab selektiert werden. Im Folgenden soll eine detaillierte Beschreibung der für die in dieser Arbeit verwendeten Merkmale erfolgen.

Dieses Kapitel begründet die Auswahl dieser Merkmale und zeigt den erwarteten Informationsgehalt und mögliche Herausforderungen auf. Grundlage dieser Erwartungen sind vorrangig englischsprachige Studien und Arbeiten, welche bereits einige dieser Merkmale untersucht haben. Insbesondere gibt die Dissertation von Nini (2015) einen Überblick über existierende Forschungen im Bereich der Profilierung.

Die Auswertungen der Merkmale in den folgenden Kapiteln sind lediglich Durchschnitts- und Summenbetrachtungen. Sie können daher lediglich als Indiz für die Wirksamkeit eines Merkmales dienen. Sie werden deshalb als Grundlage für Thesen genutzt, welche in den späteren Kapiteln Datenvorbereitung und Modellierung zu prüfen sind.

5.6.1. Bag of Words

BoW ist eine vereinfachte Darstellung eines Textes, welche bei der Verarbeitung natürlicher Sprache verwendet wird. Das Modell beschreibt einen Text anhand der enthaltenen Wörter und ihrer Häufigkeit, ungeachtet der Grammatik oder Reihenfolge. Damit nicht jede morphologische oder orthographische Variante eines Wortes als eigenes Wort behandelt wird, werden die Texte meist normalisiert. Neben der in Kapitel 4.3 beschriebenen Textnormalisierung werden zusätzlich Funktionswörter entfernt, da diese als eigenes Merkmal betrachtet werden.

Besonders häufig wird BoW in der Klassifikation von Texten in Themen wie beispielsweise „Sport“, „Politik“ oder „Wissenschaft“ verwendet, da hier allein das Vorkommen eines Wortes wie „Fußball“ oder „Bundestagswahlen“ einen Indikator für das Genre darstellt (Kaster, Siersdorfer, & Weikum, 2005, S. 1). Die Themenabhängigkeit ist gleichzeitig die größte Kritik an diesem Merkmal in der Autorschaftsattribution (vgl. Kapitel 5.5). Für das Bestimmen des Autors sind nur die Wörter relevant, welche ein Autor themenunabhängig in einer anderen Häufigkeit als andere Autoren verwendet. Gründe hierfür können Dialekte oder die soziale Klasse des Autors sein.

Die Anzahl der ermittelten Wörter der Textsammlung beträgt 70.886 verschiedene Wörter. Vergleichsweise führt Boukkouri (2018) bei einer Textsammlung von 50.000 IMDb Film Kritiken einen Vektor in der Länge von 80.000 Worten an. In der deutschen Sprache ist von einer größeren Anzahl von Wörtern auszugehen. Dies liegt beispielsweise an einer Vielzahl von gebildeten Komposita, der Verwendung von Neologismen und englischen Worten, welche beispielsweise durch Bandnamen oder Anglizismen in den Texten vorkommen.

In der Betrachtung der fünf meist verwendeten Wörter der Autoren zeigen sich nur geringe Unterschiede, beispielsweise themenspezifische Worte, wie „Album“, „Musik“ oder „Song“ (vgl. Tabelle 2). Da für diese Worte wenige Synonyme existieren, sind diese bei jedem Autor in einer hohen Häufigkeit vertreten. Die Häufigkeit des Wortes „Fazit“ erklärt sich durch das Fazit am Ende vieler Artikel, welches hierdurch eingeleitet wird. Englische Wörter sind in den deutschen Stoppwörtern nicht enthalten und werden als Teil eines Bandnamens auch nicht einer der exkludierten Wortarten zugeordnet. Dies hat zur Folge, dass das Wort „the“ sehr häufig enthalten ist.

Tabelle 2: Top 5 der häufigsten Wörter nach Autoren – Eigene Darstellung

		Position				
		1	2	3	4	5
Autor	1	the	album	musik	fazit	band
	2	the	of	album	fazit	band
	3	the	album	jahr	musik	song
	4	album	song	the	band	fazit
	5	album	band	the	song	metal
	6	metal	band	album	the	klingen
	7	album	the	band	titel	jahr
	8	song	the	album	band	geben
	9	album	song	band	the	fazit
	10	the	album	mal	fazit	jahr
	11	the	album	fazit	of	stück
	12	song	album	band	the	fazit
	13	the	band	song	mal	album
	14	the	of	band	album	fazit
	15	band	album	song	death	the
	16	song	album	band	the	fazit
	17	fazit	band	the	song	album
	18	band	metal	album	the	of
	19	song	album	band	the	metal
	20	album	song	band	fazit	geben
	21	the	album	song	metal	stück
	22	band	klingen	album	song	the
	23	the	song	album	band	klingen
	24	the	band	album	fazit	jahr
	25	the	album	band	song	of

Insgesamt besitzt dieses Merkmal eine sehr hohe Anzahl an Dimensionen. Dies wirkt sich entsprechend auf die benötigte Rechenleistung des Modells aus. Des Weiteren ist durch den hohen Anteil von irrelevanten Dimensionen mit einem beeinträchtigten Ergebnis zu rechnen. Aus diesem Grund müssen die Dimensionen in der Datenvorbereitung minimiert werden.

5.6.2. Wort-N-Gramme

BoW stellt eine einfache Möglichkeit dar, ignoriert dabei aber die Reihenfolge der Wörter und dadurch gegebenenfalls Informationen über den Kontext. Aus diesem Grund werden oftmals Wort-N-Gramme als Merkmal vorgeschlagen (Peng, Schuurmans, & Wang, 2004; Guenter & Sanderson, 2006). Halvani, Steinebach, & Neitzel (2014) führen Kollokationen und Redewendungen als explizites Merkmal an. Eine bekannte Redewendung wäre zum Beispiel „starker Tobak“. Eine Textnormalisierung erfolgt für das Extrahieren der Wort-N-Gramme gleichermaßen zum vorherigen Kapitel.

Die durch Wort-N-Gramme erzielte Klassifikationsgenauigkeit ist jedoch nicht immer besser als die einzelnen Worte (Sanderson & Guenther, 2006; Coyotl-Morales, et al., 2006). Als Kritik werden von Stamatatos (2009) die folgenden Punkte angeführt.

Die Dimensionalität nimmt mit dem Verwenden von Wort-N-Grammen erheblich zu, da alle möglichen Kombinationen von Worten berücksichtigt werden. Dies kann anhand der verwendeten Datengrundlage bestätigt werden (vgl. Tabelle 3). Dabei ist anzumerken, dass die Dimensionalität bei größeren N-Grammen unwesentlich steigt, da die meisten N-Gramme ohnehin einmalig in den Artikeln zu finden sind.

Dies führt auch zum nächsten Kritikpunkt, dass die meisten Wortkombinationen nur selten in mehreren Texten vorkommen (ebd.). Für einen Klassifikationsalgorithmus ist es deshalb schwierig, diese Merkmale effektiv zu behandeln. Dieser Punkt steht im Konflikt mit der erwähnten steigenden Dimensionalität bei einem größeren N. Durch ein Eliminieren aller Wort-N-Gramme, welche nur einmalig über alle Artikel vertreten sind, können die

Ausprägungen des Merkmals um ein Vielfaches reduziert werden, wodurch auch die Dimensionalität bei einem größeren N sinkt.

Tabelle 3 zeigt die Anzahl der verschiedenen N-Gramme, die Summe der N-Gramme, welche öfter als einmal vorkommen und die fünf häufigsten N-Gramme für n von zwei bis sechs. Die N-Gramme werden dabei durch einen senkrechten Strich („|“) getrennt.

Tabelle 3: Top 5 Wort-N-Gramme – Eigene Darstellung

	Anzahl	Summe > 1	Top 5 Wort-N-Gramme
2-Gramme	399.649	33.469	1. of|the 2. death|metal 3. black|metal 4. neu|album 5. heavy|metal
3-Gramme	466.107	6.151	1. al|di|meola 2. rock|n|roll 3. melodic|death|metal 4. of|the|world 5. pain|of|salvation
4-Gramme	470.928	2.215	1. di|meola|al|di 2. meola|al|di|meola 3. al|di|meola|al 4. rock|´|n|´ 5. ´|n|´|roll
5-Gramme	470.136	1.213	1. al|di|meola|al|di 2. di|meola|al|di|meola 3. rock|´|n|´|roll 4. arrangements|and|additional|music|al 5. and|additional|music|al|di
6-Gramme	468.275	832	1. al|di|meola|al|di|meola 2. and|additional|music|al|di|meola 3. arrangements|and|additional|music|al|di 4. additional|music|al|di|meola|al 5. music|al|di|meola|al|di

Zuletzt wird die Themenabhängigkeit der N-Gramme kritisiert (Stamatatos, 2009). Anstatt dem Erfassen von stilistischen Informationen werden unter Umständen ausschließlich inhaltsspezifische Informationen aufgenommen. Auch dies wird durch Tabelle 3 im hier verwendeten Korpus bestätigt. Unter den meist

vorkommenden N-Grammen befinden sich vorrangig Genres, Band- oder Sängernamen. Zusätzlich zeigen sich auch Fehler des PoS-Taggings bei N-Grammen wie „rock|´|n|´|roll“, bei denen Sonderzeichen enthalten sind. Die hohe Häufigkeit des Jazz Gitarristen „Al Di Meola“ lässt sich durch die angefügten Credits in einem Artikel erklären, in denen Songtexte des Musikers aufgelistet werden. Dabei wird sein Name abzüglich von Zeilenumbrüchen und Sonderzeichen zweimal hintereinandergeschrieben.

Auch dieses Merkmal besitzt, unabhängig von der Länge der N-Gramme eine zu große Anzahl an Dimensionen, welche mit einer erhöhten Rechenleistung einhergeht. Daher sind auch hier die Dimensionen in der Datenvorbereitung zu minimieren. Ebenfalls gilt es zu prüfen, welche Größe von N die besten Ergebnisse verspricht.

5.6.3. Anzahl der Wörter

Die Anzahl der Wörter bzw. die Textlänge kann in Abhängigkeit vom Alter oder der sozialen Klasse des Autors variieren (Nini, 2015, S. 167). Bromley (1991) fand in seiner Studie Ergebnisse in Bezug auf das Bildungsniveau. Die Grundlage seiner Studie waren 9 Essays, welche von 240 Erwachsenen im Alter von 20 bis 86 Jahren geschrieben wurden. Seine linguistische Analyse ergab, dass lexikalische Variablen wie Textlänge, Anzahl der Wörter mit mehr als 10 Buchstaben, durchschnittliche Wortlänge und der Flesch-Lesbarkeitsscore eine Kombination aus Status und Niveau des Vokabulars voraussagen konnten. Der Status war hierbei eine Kombination aus Bildungsniveau und beruflichem Status.

Hinweise auf eine Abhängigkeit der Textlänge zum Alter der Autoren bestätigt Nini (2015, S. 145). In ihrer Dissertation prüfte sie unter anderem die Textlänge als linguistische Variable an dem Korpus „The Fabricated Malicious Texts“, welcher 287 Texte von 96 Subjekten beinhaltet. Sie beschrieb die Zunahme der Prägnanz als ein weiteres Muster linguistischer Variation, welche das Alter charakterisiert. Diese zeigt sich vorrangig durch die Abnahme der Textlänge. Eine mögliche Erklärung dafür ist, dass ältere Probanden versuchen, prägnant und direkt zu sein, und dass sie andere Formen der Komplexität bevorzugen

(ebd.). Eine andere Theorie nach Kemper, Kynette, Rash, O'Brien, & Sprott (1989) lautet, dass die verringerte Textlänge mit dem Verlust des Arbeitsgedächtnisses zusammenhängt. Für einen genauen Nachweis hätte allerdings das Messen des Gedächtnisses erfolgen müssen.

Betrachtet man die einzelnen Autoren, lassen sich klare Unterschiede in der durchschnittlichen Textlänge erkennen. Um diese besser aufzuzeigen, wurde der Datensatz standardisiert. In Bild 2 entspricht dies den Balken. Die durchschnittlich verwendete Anzahl der Wörter wird durch die Linie und zweite Skala auf der Ordinate dargestellt.

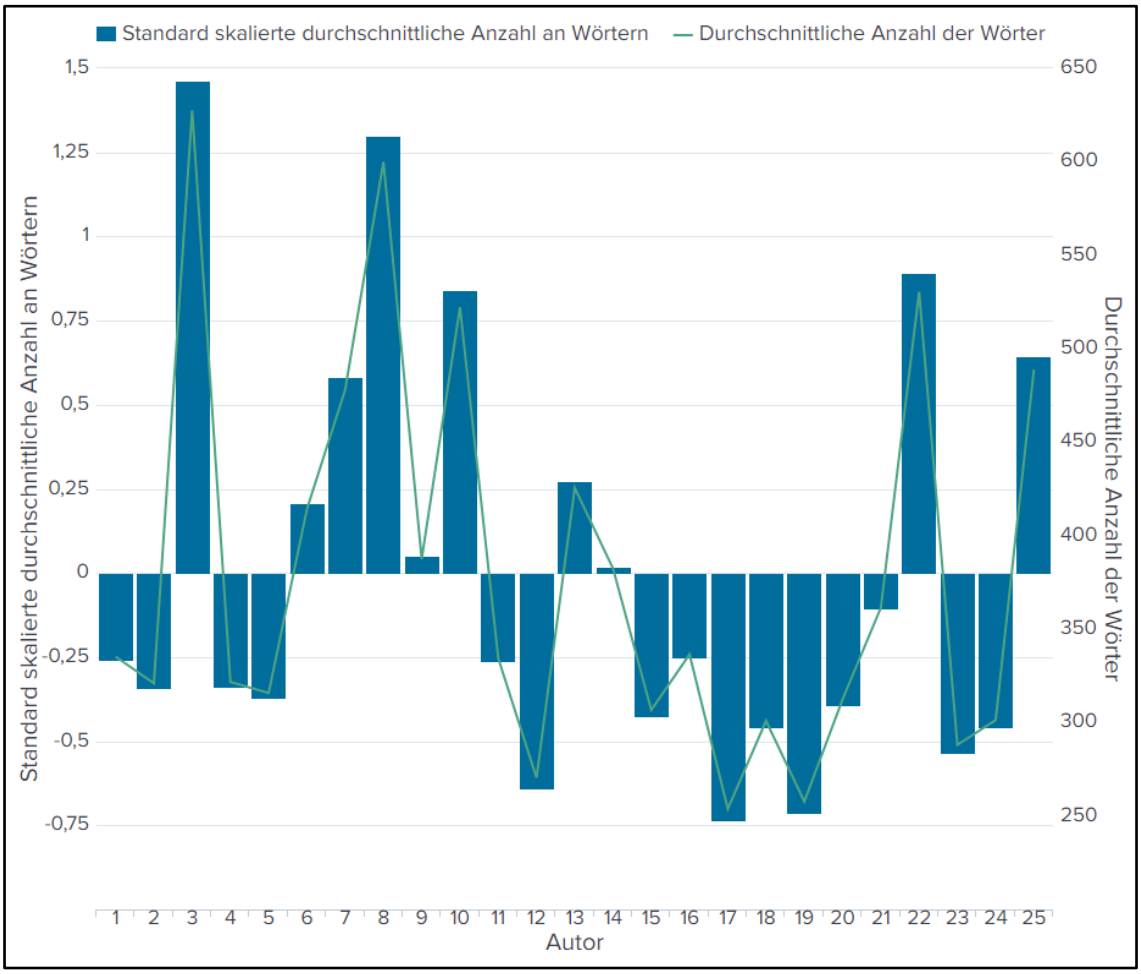


Bild 2: Durchschnittliche Anzahl der Wörter - Eigene Darstellung

Autor 3 verwendet mit durchschnittlich 627,54 die meisten Wörter. Im Gegensatz dazu verwendet Autor 17 durchschnittlich nur rund 40% der Wörter (253,82). Insgesamt existiert eine geringere Anzahl von Autoren mit Texten größer als dem

Mittelwert. Dieses Ungleichgewicht wird durch die Autoren kompensiert, welche dazu tendieren, sehr lange Texte zu schreiben.

5.6.4. Wortlänge

Die Wortlänge wird in vier von fünf der genannten Arbeiten verwendet, alleinig Potha & Stamatatos (2014) verzichten auf das Verwenden des Merkmals. Die häufige Nutzung lässt sich durch die Ergebnisse einiger Studien erklären. In einer Einführung in die Forschung zur Lesbarkeit enthält DuBay (2004) unter anderem einen historischen Rückblick. Er zeigt, dass bereits die US-Armee bei ihren Forschungen zur Lesbarkeit feststellten, dass die Messungen der Lese- und Schreibfähigkeit als auch der Lesbarkeit mit der Intelligenz, dem Wissen und der erreichten Ausbildungsdauer korrelierten. Dazu anlehnend bemerkte Kitson (1921), dass signifikante Unterschiede bezüglich der durchschnittlichen Satzlänge und der Wortlänge zwischen Zeitschriften in Abhängigkeit zu ihrer Leserschaft und damit die Art der sozialen Gruppen existieren, die eine bestimmte Zeitschrift lesen würden. Aufgrund der Annahme, dass eine Person nur Texte in einer Sprache produzieren kann, die sie auch lesen kann, lassen sich die Gruppierungen auf die schreibenden Personen übertragen. (Nini, 2015, S. 60)

Zu den bisher formulierten Hypothesen testete Bernstein (1962) sein Modell an einem Datensatz, der aus Interviews mit 61 männlichen Probanden aus der Arbeiterklasse und 45 aus der Mittelschicht, alle im Alter von 15 bis 18 Jahren, besteht. Eine Einteilung in soziale Klassen erfolgte nur aufgrund der Ausbildung der Probanden. Durch seine Analyse konnte er einen Unterschied in der Wortlänge zwischen den beiden sozialen Klassen nachweisen. Existierende Unterschiede durch den IQ der Teilnehmenden konnte er dabei ausschließen.

In der Literatur finden sich verschiedene Ansätze für das Bestimmen der Wortlänge. Beispielsweise wird sie in Nini (2015, S. 225) als durchschnittliche Wortlänge, also der Anzahl der Buchstaben eines Textes geteilt durch dessen Wörter, definiert. In anderen Arbeiten, wie Narayanan, et al., (2012) oder Abbasi & Chen (2008), wird die Häufigkeit von Wörtern bestehend aus einem bis zu

zwanzig Buchstaben bestimmt. Dieser Ansatz verspricht durch einen höheren Informationsgehalt. Beispielsweise wird die Verwendung von Personalpronomen in der englischen Sprache mit einer Länge von zwei oder drei und im deutschen drei bis vier Buchstaben erkannt. Es können aber auch Autoren erkannt werden, welche einen Hang zu langen Worten besitzen. Grundsätzlich ist in der deutschen Sprache, beispielsweise wegen langer Komposita, mit längeren Worten zu rechnen.

Bild 3 zeigt die Summe der Wörter aufgeteilt nach deren Wortlänge. Offensichtlich ist die Dominanz der Wörter mit einer Länge von drei Zeichen. Wörter mit einer Länge von einem Zeichen lassen sich durch englische Bandnamen oder Liednamen erklären (Beispielsweise „a dance“). Bis zu den Wörtern mit einer Länge von über 20 Zeichen zeigt sich ein streng monoton fallender Verlauf. Diese werden aufgrund ihrer vergleichsweise geringen Anzahl zusammengefasst.

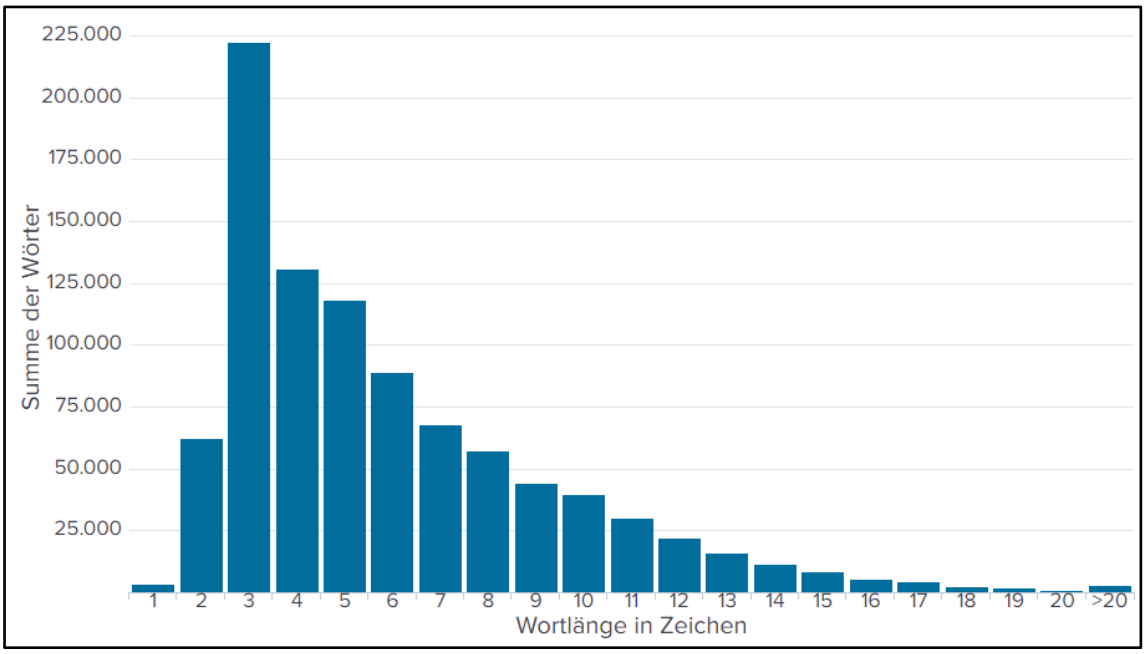


Bild 3: Wortlängen der Artikel - Eigene Darstellung

Es existieren in der Textgrundlage sehr lange Wörter mit bis zu 76 Zeichen. Diese entstehen teilweise durch Neologismen oder Komposita mit oder ohne Einsatz von Bindestrichen, weshalb die Wortlänge auch in Zeichen und nicht in Buchstaben gemessen wird. Beispiele für lange Wörter sind

„Hochspannungsalleinstellungsmerkmal“ mit 35 Buchstaben, „viertausendsechshundertdreiundfünfzigste“ mit 41 Buchstaben oder ein Satz-Adjektiv-Komposita wie "unsere-Musik-klingt-wie-diese-majestetisch-furchteinflößende-Skyline-aussieht" mit 76 Zeichen. Die Kenntnis und Neigung zur Verwendung von langen Wörtern, insbesondere langer Komposita unter der Verwendung von Bindestrichen zeigen eine Eigenschaft, welche nicht jedem Autor zuzuordnen ist. Da hier ein deutlicher Unterschied zur englischen Sprache besteht, werden nicht nur die Wortlängen bis 20 Zeichen als Merkmale in Betracht gezogen.

In der Theorie existiert keine maximale Wortlänge und es gibt eine unendliche Anzahl an möglicher Merkmale. Fraglich ist allerdings, in wie weit die genaue Länge eines Wortes bei sehr langen Worten ausschlaggebend ist. Es gilt die These zu prüfen, ob ein Zusammenfassen der langen Wörter in Intervalle einen höheren Informationsgewinn liefert als jede Wortlänge einzeln zu betrachten. Möglicherweise ist die Neigung eines Autors, sehr lange Wortkombinationen zu verwenden, aussagekräftiger als die tatsächliche Wortlänge.

Im Hinblick auf die durchschnittliche Wortlänge der Autoren besteht die Vermutung, dass diese einen geringeren Informationsgehalt besitzt. Begründet wird diese Vermutung mit der dominanten Anzahl von Wörtern mit drei oder vier Zeichen. Das Verwenden langer Wörter sollte durch die verhältnismäßig geringe Anzahl nicht im Durchschnitt erkennbar sein. Daraus folgen drei zu prüfende Thesen. Die Durchschnittsbildung

- liefert durch den geringeren Informationsgehalt schlechtere Ergebnisse
- verringert durch das Zusammenfassen irrelevante Merkmalsausprägungen und verbessert dadurch die Ergebnisse
- liefert ungefähr gleiche Ergebnisse mit einem geringeren Rechenaufwand

5.6.5. Wortschatzreichtum

Susan Kemper und Kollegen betrieben Untersuchungen über den Alterungsprozess und dessen Auswirkungen auf die Sprache im Alter. Auch

wenn es nicht der Hauptuntersuchungsgegenstand in der Psycholinguistik ist, ergaben sich einige Ergebnisse bezüglich des Wortschatzes und dessen Korrelation zum Bildungsniveau und Alter der Probanden.

1989 führten Kemper, Kynette, Rash, O'Brien, & Sprott ein Experiment an 30 jungen Erwachsenen (im Alter 18 bis 28 Jahren) und 78 älteren Erwachsenen (im Alter von 60 bis 92 Jahren) durch. Die Teilnehmer mussten ein strukturiertes Interview sowie eine mündliche offene Frage beantworten und einen kurzen schriftlichen Aufsatz verfassen. Dabei Fand eine Kontrolle der meisten Merkmale mit einem Einfluss auf die Satzkomplexität statt, wie beispielsweise dem Bildungsniveau. Ein Ergebnis war, dass Personen mit mehr Bildungsjahren dazu neigten, beim Vokabeltest eine höhere Punktzahl zu erzielen.

Dies bestätigte sich in der Analyse eines Datensatzes, welcher aus mündlichen Erzählungen älterer Erwachsener im Alter von 60 bis 90 Jahren besteht (Cheung & Kemper, 1992). Das Ziel der Studie war das Finden eines theoretischen Modells, welches die Zusammenhänge zwischen Alter, Arbeitsgedächtnis, verbalen Fähigkeiten und sprachlicher Komplexität erfolgreich vorhersagen kann, dennoch fanden die Autoren heraus, dass Messungen des Vokabulars nur mit dem Bildungsniveau und nicht mit dem Alter positiv korrelierten.

Im Gegensatz dazu fanden Kemper & Sumner (2001) in einer Untersuchung einer Stichprobe von 100 jungen Erwachsenen (im Alter von 18 bis 28 Jahren) und 100 älterer Erwachsener (im Alter von 63 bis 88 Jahren) heraus, dass ältere Erwachsene unter anderem einen reicheren Wortschatz besitzen. Diese Ergebnisse konnte Rabaglia & Salthouse (2011) in einer neueren Replikation von Kempers Studie bestätigen. Hierfür sammelte sie Sprachproben, die aus kurzen Texten von etwa 900 Erwachsenen im Alter von 18 bis 90 Jahren bestanden. Sie bezeichneten den Vokabeltest als einen Vertreter für gesammeltes Weltwissen und schlugen als Begründung vor, dass mit zunehmendem Alter das Weltwissen der Menschen steigt. (Nini, 2015, S. 62 ff.)

Eine Möglichkeit die Größe des Wortschatzes zu bestimmen, ist das Ermitteln von einmalig im Text vorkommenden Wörtern, dem sogenannten Hapax legomenon. Hier besteht allerdings eine starke Abhängigkeit zur Länge des

Textes und somit eine Abhängigkeit zum Inhalt, die es zu vermeiden gilt (vgl. Kapitel 5.5). Im Gegensatz dazu stehen Konstanzmaße für Texte in natürlicher Sprache. Sie werden als ein rechnerisches Maß definiert, welches den lexikalischen Reichtum literarischer Texte misst und dabei für eine bestimmte Textmenge zu einem Wert konvergiert (Aihara & Tanaka-Ishii, 2015, S. 1). Es bleibt dabei für jede größere Textmenge invariant. Ein solches Textkonstanzmaß wurde von Yule (1944) in Form seines Maßes K eingeführt. K wird dabei nach Aihara & Tanaka-Ishii (2015, S. 3) folgendermaßen definiert:

$$K = C\frac{S_{2-}S_1}{S_1^2} \tag{5.6.5.1}$$

, wobei $S1 = N = \sum_m mV(m,N)$, $S2 = m^2V(m,N)$ und $C = 10^4$

$$= 10^4\left[-\frac{1}{N} + \sum_{m=1}^{m_{max}} V(m,N)\left(\frac{m}{N}\right)^2\right] \tag{5.6.5.2}$$

Dabei entspricht „N“ den gesamten Wörter eines Textes und „$V(m, N)$“ der Anzahl der Wörter, welche „m“ mal im Text auftauchen. K soll den Wortschatzreichtum eines Textes messen: Je größer Yules K, desto weniger reichhaltig ist der Wortschatz.

Bild 4 stellt den durchschnittlichen Wert von Yules K für die einzelnen Autoren im Vergleich zum Gesamtdurchschnitt dar. Die Texte wurden vor dem Berechnen dahingehend normalisiert, dass alle Wörter klein geschrieben und lemmatisiert wurden. Von einer Eliminierung der Stoppwörter wurde bewusst abgesehen, da sich auch hier Variationen der Wörter zeigen können. Die Mehrheit der Autoren besitzt einen unterdurchschnittlichen Wert von K bzw. einen überdurchschnittlichen Wortschatz. Drei Autoren dominieren die Statistik mit einem besonders hohen Wert.

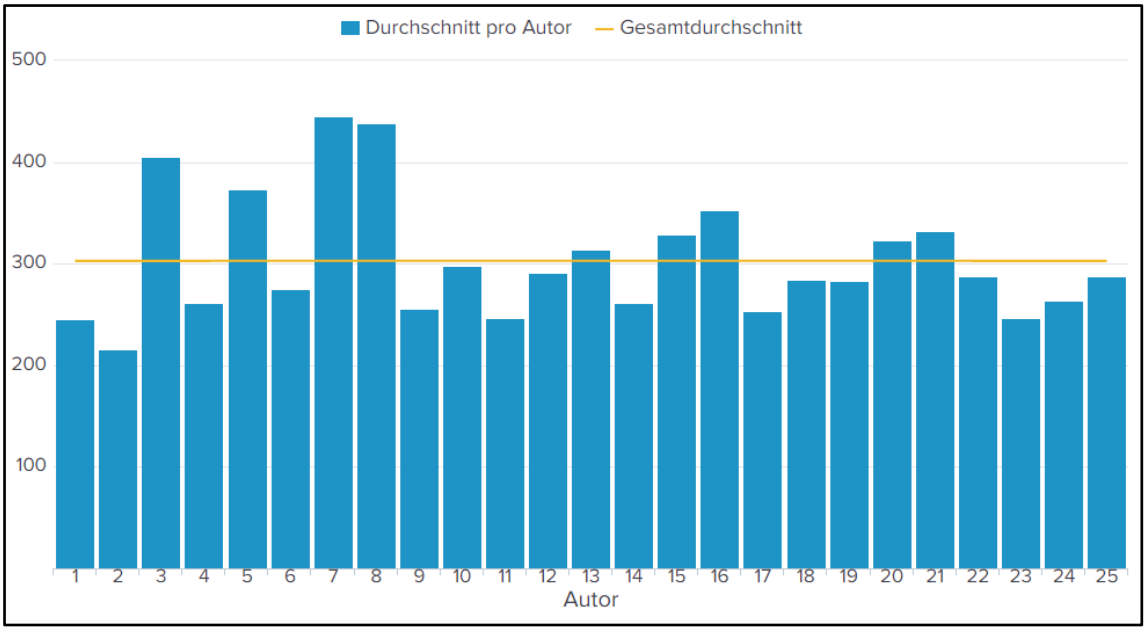


Bild 4: Wortschatzreichtum der Autoren im Vergleich - Eigene Darstellung

Verdeutlicht wird dieser Sachverhalt in Bild 5. Hier werden alle Werte von Yules K für die Autoren auf den Ordinaten aufgetragen. Die Autoren werden auf der Abszisse in einem Intervall von Label[9] abzüglich eins und Label definiert: Beispielsweise werden Artikel des Autoren 1 zwischen null und eins dargestellt. Die obere und untere Grenze für Ausreißer wird als das Doppelte der Standardabweichung plus/minus den Mittelwert definiert. In einem Vergleich von Bild 4 und Bild 5 zeigt sich, dass die durchschnittlichen Werte kaum durch Ausreißer dominiert werden und sich die Autoren meist in ihrem festgelegten Bereich bewegen. Eine Ausreißer-Quote von durchschnittlich 4,12% spricht ebenfalls für eine Normalverteilung der Werte, bei der näherungsweise gilt, dass 99,45% der Werte innerhalb einer Abweichung von plus/minus der zweifachen Standardabweichung liegen. Mit einer höheren Standardabweichung oder Anzahl von Ausreißern sinkt der Informationsgehalt des Merkmals in Bezug auf die Klassifikation eines Autors.

[9] Das Label entspricht der pseudonymisierten Nummer eines Autors.

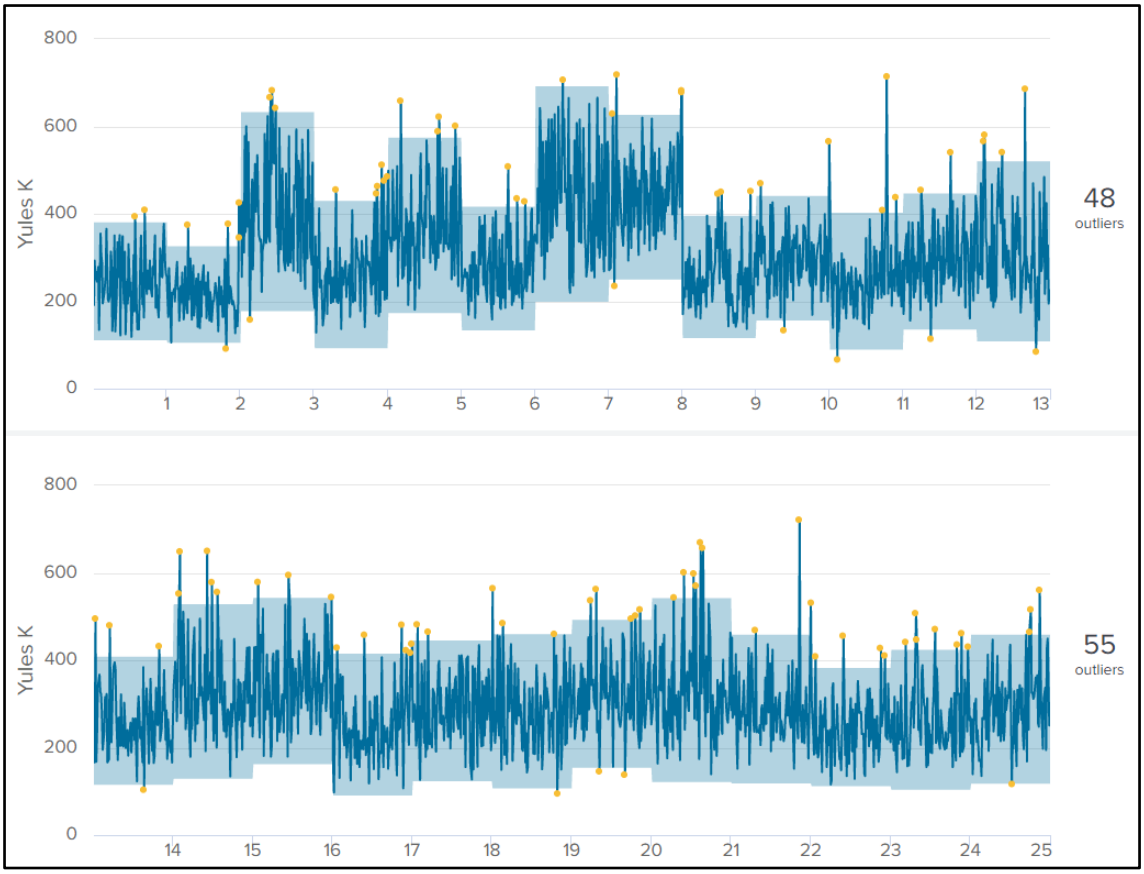


Bild 5: Ausreißer Wortschatzreichtum - Eigene Darstellung

5.6.6. Sonderzeichen

Das Merkmal beschreibt die Häufigkeit der verwendeten Sonderzeichen in einem Text. Das Verwenden von Sonderzeichen ist weniger abhängig vom Inhalt des Textes, wodurch Aussagen über den individuellen Schreibstil eines Autors möglich sind (Sapkota, Bethard, Montes, & Solorio, 2015, S. 93). Beispielsweise tendieren weibliche Autoren eher zum Setzen von Kommas vor Konjunktionen. Halvani & Steinebach (2014) interpretieren daraus, dass diese mehr Gebrauch von den optionalen Regeln der neuen Rechtschreibreform von 2006 machen als männliche Autoren. Dies würde in einer erhöhten Häufigkeit von Kommata resultieren. Meist ungeachtet in der Themenabhängigkeit der Sonderzeichen sind Smileys in beispielsweise E-Mails oder Direktnachrichten. Hierbei ist fraglich, wie sich dies auf die Performanz des Merkmals innerhalb der Domäne und domänenübergreifend auswirkt, ein Prüfen dessen ist anhand der Datengrundlage nicht möglich.

Die in der Arbeit verwendeten Sonderzeichen beinhalten alle Interpunktionszeichen und werden um einige gängige Sonderzeichen erweitert. Zwar existieren innerhalb des Textes einige weitere spezielle Sonderzeichen wie „‰“ oder „¿“, allerdings wird von einer automatischen Generierung der extrahierten Sonderzeichen abgesehen, da diese als sehr spezifisch für die Datengrundlage dieser Arbeit angesehen werden und dadurch keine Relevanz für andere Arbeiten vermutet wird, die sich mit der deutschen Sprache befassen. Daraus ergibt sich die folgende Liste an Sonderzeichen:

! „ “ § $ % & / ( ) = ? { } [ ] \ @ # , ‘ - _ + * . , ;

Tabelle 4: Hitzekarte Sonderzeichen - Eigene Darstellung

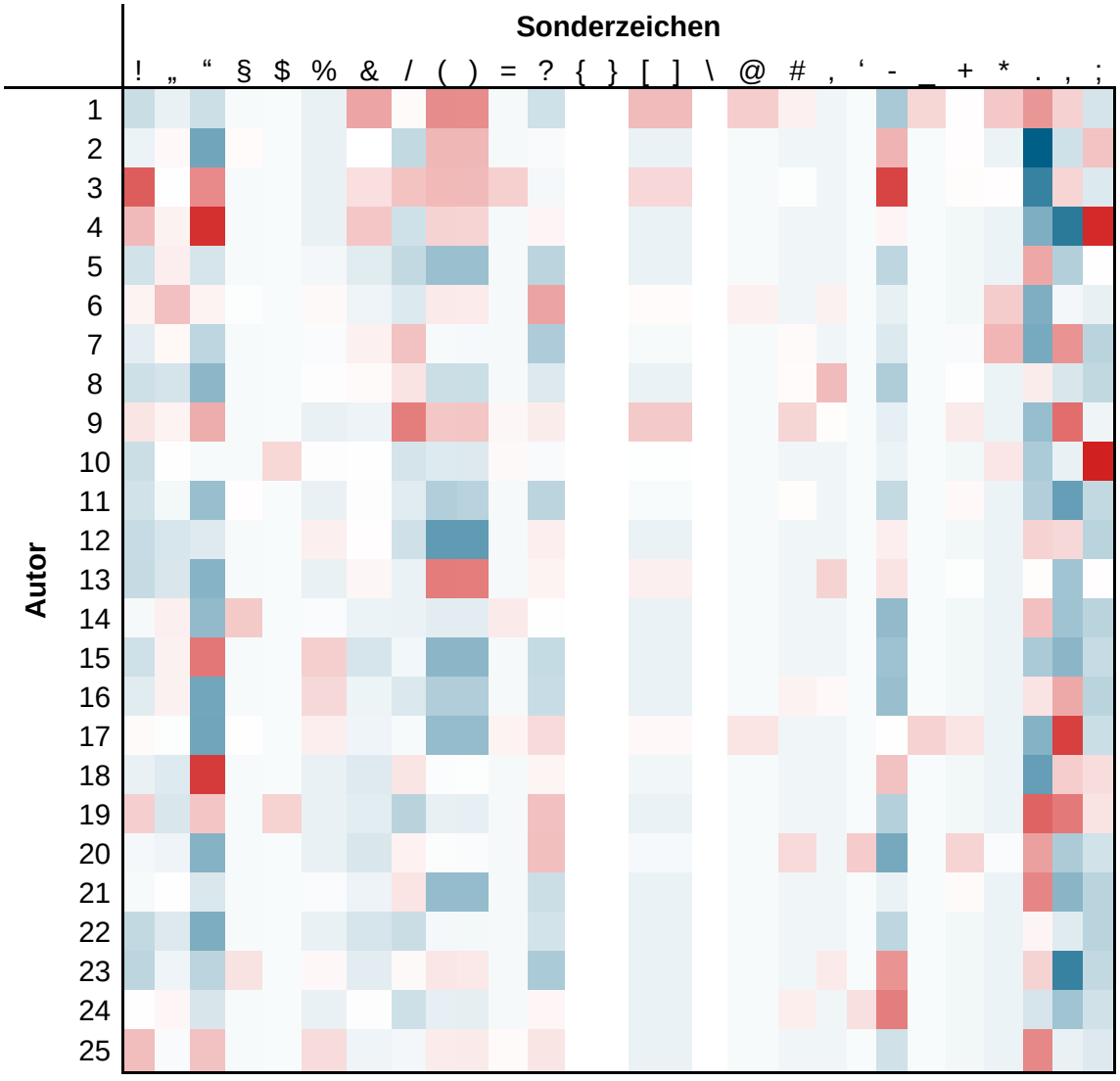


Tabelle 4 beschreibt die verwendeten Sonderzeichen der Autoren in Form einer Hitzekarte. Hohe Werte werden rot, tiefe Werte blau und die Mitte weiß dargestellt. Auffällig zeigt sich die häufige Verwendung von Bindestrichen bei Autor 3, dies lässt sich mit seiner in Kapitel 5.6.4 festgestellten Neigung erklären, sehr lange Komposita zu bilden. Weiterhin lassen sich klare Differenzen bei Punkten, Kommata, Semikola und Anführungszeichen erkennen. Einige Autoren scheinen nur obere Anführungszeichen zu verwenden. Bei wiederum anderen Zeichen lassen sich keine Unterschiede erkennen. Beispielsweise werden geschweifte Klammern oder Paragraphenzeichen augenscheinlich von den Autoren nicht verwendet. Insgesamt sind deutlichere Unterschiede bei den Häufigkeiten der Interpunktionszeichen zu erkennen.

5.6.7. Zeichen-N-Gramme

Zeichen-N-Gramme gehören zu den umstrittensten Merkmalen in der Autorschaftsanalyse. Wie auch bei BoW oder Wort-N-Grammen steht die hohe Wirksamkeit der Themenabhängigkeit gegenüber. So greifen alle Autoren außer Narayanan, et al. (2012) der in Kapitel 5.3 genannten Literatur auf das Merkmal zurück. Narayanan, et al. (2012) sprechen sich hier explizit gegen die Verwendung des Merkmals aus, wohingegen Potha & Stamatatos (2014) ausschließlich N-Gramme verwenden. Nach Koppel, Schler, & Argamon (2011) handelt es sich um das Merkmal mit den besten Ergebnissen. Trotz dieser Kontroverse wurden die Hintergründe für die Wirksamkeit, als auch die Themenabhängigkeit nicht vollständig verstanden (Sapkota, Bethard, Montes, & Solorio, 2015). Eine Hypothese besteht darin, dass das Feature ein wenig von allen Informationen aufnimmt: Lexikalischer Inhalt, Syntaktik, Interpunktion und Leerzeichen (ebd.). Dies führte zu einer genaueren Studie ihrerseits. Aufgrund der bedeutsamen Ergebnisse wird diese im Folgenden erläutert.

Zuerst wurde eine grundlegende Unterteilung der N-Gramme in drei Supergruppen aufgrund der linguistischen Aspekte vorgenommen, von denen angenommen wird, dass sie durch Zeichen-N-Gramme repräsentiert werden: Morpho-Syntax (repräsentiert durch Zeichen-Affix-N-Gramme), thematischer

Inhalt (repräsentiert durch Zeichen-Wort-N-Gramme) und Stil (repräsentiert durch auf Zeichen-Interpunktion-N-Gramme).

Zeichen-Affix-N-Gramme können einen gewissen Grad der Morphologie widerspiegeln, indem sie beispielsweise Derivationen[10] repräsentieren. Sie beinhalten die folgenden Merkmale:

- Präfix: Beinhaltet die ersten n Zeichen eines Wortes, welches mindestens n+1 Zeichen lang ist.
- Suffix: Beinhaltet die letzten n Zeichen eines Wortes, welches mindestens n+1 Zeichen lang ist.
- Leerzeichen-Präfix: Ein N-Gramm, welches mit einem Leerzeichen beginnt.
- Leerzeichen-Suffix: Ein N-Gramm, welches mit einem Leerzeichen endet.

Auch Halvani, Steinebach, & Neitzel (2014, S. 241) verwenden Prä-, In- und Suffixe als gesonderte Merkmale.

Zeichen-Wort-N-Gramme sind oft zu kurz, um ganze Worte zu erfassen, weshalb sie nur kurze Wörter oder Teile von Wörtern erfassen. Sie unterteilen sich in:

- Gesamt-Wort: Beinhaltet ein Wort mit genau n Buchstaben.
- Mittel-Wort: Enthält n Zeichen eines Wortes, welches mindestens n+2 Zeichen lang ist. Es darf weder den ersten, noch den letzten Buchstaben beinhalten.
- Multi-Wort: Erfassen Teile mehrerer Wörter und besitzen ein Leerzeichen in der Mitte.

Zeichen-Interpunktion-N-Gramme können die wichtigsten stilistischen Entscheidungen eines Autors erfassen. Dazu gehören Interpunktionsmuster oder gewählte Präfixe nach einem Satzzeichen. Die Kategorie unterteilt sich in:

[10] Derivation bezeichnet Wortbildung durch Kombination von Wortstämmen und Affixen.

- Anfangs-Interpunktion: Beim ersten Zeichen handelt es sich um ein Interpunktionszeichen, bei den mittleren aber nicht.
- Mittel-Interpunktion: Bei mindestens einem mittleren Zeichen (nicht das erste oder letzte Zeichen) handelt es sich um ein Interpunktionszeichen.
- End-Interpunktion: Beim letzten Zeichen handelt es sich um ein Interpunktionszeichen, bei den mittleren aber nicht.

Als Grundlage der Auswertung wurden zwei verschiedene Datengrundlagen verwendet. Der „CCAT“, eine Teilmenge des „Reuters Corpus Volume 1“ besitzt nur Texte über ein Thema (Lewis, Russel-Rose, Yang, Li, & Dietterich, 2004). Der zweite Korpus hingegen umfasst Texte von 13 verschiedenen Autoren über vier sich unterscheidenden Themen, veröffentlicht in der „Guardian Newspaper“ (Stamatatos, 2013). Dieser Korpus wird nachfolgend Guardian genannt.

Die Analyse der Merkmale über die beiden Korpora ergab, dass unabhängig vom verwendeten Algorithmus Präfix, Suffix, Leerzeichen-Präfix und Mittel-Wort für eine Domäne und Präfix, Leerzeichen-Präfix, Anfangs-Interpunktion, und Mittel-Interpunktion für mehrere Themen die informativsten Merkmale sind. Zusammenfassend konnten die Autoren feststellen, dass Zeichen-Affix-N-Gramme in beiden Szenarien, Zeichen-Interpunktion-N-Gramme in themenüberschneidenden Szenarien und Zeichen-Wort-N-Gramme nur bei Auswertungen über den Guardian effektiv sind. Überschnitten sich die Themen der Artikel, ergaben die Messungen, dass jedes Merkmal der anderen Kategorien besser geeignet war als das beste Merkmal der wortähnlichen N-Gramme.

Die Autoren stellten sich nach diesem Ergebnis die Frage über die Relevanz einiger Merkmale und testeten zwei Modelle:

- Alle Kategorien
- Zeichen-Affix-N-Gramme kombiniert mit Zeichen-Interpunktion-N-Grammen

Bei den Dokumenten des CCAT erzielte das Modell ohne die wortähnlichen Merkmale gleiche oder bessere Ergebnisse als mit diesen. Dies war aufgrund der Werte aus dem ersten Versuch wenig überraschend. Überraschender war, dass

dieses Ergebnis bei den Dokumenten des Guardian wiederholt werden konnte, obwohl das am besten bewertete Merkmal „Mittel-Wort“ eliminiert wurde. Dies wurde dahingehend interpretiert, dass einige Merkmale der anderen Kategorien zumindest teilweise dieselben Informationen beinhalten müssen. Dies festigte die These, dass Zeichen-Affix- und Zeichen-Interpunktion-N-Gramme wichtiger sind als Zeichen-Wort-N-Gramme. Dieses Ergebnis gilt es in der Bewertung der Merkmale zu prüfen.

Eine wichtige Frage bei der Verwendung von N-Grammen ist die Definition von N. Ein größeres N erfasst eher lexikalische und kontextbasierte Informationen, erhöht aber dabei auch die Dimensionalität des Merkmals (vgl. Tabelle 5). Ein kleines N hingegen ist für die Darstellung der Kontextinformationen nicht geeignet, steigert aber durch die verringerte Anzahl der Dimensionen die Geschwindigkeit des Modells (Stamatatos, 2009, S. 7). Die Bewertung, ob ein Erfassen der Kontextinformationen positiv oder negativ ist, bleibt an dieser Stelle noch aus. Sapkota, Bethard, Montes, & Solorio (2015) erzielten die besten Ergebnisse mit einer Länge von N gleich drei. Die Definition von N ist allerdings sprachenabhängig und ein größeres N erscheint in der deutschen Sprache, welche durchschnittlich längere Wörter als die englische Sprache besitzt, durchaus sinnvoll (vgl. Kapitel 5.6.4). Aus diesem Grund ist in der Datenvorbereitung die optimale Größe von N zu evaluieren. Im weiteren Verlauf werden alle N-Gramme mit einem N von zwei bis fünf betrachtet.

Tabelle 5 zeigt die fünf meist verwendeten N-Gramme und die Gesamtzahl der extrahierten N-Gramme für die einzelnen Superklassen von n=2 bis n=5. Leerzeichen werden für eine bessere Lesbarkeit als Unterstrich („_“) dargestellt. Die Häufigkeit dient hier lediglich als Veranschaulichung der extrahierten N-Gramme und besitzt keine Aussagekraft über den Informationsgehalt. Beispielsweise steht an jedem Ende eines Artikels „Fazit:“, weshalb zwei der häufigsten Zeichen-Interpunktion-5-Gramme „zit:_“ und „azit:“ sind. Da dieses Merkmal aber bei fast allen Texten und Autoren gefunden wird, ist anzunehmen, dass es keine Varianz für einzelne Autoren und dadurch keine Aussagekraft besitzt.

Tabelle 5: Zeichen-N-Gramme - Eigene Darstellung

	2-Gramme	3-Gramme	4-Gramme	5-Gramme
Zeichen-Affix-N-Gramme	_n en d_ _e _r	en_ er_ _de ch_ nd_	_die und_ _ein der_ _und	_und_ _und_ _die_ _die_ _der_
Zeichen-Wort-N-Gramme	Ch er in re ei	n_d und ich die sch	en_d en_s isch en_a sche	Nicht n_die ische n_und en_un
Zeichen-Wort-N-Gramme	,_ ._ "_ :_ t:	,_d n,_ n._ t,_ t._	en,_ en._ ,_da ,_di ,_de	,_die ,_das en,_d zit:_ azit:
Summe Affix	1.881	11.019	39.592	91.968
Summe Wort	2.053	14.766	70.498	203.003
Summe Interpunktion	1.142	11.123	50.764	144.305
Summe Gesamt	5.076	36.908	160.854	439.276

Wie in Tabelle 5 verdeutlicht wird, entstehen gerade für lange N-Gramme eine sehr große Anzahl von Merkmalsausprägungen. Dies steigert den benötigten Rechenaufwand des Modells. Aus diesem Grund ist die Dimensionalität des Merkmales in der Vorverarbeitung zu reduzieren.

5.6.8. Ziffern

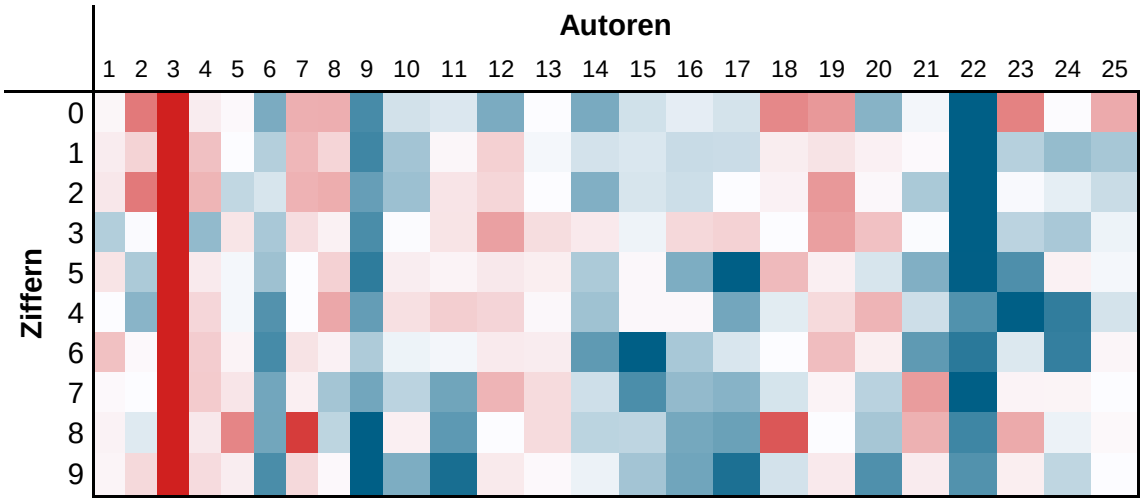
Die Häufigkeit der verwendeten Ziffern von null bis neun wird in der Literatur oftmals als Merkmal angeführt. So wird es in drei der fünf in Verweis erwähnten Arbeiten verwendet. Eine Erklärung für die Verwendung des Merkmals oder einen versprochenen Nutzen führt dabei keiner der Autoren auf. Auch Studien lassen sich diesbezüglich nicht auffinden.

Eine mögliche Erklärung hierfür wäre, dass Zahlen von 1-10 im Englischen üblicherweise ausgeschrieben werden und sich Autoren dahingehend

unterscheiden, inwieweit sie dieser Regelung folgen. In der Literatur wird dabei ausschließlich von Ziffern und nicht von Zahlen gesprochen.

Tabelle 6 beschreibt die verwendeten Ziffern der Autoren in Form einer Hitzekarte. Hohe Werte werden rot, tiefe Werte blau und die Mitte weiß dargestellt. Tatsächlich lässt sich eine gewisse Aussagekraft der Ziffern vermuten. Autor 3 verwendet durchgängig jede Ziffer am häufigsten, wohingegen Autor 9 und 22 sehr wenige Ziffern zu verwenden scheinen. Erkennen lässt sich auch eine Korrelation zwischen der Verwendung der einzelnen Ziffern. Augenscheinlich tendieren Autoren dazu eher alle Ziffern besonders häufig oder selten zu verwenden. Ein häufiges Verwenden einer Ziffer und das seltene einer anderen ist kaum zu erkennen. Aus diesem Grund gilt es in der Datenvorbereitung zu prüfen, ob ein informationeller Mehrwert dadurch entsteht, alle Ziffern auf ihre Häufigkeit zu prüfen oder ob es ausreicht, lediglich das Verwenden von Ziffern auszuwerten.

Tabelle 6: Hitzekarte Ziffern - Eigene Darstellung



5.6.9. Funktionswörter

Die Anzahl der verwendeten Funktionswörter gehört zu den meist genutzten Merkmalen in der Autorschaftsanalyse. Durch ihre hohe Häufigkeit sind Funktionswörter oftmals Teil der Stoppwörter und werden in anderen Anwendungsgebieten meist durch die Normalisierung eliminiert. In der

Autorschaftsanalyse hingegen werden sie als unerlässlich angesehen, da die unbewusste Verwendung und Inhaltsunabhängigkeit entscheidende Merkmale darstellen (Kaster, Siersdorfer, & Weikum, 2005, S. 28). Koppel, Akiva, & Dagan (2006) führen ihnen ebenfalls eine hohe Bedeutung zu, und bezeichnen die verwendeten Varianten als eine starke Reflektion stilistischer Entscheidungen.

Die verwendeten Funktionswörter unterscheiden sich dabei stark zwischen den verschiedenen Arbeiten. Bereits innerhalb der englischen Sprache gibt es hier deutliche Unterschiede (Stamatatos, 2009, S. 542):

- Abbasi & Chen (2005): 150 Wörter
- Argamon, Saric, & Stein (2003): 303 Wörter
- Zhao & Zobel (2005): 365 Wörter
- Koppel & Schler (2003): 480 Wörter
- Argamon, et al. (2007): 675 Wörter

In dieser Arbeit werden alle Funktionswörter extrahiert, welche den in Kapitel 4.3.2 beschriebenen Gruppen angehören. Dies entspricht 587 Wörtern. Tabelle 7 zeigt die fünf meist verwendeten Funktionswörter jedes Autoren. Dabei zeigt sich, dass die Funktionswörter durch einzelne Wörter dominiert werden, welche bei allen Autoren auf demselben Rang stehen. Die einzige Individualität zeigt Autor 16 mit dem Wort „man“.

Tabelle 7: Top 5 Funktionswörter der Autoren - Eigene Darstellung

Autor	Top 5 Funktionswörter	Autor	Top 5 Funktionswörter
1	der, und, sein, einen, mit	**14**	der, und, sein, mit, in
2	der, und, sich, in, einen	**15**	der, und, sein, zu, sich
3	der, und, ich, in, sein	**16**	der, und, sein, einen, man
4	der, und, sein, mit, in	**17**	der, und, sein, sich, in
5	der, und, sein, ich, einen	**18**	der, und, sein, einen, in
6	der, und, sich, ich, in	**19**	der, und, ich, sein, einen
7	der, und, mit, sein, in	**20**	der, und, sein, sich, nicht
8	der, und, sein, ich, zu	**21**	der, und, ich, sein, sich
9	der, und, sich, sein, einen	**22**	der, und, einen, zu, ich
10	der, und, sein, sich, ich	**23**	der, und, sein, einen, sich
11	der, und, sich, zu, mit	**24**	der, und, mit, einen, sich
12	der, und, sein, einen, mit	**25**	der, und, sein, einen, ich
13	der, und, sein, ich, mit		

In Anbetracht der Ähnlichkeit der meist verwendeten Funktionswörter besteht die Vermutung, dass diese keine Aussage über den Autoren treffen können. Innerhalb der Datenvorbereitung gilt es deshalb, eine Selektion der relevanten Funktionswörter durchzuführen.

5.6.10. Wortarten

Grundlage der syntaktischen Konstruktion eines Textes bilden die verwendeten Wortarten. Wortarten beschreiben die Art eines Wortes in Abhängigkeit zu einer grammatikalischen Funktion oder dem Kontext (vgl. Kapitel 4.2). Die grundlegende Annahme nach Stamatatos (2009) ist dabei, dass die Autoren unbewusst ähnliche syntaktische Muster verwenden. Demzufolge betrachtet er die syntaktische Struktur als zuverlässiger als lexikalische Informationen.

Dem entgegen steht eine Studie von Koppel, Argamon, & Shimoni (2003), in welcher sie als Merkmale 405 gemeinsame Funktionswörter Wortarten und Wortarten-N-Gramme verwendeten. Bei einem Korpus von 566 Belletristik- und Sachbuchdokumenten, die zu gleichen Teilen in Geschlechtergruppen verteilt sind, schnitt der Algorithmus des maschinellen Lernens bei der Klassifizierung der Texte nach Genre und nicht nach Geschlecht wesentlich besser ab. Nach der Kontrolle des Genres erreichte der Klassifizierungsalgorithmus für das Geschlecht eine Genauigkeit von etwa 80%. Bestätigende Ergebnisse lieferte eine ähnliche Studie, in der Argamon, Fine, & Shimoni (2003) eine analoge Stichprobe, die aus 246 fiktionalen und 358 nicht-fiktionalen Dokumenten bestand, analysierten. Die Autoren der Dokumente sind jeweils zu gleichen Teilen Männer und Frauen. Ihr Ziel war es zu überprüfen, ob Geschlechtsunterschiede auch in formalen schriftlichen Dokumenten zu finden sind. Sie verwendeten ebenfalls Funktionswörter, Wortarten und Wortarten-N-Gramme. Weiterhin führen sie an, dass Männer sich als informationsreicher zeigten und daher mehr Determinanten und Präpositionen verwendeten, während Frauen stärker involviert waren und daher mehr Pronomen verwendeten (Nini, 2015, S. 30). Schmid (2002, S. 218) spricht hier von einer

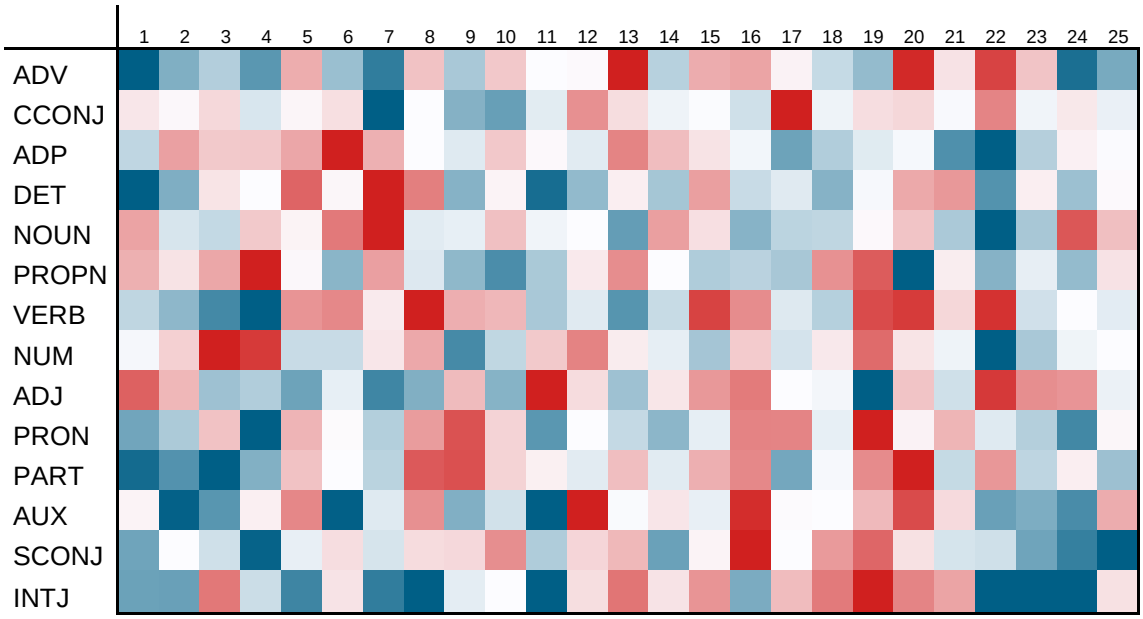
geringeren emotionalen Distanz des weiblichen Geschlechts. Pennebaker & Stone (2003) erkennen eine weitere Korrelation zwischen der Anzahl der verwendeten Personalpronomen und des Alters. Sie beschreiben, dass mit zunehmendem Alter die Verwendung dieser abnimmt.

Daraus folgt zum einen, dass syntaktische Merkmale wie bisher angenommen nicht unbedingt inhaltsunabhängig sind und zum anderen, dass den Wortarten Korrelationen zu verschiedenen Eigenschaften der Autoren zugeschrieben werden.

van Halteren, Tweedie, & Bayeen (1996) verwendeten als erste syntaktische Informationen in der Autorenschafts-Attribution. Sie stellten jeden Satz als einen Syntaxbaum dar, wodurch sie die Häufigkeiten von „Neuschreiberegeln“ erfassen konnten. Eine solche Regel wird definiert durch $A \rightarrow X$, wobei A eine syntaktische Kategoriebezeichnung wie ein Satz ist und X eine Folge solcher Bezeichnungen oder Morpheme ist, die die Tatsache ausdrückt, dass A bei der Erzeugung der konstituierenden Struktur eines Satzes durch X ersetzt werden kann (Collins English Dictionary, 2020). Ein weniger komplexer Ansatz ist das Verwenden eines PoS-Taggers, welcher die Wortarten der einzelnen Wörter in einem Satz bestimmt (vgl. Kapitel 4.2). Dieser Ansatz wird als vielversprechender empfunden, da heutige Methoden sehr präzise sind und weniger Fehlerpotential aufweisen (Stamatatos, 2009, S. 8). In dieser Arbeit wird der PoS-Tagger von SpaCy verwendet (vgl. Kapitel 4.2). Da eine Auswertung der Sonderzeichen gesondert erfolgt, werden Sonderzeichen, Leerzeichen und durch die Bibliothek nicht definierbare Wortarten ausgeschlossen.

Tabelle 8 beschreibt die verwendeten Wortarten der Autoren in Form einer Hitzekarte. Hohe Werte werden rot, tiefe Werte blau und die Mitte weiß dargestellt. Es zeigt sich, dass einige Autoren einige Wortgruppen besonders oft oder selten verwenden. Dabei scheint sich das Muster keiner Autoren zu gleichen. Dies spricht zum einen für eine Varianz und zum anderen für eine Individualität des Merkmals.

Tabelle 8: Hitzekarte Wortarten - Eigene Darstellung



5.6.11. Wortarten-N-Gramme

Wie bereits im vorherigen Kapitel erläutert, werden Wortarten extrahiert, um die syntaktische Struktur des Textes zu erfassen. In der Grammatik spielen allerdings auch Kombinationen von Wortgruppen, sogenannte Phrasen, eine wichtige Rolle. Diese beschreiben die wesentliche Struktur eines Satzes. Zum Ermitteln dieser Phrasen werden in der Linguistik syntaktische Tests, sogenannte Konstituententests, durchgeführt (Imo, 2016, S. 116). In der Autorschaftsanalyse kann stattdessen, um die Komplexität möglichst gering zu halten, auf N-Gramme zurückgegriffen werden. Diese erfassen alle Phrasen eines Textes und sollen so Unterschiede zwischen Autoren zeigen.

In der Literatur wird für N meist eine Länge von zwei bis drei gewählt, da mit steigender Länge die möglichen Merkmale um ein Vielfaches erhöhen (vgl. Tabelle 9). Es gilt zu evaluieren, ob mit einem höheren N bessere Ergebnisse erzielt werden können und ob dies den steigenden Rechenaufwand des Modells rechtfertigt.

Tabelle 9 zeigt die meist verwendeten 2 bis 5-Gramme. Die Bezeichnung der Wortarten richtet sich dabei nach der in Kapitel 4.2 beschriebenen Annotation

von spaCy (Explosion, 2020a). Auffällig sind lange Ketten von Eigennamen (engl. „proper noun“), welche durch lange Band- oder Albennamen entstehen. Zu den häufigsten Kombinationen gehören klassische Artikel-Nomen oder Adjektiv-Nomen Kombinationen, aber auch längere Kombinationen wie Nomen-Adposition-Artikel-Adjektiv-Nomen sind häufig zu finden, wie zum Beispiel „einen Schlussstrich unter einen selbstzerstörerischen Lebensabschnitt“.

Tabelle 9: Wortarten-N-Gramme - Eigene Darstellung

	Top 5 N-Gramme	**Vorhandene N-Gramme**	**Mögliche N-Gramme**
2-Gramm	DET, NOUN ADJ, NOUN PROPN, PROPN NOUN, ADP NOUN, VERB	191	196
3-Gramm	DET, ADJ, NOUN ADP,DET,NOUN PROPN,PROPN,PROPN DET,NOUN,ADP ADJ,NOUN,ADP	2158	2744
4-Gramm	ADP, DET, ADJ, NOUN PROPN, PROPN, PROPN, PROPN NOUN, ADP, DET, NOUN DET, ADJ, NOUN, ADP DET, NOUN, DET, NOUN	18070	38416
5-Gramm	PROPN, PROPN, PROPN, PROPN, PROPN DET, NOUN,ADP,DET,NOUN NOUN,ADP,DET,ADJ,NOUN ADJ,NOUN,ADP,DET,NOUN DET,ADJ,NOUN,DET,NOUN	88839	537824

Die theoretische Anzahl der möglichen N-Gramme erhöht sich mit einem höheren N exponentiell mit $\mathrm{Wortarten}^N$. In der Praxis wird diese Anzahl durch untypische Wortkombinationen zwar nicht erreicht werden, allerdings ist die Anzahl der vorhandenen Dimensionen des Merkmales dennoch sehr groß. Wie bereits im vorherigen Kapitel angeführt, sollen deshalb in der Datenvorbereitung irrelevante Merkmale ausgeschlossen und somit die Anzahl der Merkmale reduziert werden.

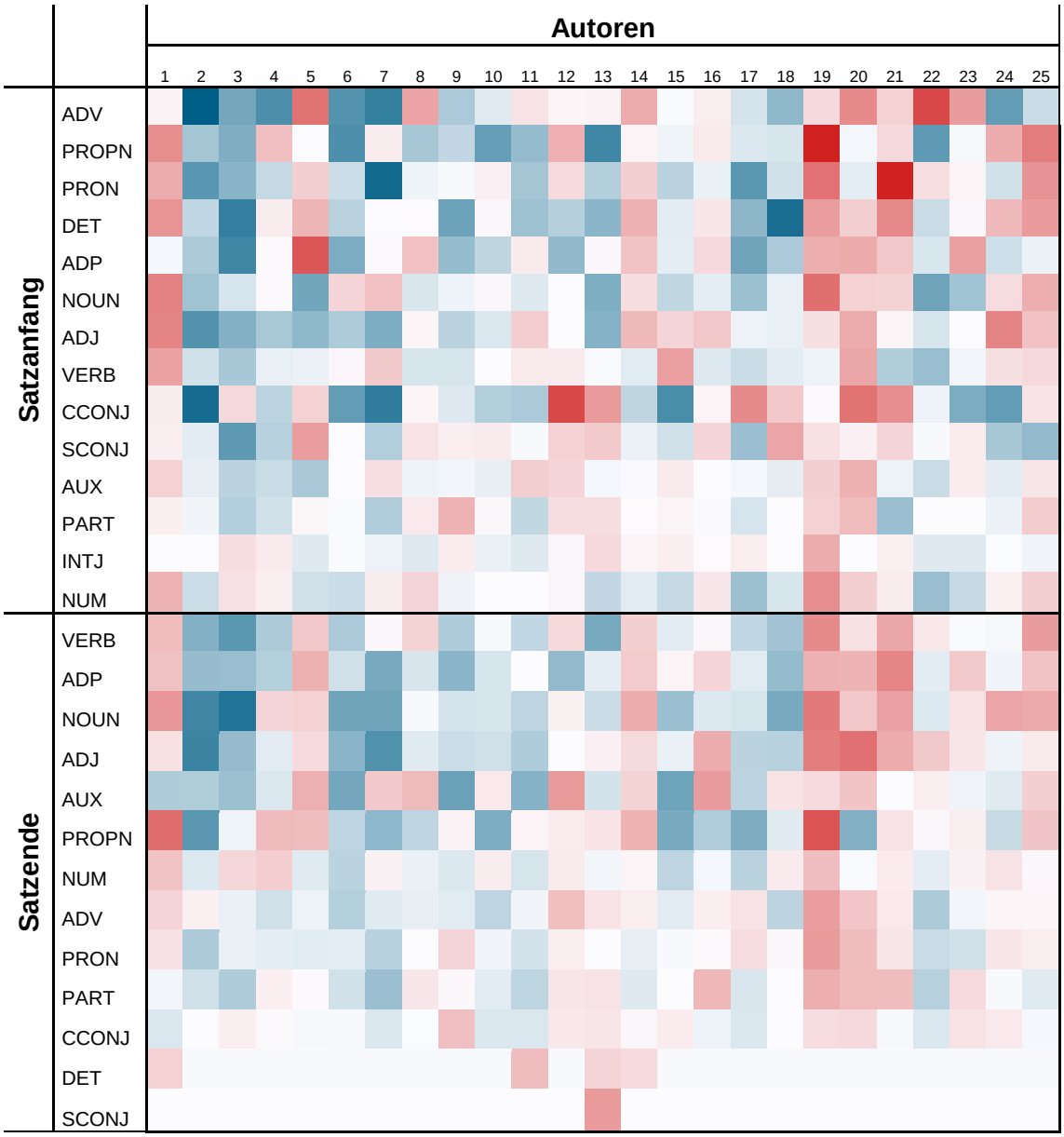
5.6.12. Satz-Anfänge/-Endungen

Die englische Sprache weist die Wortstellung Subjekt – Prädikat – Objekt auf. Die einzige Ausnahme sind Entscheidungsfragesätze, in welchen der Prädikatsteil vor dem Subjekt steht. Im Gegensatz dazu gibt es im deutschen keine feste Abfolge von Satzgliedern, sondern Satzgrundmuster, welche durch die Stellung des Verbs gekennzeichnet werden (Imo, 2016, S. 199). Für die korrekte Anordnung dieser Muster existieren grammatikalische Regeln, welche zu Fehlern führen können. Die Komplexität der deutschen gegenüber der englischen Grammatik zeigt sich gleichzeitig als vermutete Begründung, warum das Merkmal in englischen Autorschaftsattribution nicht verwendet wird. Für den Erfolg des Merkmales wird sich versprochen, dass die Wahl der Satzkonstruktion eine individuell stilistische Entscheidung eines Autors ist. Weiterhin wird davon ausgegangen, dass diese stilistische Wahl durch das Erfassen der Wortart des ersten und des letzten Wortes eines Satzes abgebildet werden kann.

Rexha, Kröll, Ziak, & Kern (2018) beschreiben, dass die ersten Wörter eines Satzes Aussagen über den Schreibstil eines Autors treffen lassen. So verwenden manche Autoren Phrasen wie „allerdings“, „obgleich“, „weiterhin“, „demnach“ oder viele mehr um zwei Sätze miteinander zu verbinden. Da dies einen Satz mit beginnendem Adverb zufolge hat, könnten auch so der Stil eines Autors erfasst werden.

Tabelle 10 beschreibt die verwendeten Wortarten der Satzanfänge und Satzendungen in Form einer Hitzekarte. Hohe Werte werden rot, tiefe Werte blau und die Mitte weiß dargestellt. Auch hier werden Sonderzeichen, Leerzeichen und Tags, welche nicht definiert werden können ausgeschlossen. Die Bezeichnung der Wortarten richtet sich dabei nach der in Kapitel 4.2 beschriebenen Annotation von spaCy (Explosion, 2020a).

Tabelle 10: Wortarten Satzanfänge und –endungen - Eigene Darstellung



Offensichtlich existieren einige Eigenheiten der Autoren. Beispielsweise verwendet Autor 19 häufig Eigenenamen (PROPN) und Autor 21 Substantive (PRON) als Satzanfang, wobei Autor 21 auch gerne mit diesen endet. Eine besondere Eigenheit kann bei Autor 12 erkannt werden. Dieser verwendet häufig nebengeordneten Konjunktionen wie „und“ als Satzanfang, anstatt die Sätze mit diesen zu verbinden, wohingegen viele andere Autoren stark davon absehen. Untergeordnete Konjunktionen am Ende eines Satzes, wie sie Autor 13 verwendet, kommen durch eine rhetorische Frage an den Leser durch einen

angefügten Nebensatz am Ende zustande, beispielsweise „…, oder?“. Insgesamt scheint das Merkmal einige stilistischen Variationen der Autoren wiederzugeben.

5.6.13. Verständlichkeitsindizes

Die bereits in Kapitel 5.6.4 erwähnte Forschung von DuBay (2004) zeigt, dass bereits die US-Armee bei ihren Forschungen feststellte, dass die Messungen der Lese- und Schreibfähigkeit und der Lesbarkeit mit der Intelligenz, dem Wissen und der erreichten Ausbildungsdauer korrelierten. Die bereits angeführte Studie von Kitson (1921) zeigte, dass signifikante Unterschiede bezüglich der durchschnittlichen Satzlänge und der Wortlänge zwischen Zeitschriften in Abhängigkeit zu ihrer Leserschaft bestehen und damit die Art der sozialen Gruppen beschreiben, die eine bestimmte Zeitschrift lesen würden. Dies folgte in einigen Studien der US-Armee und der Psychologie, die darauf abzielten, eine Lesbarkeitsformel zu finden, die die Lesbarkeit eines Textes vorhersagen konnte. Mit der Zeit nahm die Komplexität der Formeln zu, obwohl das Hinzufügen weiterer Variablen ihre Effizienz nur geringfügig erhöhte. Im Allgemeinen stellten die Forscher jedoch fest, dass die Anzahl der Morpheme und die Anzahl der syntaktischen Verzweigungen die nützlichen Variablen sind, um die semantische (lexikalische) Komplexität bzw. die syntaktische Komplexität zu messen und eine Schätzung der Gesamtkomplexität eines Textes zu geben. Da die Anzahl der Morpheme stark mit der Anzahl der Silben und die Anzahl der syntaktischen Verzweigungen stark mit der Satzlänge korreliert, war die erfolgreichste Lesbarkeitsformel letztendlich nur eine Kombination von Silben- und Wortzählungen. Die bekannteste dieser Formeln ist wahrscheinlich der FRE (Flesch Reading Ease Score), die als Funktion der Wort- und Satzlänge berechnet wird. (Nini, 2015, S. 60)

Die deutsche Sprache besitzt durchschnittlich längere Worte, dabei aber eine vergleichbare Satzlänge, weshalb Amstad (1978) die Faktoren des von Flesch (1948) entwickelten Verfahrens anpassen musste. Er definierte die Formel folgendermaßen:

$$FRE = 180 - ASL - (58{,}5 * ASW) \quad (5.6.13.1)$$

In der Formel steht „ASL“ für die durchschnittliche Satzlänge und „ASW“ für die durchschnittliche Silbenzahl der Wörter. Das Maximum des Wertes liegt bei 121,22 und ein Minimum existiert nicht. Negative Werte sind möglich. Dabei lässt sich der errechnete Wert durch die folgende Tabelle übersetzen:

Tabelle 11: Übersetzung FRE - Eigene Darstellung - Anlehnung an (DuBay, 2006)

FRE	Lesbarkeit	Verständlich für
0–30	Sehr schwer	Hochschulabsolvent
30–49	Schwer	
50–59	Mittelschwer	
60–69	Standard	12–14-jährige Schüler
70–79	Mittelleicht	
80–89	Leicht	
90–100	Sehr leicht	10-jährige Schüler

Bezogen auf die Datengrundlage der Arbeit zeigt sich ein geringer Gesamtdurchschnitt des FRE-Wertes. Mit einem Durchschnitt von 41,3 ist ein Großteil der Texte schwer zu lesen. Der Maximalwert aller Artikel liegt bei 71,9, was einer mittelleichten Lesbarkeit entspricht. Ursächlich dafür sind die sehr langen Worten, zu denen einige Autoren neigen (vgl. Kapitel 5.6.4) und die durchschnittlich lange Textlänge. Des Weiteren ist davon auszugehen, dass die Zielgruppe der Musikreviews keine Schüler unter 15 Jahren beinhaltet, wodurch die Autoren nicht auf eine leichte Verständlichkeit ihrer Texte achten müssen.

Dennoch lassen sich Unterschiede zwischen den einzelnen Autoren erkennen, so scheinen die Autoren 2, 3, 7, 9 und 18 die am schwersten zu verstehenden Texte zu schreiben. Autor 20 und 21 sind der mittelschweren Klasse zuzuordnen. Das folgende Diagramm zeigt die durchschnittliche Verteilung über alle Autoren.

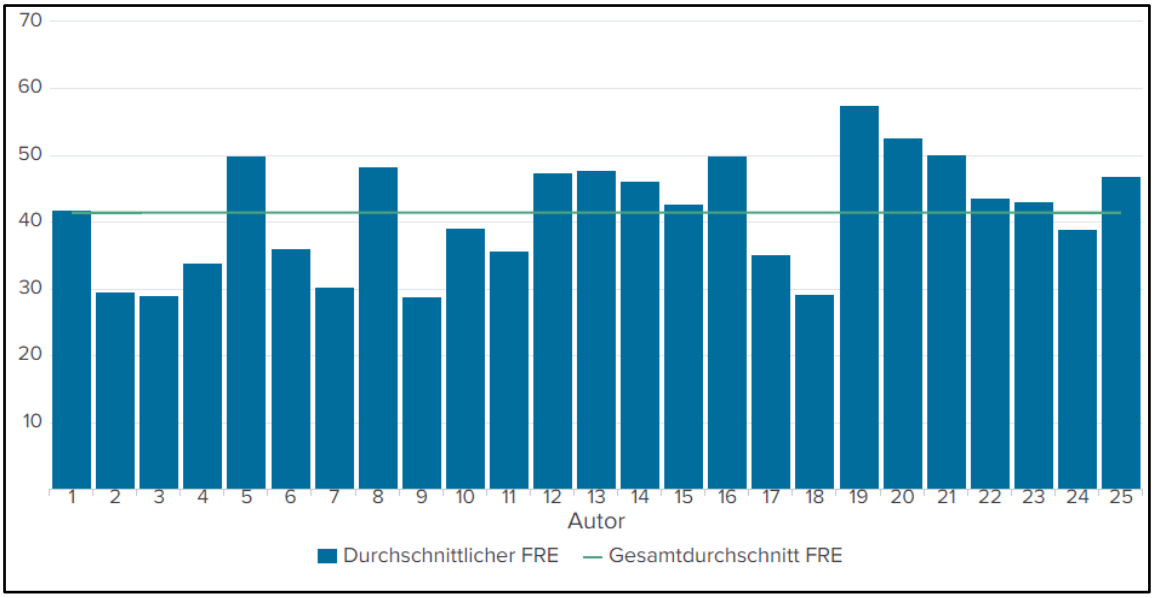


Bild 6: Durchschnittlicher FRE der Autoren - Eigene Darstellung

6. Grundlagen maschinelles Lernen

Die Analysen und Verfahren in der Datenvorbereitung und Modellierung basieren auf Methoden und theoretischen Grundlagen des maschinellen Lernens. Deshalb werden in diesem Kapitel einige dieser Grundlagen vorgestellt.

6.1. Verwendete Bibliotheken

Im Rahmen dieser Arbeit werden grundlegende Bibliotheken im Anwendungsgebiet des maschinellen Lernens verwendet. Diese werden im Folgenden erläutert.

Nach der offiziellen Dokumentation ist NumPy das grundlegende Paket für wissenschaftliches Rechnen und damit maschinelles Lernen in Python (The SciPy community, 2020). Es ist eine Python-Bibliothek, die ein mehrdimensionales Array-Objekt, verschiedene abgeleitete Objekte (z. B. maskierte Arrays und Matrizen) und eine Reihe von Routinen für schnelle Operationen auf Arrays bereitstellt, darunter mathematische, logische, Formmanipulationen, Sortieren, grundlegende lineare Algebra und grundlegende statistische Operationen (ebd.).

Pandas ist ein Python-Paket mit Datenstrukturen, welche die die Arbeit mit "relationalen" oder "beschrifteten" Daten erleichtert. Durch Pandas wird unter anderem eine einfache Lesbarkeit und Analyse der Datengrundlage in Form einer CSV-Datei ermöglicht. Pandas basiert auf NumPy und ermöglicht die Anwendung weitergehender Funktionen auf die enthaltenen Daten. (pandas development team, 2020)

Scikit-learn ist eine open-source Bibliothek für überwachtes und unüberwachtes Lernen. Weiterhin beinhaltet sie einige Werkzeuge für die Modellierung und die Datenverarbeitung (scikit-learn developers, 2020h). Nach Liu (2019, S. 41) handelt es sich dabei um eine der am besten entworfenen Bibliotheken für maschinelles Lernen. Bekannt ist sie durch die schnellen Laufzeiten des Codes (ebd.).

Die Python-Bibliothek matplotlib wird verwendet, um Graphen, Diagramme und Figuren zu erstellen. Die Bibliothek ist umfangreich und in der Lage, kleine Details einer Figur zu verändern (Edpresso Team, 2020). In dieser Arbeit wird die Bibliothek verwendet, um grundsätzliche Graphen innerhalb der Entwicklung zu erstellen. Beispielhafte Grafiken, welche in der Arbeit aufgezeigt werden, unterliegen grundsätzlich einer weiteren Aufbereitung.

6.2. Gruppierung von Funktionen

Die Hauptaufgabe von maschinellem Lernen ist das Entdecken und Konstruieren von Algorithmen, welche es erlauben, aufgrund von historischen Daten Aussagen oder Vorhersagen über zukünftige Daten zu treffen. Für eine datengesteuerte Lösung muss eine Bewertungsfunktion definiert werden, die als Verlust- oder Kostenfunktion bezeichnet wird und misst, wie gut die Modelle lernen. Die Definition dieser Funktion kann gleichermaßen durch einen Algorithmus erfolgen. (Liu Y. H., 2019, S. 14)

Die grundlegendenden Ansätze für maschinelles Lernen sind überwachtes (engl. supervised), unüberwachtes (engl unsupervised), bestärkendes (engl. reinforcement) und mehrschichtiges (engl. deep) Lernen (engl. learning).[11]

Supervised learning wird in der Regel im Kontext der Klassifikation, dem Zuordnen von Eingabedaten zu einer festgelegten Kategorie oder in der Regression, dem Prognostizieren von abhängigen Variablen verwendet. Bei den Eingabedaten handelt es sich um beschriftete Daten, da sie neben den indikativen Signalen eine Beschreibung, Ziele oder eine gewünschte Ausgabe besitzen. Die Funktion hat zum Ziel, unbeschriftete Daten einer solchen Beschreibung zuzuordnen (ebd.). Durch einen Lehrer oder Leiter (engl. supervisor) wird eine präzise Messung des Fehlers zurückgegeben, welche direkt mit dem bekannten Ziel verglichen werden kann. Mit dieser Information können die Parameter so angepasst werden, dass das Ausmaß der globalen

[11] Im Jargon des maschinellen Lernens werden üblicherweise die englischen Begriffe verwendet. Auch in dieser Arbeit werden deshalb die englischen Begriffe den deutschen vorgezogen.

Verlustfunktion minimal wird. Jede Iteration präzisiert die Funktion. (Bonaccorso, 2017, S. 10)

Im unsupervised learning besteht die Eingabe nur aus indikativen Signalen, ohne einer Beschreibung oder einem Ziel und besitzt demzufolge keinen Leiter oder absolute Fehlermessungen. Die Funktion wird verwendet, um die Daten nach einem statistischen oder geometrischen Ähnlichkeitskriterium zu untersuchen. Folglich wird eine inhärente Struktur der Daten gesucht. Hierdurch können auch von der erwarteten Struktur abweichende Daten (Anomalien) gefunden werden. (Igual & Seguí, 2017, S. 67)

Reinforcement learning basiert wie das supervised learning auf Rückgabewerten. Es unterscheidet sich dadurch, dass dieses Feedback nicht durch einen Leiter erfolgt und es sich nicht um präzise Fehlermessungen handelt. Das Feedback ist qualitativ und wird als „Belohnung“ (engl. reward) bzw. „Strafe“ (engl. penalty) definiert. Die Aufgabe des Feedbacks ist zu verstehen, ob eine bestimmte Handlung in einem Zustand positiv oder negativ ist. Die Sequenz mit den meisten nützlichen Handlungen ist die, welche das System lernen muss. Der Ansatz basiert darauf, dass das rationale System immer die Entscheidung mit der größten Belohnung, bzw. der geringsten Strafe trifft. Reinforcement learning ist besonders effizient, wenn die Umgebung nicht deterministisch oder dynamisch ist, und wenn es nicht möglich ist eine präzise Messung des Fehlers vorzunehmen. (Bonaccorso, 2017, S. 14)

Deep learning beschäftigt sich mit Problemen, welche für den Menschen einfach durchzuführen sind, sich aber schwer mit mathematischen Modellen beschreiben lassen, wie etwa dem Erkennen von menschlicher Sprache. Als Lösung für diese intuitiven Probleme soll es Computern ermöglicht werden, aus Erfahrungen zu lernen und komplexe Sachverhalte in einer Hierarchie von Konzepten darzustellen. Die Struktur entsteht dadurch, dass jedes Konzept durch seine Beziehung zu einfacheren Konzepten definiert wird. Mittels Sammeln von Erfahrungswissen vermeidet dieser Ansatz die Notwendigkeit, das gesamte vom Computer benötigte Wissen durch menschliche Operatoren formal zu spezifizieren. Graphisch dargestellt resultiert daraus ein tiefer Graph mit vielen

Schichten. Jede Schicht leitet ihre Ausgaben an die nächste Schicht weiter. Aus diesem Grund wird dieser Ansatz auch „mehrschichtiges“ oder „tiefes“ Lernen genannt. (Goodfellow , Bengio , & Courville, 2016)

In einem künstlichen neuronalen Netz bestehen die Schichten aus mehreren Knoten, den sogenannten Neuronen. Diese können Daten von außen oder von anderen Neuronen aufnehmen sowie diese an andere weiterleiten oder als Endergebnis ausgeben. Die Verbindung der Neuronen erfolgt über sogenannte Kanten. Jede Kante besitzt abhängig von der Bedeutung des Inputs eine Gewichtung, welche im Lernprozess bei mehreren Iterationen angepasst werden kann. (Luber & Litzel, 2018)

Bei der Frage nach der Autorschaft stellen Texte mit bekanntem Autor die beschrifteten Daten dar. Demnach ist es möglich, eine präzise Messung eines Fehlers vorzunehmen und ein Modell zu erstellen, welches eine Zuordnung eines Autors bei unbeschrifteten Daten (Texte ohne bekanntem Autor) vornimmt. Metriken, um diese Fehler zu messen, werden im weiteren Verlauf der Arbeit vorgestellt. Die Autorschaftsattribution lässt sich demnach am besten dem Ansatz des supervised learnings zuordnen. Es existieren auch Ansätze im Bereich des deep learnings. Da Anwendungen in diesem Bereich deutlich komplexer aufgebaut sind und die Transparenz des Verfahrens in dieser Arbeit als unerlässlich angesehen wird, werden diese Modelle im Weiteren nicht betrachtet.

6.3. Mathematische Grundlagen des überwachten Lernens

Ein jeder der in kommenden Kapiteln vorgestellten Algorithmen besitzt grundsätzlich dieselbe Basis. Diese Grundlagen werden anlehnend an Bonaccorso (2017, S. 20) im Folgenden dargestellt.

Der Ausgang ist immer ein Datensatz, definiert als eine endliche Menge von Vektoren mit jeweils m Merkmalen.

$$X = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n\} \, wobei \, \bar{x}_i \in \mathbb{R}^m \tag{6.3.1}$$

Ein jeder Vektor wird durch seine Merkmale dargestellt. Im Rahmen dieser Arbeit entspricht dies den in Kapitel 5.6 vorgestellten Merkmalen eines Textes.

$$\bar{x}_i = [x_1, x_2, \dots, x_m] \tag{6.3.2}$$

Unter der Annahme, dass der Ansatz immer wahrscheinlichkeitstheoretisch ist, muss jedes X als aus einer statistisch multivariaten Distribution D abgeleitet betrachtet werden. Eine weitere Bedingung ist, dass alle Stichproben des Datensatzes X unabhängig und identisch verteilt sind. Demzufolge gehören alle derselben Verteilung D an, wodurch für eine willkürliche Untermenge von m resultiert, dass die Gesamtwahrscheinlichkeit durch das Produkt der einzelnen Vektor-Wahrscheinlichkeiten ermittelt werden kann.

$$P(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m) = \prod_{i=1}^{m} P(\bar{x}_i) \tag{6.3.3}$$

Die entsprechenden Ausgabewerte können sowohl numerisch, als auch kategorial sein. Im ersten Fall wird der Prozess als Regression bezeichnet, im zweiten Fall als Klassifikation. Numerische Werte können einer positiven reellen oder einer Zahl zwischen null und eins entsprechen, wohingegen kategoriale Ausgaben beispielsweise Werte wie Marken oder Autorennamen annehmen können.

Im Falle der Regression wird eine vektoriell bewertete Funktion als generischer Regressor r definiert, die einen Eingabewert mit einer kontinuierlichen Ausgabe verknüpft. Bei der Klassifikation handelt es sich um eine Vektorwertfunktion als Klassifikator (c), deren vorausgesagte Ausgabe kategorial ist. Besteht eine Abhängigkeit zu einem internen Parametervektor ($\bar{\theta}$), der die tatsächliche Instanz eines generischen Prädiktors bestimmt, wird der Ansatz parametrisches Lernen genannt.

$$\tilde{y} = r(\bar{x}, \bar{\theta}) \tag{6.3.4}$$

$$\tilde{y} = c(\bar{x}, \bar{\theta}) \tag{6.3.5}$$

Die interne Dynamik und die Interpretation aller Elemente sind jedem Algorithmus eigen. Ein generischer parametrischer Trainingsprozess muss den besten Parametervektor finden, der den Regressions-/Klassifikationsfehler bei einem bestimmten Trainingsdatensatz minimiert. Außerdem sollte er auch einen Prädiktor[12] generieren, der korrekt verallgemeinern kann, wenn unbekannte Stichproben zur Verfügung gestellt werden. Ein in der Literatur allgemein beschriebener Ansatz ist es, diese Fehler zu messen und zu minimieren. In diesem Zusammenhang wird auch von der Fehlerfunktion gesprochen (Liu Y. H., 2019, S. 17). Eine andere Interpretation drückt die Fehler in Form von additivem „Rauschen“ aus. Dieses Rauschen wird näher in Kapitel 6.6 erläutert.

6.4. Multi-Label-Klassifikation

Im vorherigen Kapitel bestand die Annahme, dass das Ergebnis einer Klassifikation aus einer einzigen Klasse, beziehungsweise der Entscheidung „zugehörig oder nicht zugehörig zur Klasse“ ($y(i) \in \{0,1\}$) besteht.

Diese Art der Klassifikation wird als binäre Klassifikation bezeichnet. Besteht das mögliche Ergebnis der Klassifikation aus mehreren Klassen, wird diese als Multi-Label-Klassifikation bezeichnet. ($y(i) \in \{1, \dots, K\}$, wobei K für die Anzahl der Klassen steht) Um eine Multi-Label-Klassifikation zu lösen, werden diese mithilfe von heuristischen Methoden in mehrere binäre Klassifikationen unterteilt. Beispiele für diese Methoden sind OvR (One-vs-Rest) oder OvO (One-vs-One) (Brownlee, 2020).

Die Annahme ist, dass vier mögliche Autoren für einen Text existieren: $y(i) \in \{A1,\ A2, A3, A4\}$. OvR unterteilt dieses Problem in mehrere binäre Klassifizierungsprobleme, in welchen eine Klasse gegen alle anderen Klassen vorhergesagt wird:

1. Klassifizierungsproblem: A1 gegen [A2, A3, A4]
2. Klassifizierungsproblem: A2 gegen [A1, A3, A4]

[12] Variable, die benutzt wird, um die Werte einer anderen Variable vorherzusagen

3. Klassifizierungsproblem: A3 gegen [A1, A2, A4]
4. Klassifizierungsproblem: A4 gegen [A1, A2, A3]

Als Folge muss für jede Klasse ein Model erstellt werden, was abhängig von der Anzahl der Klassen in einem hohen Rechenaufwand resultieren kann. Vorrausetzung für diesen Ansatz ist, dass jedes Model einen Wahrscheinlichkeitswert für die Zugehehörigkeit ausgibt. Die Klasse mit dem höchsten Wert wird dann von der Strategie ausgewählt. (Murphy, 2012)

Bei OvO wird jede gegen jede Klasse geprüft. Daraus folgt eine Vielzahl an Modellen. Die Anzahl wird mit der Formel $\left(K \times (K-1)\right) \div 2$ berechnet.

1. Klassifizierungsproblem: A1 gegen A2
2. Klassifizierungsproblem: A1 gegen A3
3. Klassifizierungsproblem: A1 gegen A4
4. Klassifizierungsproblem: A2 gegen A3
5. Klassifizierungsproblem: A2 gegen A4
6. Klassifizierungsproblem: A3 gegen A4

Jedes Model wird für eine Klasse stimmen. Die mit den meisten Stimmen wird von der OvO-Strategie vorhergesagt (Bishop, 2006, S. 183).

6.5. Klassifikationsalgorithmen

Im Folgenden werden drei Klassifikationsalgorithmen vorgestellt. KNN (K-Nearest-Neighbor) als parametrisches Modell, Naïve Bayes als probabilistisches Modell und SVM (Support Vector Machines) als Breiter-Rand-Klassifikator (engl. Large Margin Classifer).

6.5.1. K-Nearest-Neighbor

Algorithmen für maschinelles Lernen können in parametrische und nicht parametrische Modelle gruppiert werden. Parametrische Modelle schätzen Parameter aus dem Trainingsdatensatz, um eine Funktion zu erlernen, die neue Datenpunkte klassifizieren kann ohne dass der ursprüngliche Trainingsdatensatz weiter benötigt wird. Im Gegensatz dazu können nichtparametrische Modelle

nicht durch einen festen Satz von Parametern charakterisiert werden und die Anzahl der Parameter wächst mit den Trainingsdaten. Eine Unterkategorie der nichtparametrischen Modelle sind instanzbasierte[13] Modelle. Sie zeichnen sich dadurch aus, dass sie den Trainingsdatensatz auswendig lernen. „Lazy Learners“ sind ein Sonderfall, welche mit einem Nullaufwand während des Lernens assoziiert werden. KNN ist ein klassisches Beispiel für einen Lazy Learner. (Raschka, 2016, S. 92)

Nach Dinov (2018, S. 267) lässt sich KNN in den folgenden vier Schritten beschreiben:

1. Erstellen eines Trainingsdatensets klassifiziert nach nominalen Variablen und Merkmalen mit ordinalen oder numerischen Variablen.
2. Erstellen eines nicht klassifizierten Testdatensets mit denselben Merkmalen.
3. Vergleich jedes Testdatensatzes mit K Trainingsdatensätzen, welche in ihrer Ähnlichkeit am nächsten liegen.
4. Zuweisung der Klasse, die den Großteil der K Trainingsaufzeichnungen enthält.

Ist die Anzahl der Kategorien in den K nächsten Punkten gleich verteilt können Gewichte, wie beispielsweise die Distanz dieser, festgelegt werden (scikit-learn developers, 2020e).

Eine wichtige Rolle für die Funktion des Algorithmus spielt die Definition von K. Wird ein zu großes K gewählt, wird der Testdatensatz eher der beliebtesten Klasse, als der ähnlichsten Klasse zugeordnet. Ist K dagegen zu klein, erhöht dies die Anfälligkeit für Ausreißer, was zu Fehlern in der Vorhersage führen kann. Die robusteste Methode zum Definieren von K ist es, mehrere Werte auszuprobieren und das K mit der besten Leistung im Training zu wählen. (Dinov, 2018, S. 267)

[13] Nicht zu verwechseln mit dem instanzbasierten Paradigma der Autorschaftsanalyse.

Bild 7 veranschaulicht das Hinzufügen eines neuen Datenpunktes (X) bei K=3. Um die Komplexität möglichst gering zu halten, erfolgt die Darstellung im zweidimensionalen Raum bzw. mit zwei Merkmalen. Die Abstände werden als d1 bis d3 bezeichnet. Die drei nächsten Punkte wären in diesem Fall zwei rote Quadrate und ein grünes Dreieck, was zu einer Klassifikation zu einem roten Quadrat führt.

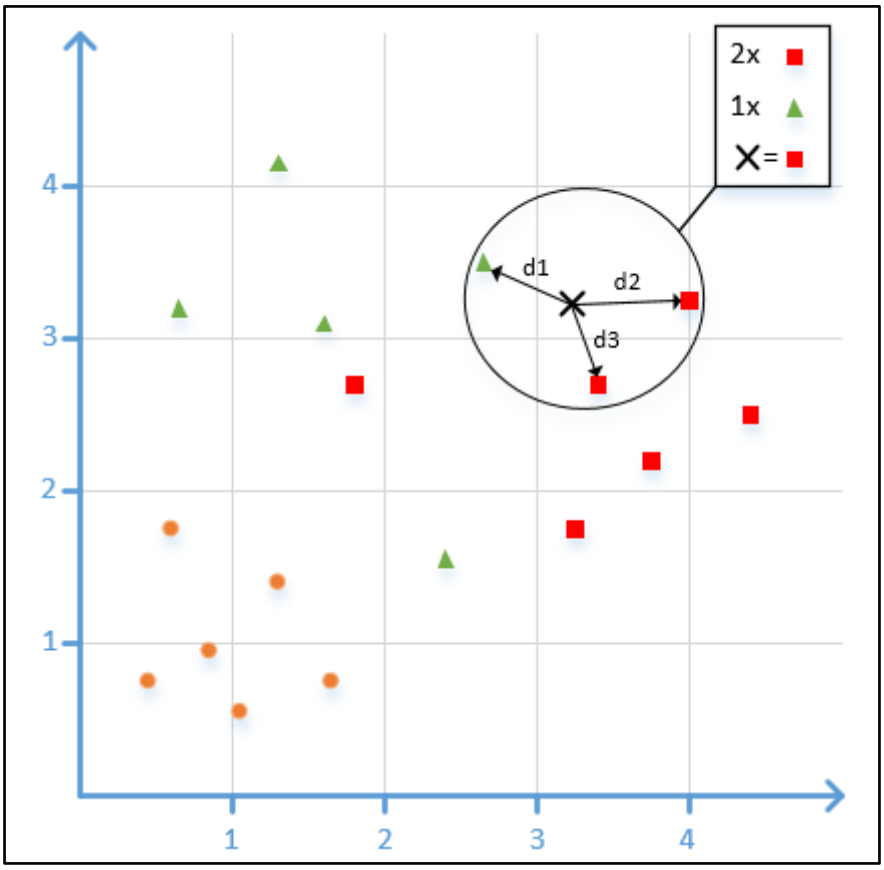


Bild 7: Hinzufügen Datenpunkt KNN - Eigene Darstellung – Anlehnung an (Bonnin, 2017, S. 73)

Das Ermitteln der Distanz kann über verschiedene Funktionen erfolgen. Meist wird hierfür der euklidische Abstand verwendet. Dieser beschreibt den direkten Abstand zweier Punkte im n-dimensionalen Raum. Angenommen „p“ und „q“ haben beide „n“ Merkmale mit den Koordinaten $(p_1, p_{2,} \ldots, p_n)$ und $(q_{1,} q_{2,} \ldots, q_n)$ wird der euklidische Abstand nach (Dinov, 2018, S. 267) definiert durch:

$$d(p,q) = \|p-q\|_2 = \sqrt{(q_1-p_1)^2+\cdots+(q_n-p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2} \qquad (6.5.1.1)$$

Das Modul „KNeighborsClassifier“ von sklearn verwendet die "Minkowski"-Distanz. Diese ist nur eine Verallgemeinerung des euklidischen Abstandes und der Manhattan-Metrik:

$$d(X,Y) = \left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{\frac{1}{p}} \qquad (6.5.1.2)$$

Demnach ist die Gleichung für $\mathrm{p}=1$ äquivalent zur Manhattan Metrik und für $\mathrm{p}=2$ zum euklidischen Abstand (Bonaccorso, 2017, S. 226).

Der Vorteil des KNN liegt in seiner Einfachheit, allein K muss als Parameter optimiert werden. Es wird kein Training benötigt, lediglich mehr Trainingsdatensätze, um das Modell zu verbessern. Der Nachteil liegt im hohen Berechnungsaufwand, da für jeden Datenpunkt die Distanz zu jedem anderen Punkt berechnet werden muss. (Bonnin, 2017, S. 73)

6.5.2. Naïve Bayes

Naïve Bayes-Klassifikatoren gehören zu den linearen Klassifikatoren. Sie sind dafür bekannt, einfach und effizient zu sein. Das probabilistische Modell basiert auf dem von Thomas Bayes entwickelten Bayes-Theorem. Es wird nach Raschka (2014, S. 3) wie folgt definiert:

$$A\ posteriori\ Wahr. = \frac{Bed.Wahrscheinlichkeit * A\ priori\ Wahr.}{Evidenz} \qquad (6.5.2.1)$$

Gegeben eines Datensatzes mit dem Merkmalsvektor $x = (x_1, x_2, \ldots, x_n)$ mit „n“ Merkmalen für jedes „x“ und „K“ möglichen Klassen $(y_1, y_2, \ldots, y_k)$ lässt sich dieser Satz nach Liu (2019, S. 113) folgendermaßen übertragen:

- Die A-priori-Wahrscheinlichkeit entspricht dem Wahrscheinlichkeitswert, welcher aufgrund von Vorwissen über die Eigenschaften des Systems erschlossen wird. In diesem Fall der Verteilung der Klassen: $P(y_k)$

- Die A-posteriori-Wahrscheinlichkeit eines Zustandes nach der Beobachtung einer Zufallsgröße. Im Falle der Klassifikation der Klasse in Abhängigkeit zum Datensatz X: $P(y_k|x)$
- Die bedingte Wahrscheinlichkeit beschreibt hier die Wahrscheinlichkeit eines Wertes für ein Merkmal x unter der Bedingung, dass der Datensatz der Klasse zugehörig ist: $P(x|y_k)$ oder für die Gesamtheit der Merkmale $P(x_1 x_2, \dots, x_n|y_k)$.
- Die Evidenz hängt allein von der Gesamtverteilung der Merkmale ab, also der Wahrscheinlichkeit für ein Merkmal x: $P(x)$

Demnach folgt der Term:

$$P(y_k|x) = \frac{P(x|y_k)P(y_k)}{P(x)} \tag{6.5.2.2}$$

Die Evidenz ist nicht spezifisch für bestimmte Klassen und daher konstant, wodurch diese keinen Einfluss auf das Bestimmen der A-posteriori-Wahrscheinlichkeit besitzt (ebd.). Aus diesem Grund kann der Term folgendermaßen vereinfacht werden:

$$P(y_k|x) = P(x|y_k)P(y_k) \tag{6.5.2.3}$$

Der Zusatz „Naive“ ist auf der naiven Annahme der bedingten Unabhängigkeit von Merkmalen begründet. Unter dieser naiven Annahme werden die bedingten Wahrscheinlichkeiten direkt aus den Trainingsdaten geschätzt, anstatt alle Möglichkeiten von x zu bewerten. So kann bei einem n-dimensionalen Merkmalsvektor x die bedingte Wahrscheinlichkeit wie folgt berechnet werden (Raschka, 2014, S. 5):

$$\begin{aligned} P(x|y_k) &= P(x_1|y_k) * P(x_2|y_k) * \dots * P(x_n|y_k) \\ &= \prod_{j=1}^{n} P(x_j|y_k) \end{aligned} \tag{6.5.2.3}$$

Das Ergebnis der Klassifikation ergibt sich aus der Klasse mit der höchsten A-posteriori-Wahrscheinlichkeit (Dinov, 2018, S. 292). Dies wird durch die argmax-Funktion ausgedrückt. Daraus ergibt sich folgende Definition für die Klassifikation mit Naïve Bayes:

$$\hat{y} = \underset{yk}{\operatorname{argmax}} P(y_k) \prod_{j=1}^{n} P(x_j|y_k) \qquad (6.5.2.3)$$

Die Bibliothek scikit-learn besitzt drei verschiedene Varianten von Naive Bayes Klassifikationen. Diese basieren auf verschiedenen probabilistischen Verteilungen: Bernoulli, Multinomiale und Gaußsche Verteilung. Die Bernoulli Verteilung ist eine Binärverteilung und wird angewandt, wenn ein Merkmal vorhanden oder nicht vorhanden sein kann. Eine Multinomiale Verteilung ist eine diskrete Verteilung und wird immer dann verwendet, wenn ein Merkmal durch eine ganze Zahl dargestellt werden muss. Gerade in der Textklassifikation basierend auf BoW wird oftmals auf dieses Modell zurückgegriffen (vgl. Kapitel 5.6.1). Die Gaußsche Verteilung ist eine kontinuierliche Verteilung, die durch ihren Mittelwert und ihre Varianz gekennzeichnet ist. Sie bietet eine Möglichkeit, eine Klassifikation stetiger Variablen durchzuführen. (Bonaccorso, 2017, S. 123) Da im Rahmen dieser Arbeit Merkmale ausschließlich stetig sind, wird auf dieses Modell zurückgegriffen.

Unter der Annahme, dass die Wahrscheinlichkeitsverteilungen der Merkmale einer normalen (Gaußschen) Verteilung folgen, kann die Bestimmung der Wahrscheinlichkeit dieser nach der folgenden Definition erfolgen:

$$P(x_j|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_j - \mu_y)^2}{2\sigma_y^2}\right) \qquad (6.5.2.4)$$

In der Formel steht μ für den Mittelwert des Datensatzes und σ für die Standardabweichung.

In der Praxis wird die Unabhängigkeitsannahme oft verletzt, dennoch tendieren Bayes-Klassifikatoren dazu, unter dieser unrealistischen Annahme sehr gut abzuschneiden (Rish, 2001). Insbesondere bei kleinen Stichprobengrößen können naive Bayes-Klassifikatoren die leistungsfähigeren Alternativen übertreffen (Domingos & Pazzani , 1997). Starke Verstöße gegen die Unabhängigkeitsannahmen und nichtlineare Klassifizierungsprobleme können zu sehr schlechten Leistungen der naiven Bayes-Klassifikatoren führen. In der Praxis wird daher empfohlen die Ergebnisse verschiedener Klassifikatoren zu vergleichen (Raschka, 2014, S. 3).

6.5.3. Support Vector Machines

Der Klassifikationsansatz von Support Vector Machines (SVM) ist, eine optimale Hyperebene zu finden, welche die Merkmale verschiedener Klassen am besten trennt (Liu Y. H., 2019, S. 138). Eine Hyperebene ist eine Ebene von n minus eins Dimensionen, die den n-dimensionalen Merkmalsraum der Beobachtungen in zwei Räume trennt. Beispielsweise ist die Hyperebene in einem zweidimensionalen Merkmalsraum eine Linie und in einem dreidimensionalen Merkmalsraum eine Fläche. Demnach handelt es sich bei SVM um einen binären Klassifikator. Bei einer Klassifikation mittels SVM spricht man auch von SVC (Support Vector Classifier). Soll Klassifikation über mehrere Klassen erfolgen, kann dies über die Ansätze „one-vs-all“ oder „one-vs-rest“ erfolgen (vgl. Kapitel 6.4). Im Folgenden wird das Verfahren auf Basis von Liu (2019, S. 138ff.) und Raschka (2016, S. 70ff.) vorgestellt.

In einem zweidimensionalen Raum kann eine Linie durch einen Steigungsvektor ω und einen Schnittpunkt b definiert werden. Ebenso kann in einem n-dimensionalen Raum eine Hyperebene durch einen n-dimensionalen Vektor ω und einen Schnittpunkt b definiert werden. Jeder Datenpunkt x auf der Hyperebene $\omega x + b = 0$ erfüllt. Eine Hyperebene ist eine trennende Hyperebene, wenn die folgenden Bedingungen erfüllt sind:

- Für jeden Datenpunkt x aus einer Klasse erfüllt sie $\omega x + b > 0$
- Für jeden Datenpunkt x aus einer anderen Klasse erfüllt er $\omega x + < 0$

Als Stützvektoren (engl. support vectors) werden die Punkte benannt, welche der Hyperebene am nächsten liegen. Der Korridor bezeichnet den Abstand der trennenden Hyperebene und der Stützvektoren. Bei SVM besteht das Optimierungsziel darin, diesen Korridor zu maximieren. Als positive bzw. negative Hyperebene werden die parallel zur trennenden Ebene und durch die Stützvektoren verlaufenden Ebenen bezeichnet. Sie werden wie folgt definiert:

$$\omega_0 + \omega^T x_{pos} = 1 \tag{6.5.3.1}$$

$$\omega_0 + \omega^T x_{neg} = -1 \tag{6.5.3.2}$$

Bild 8 visualisiert dies beispielhaft in einem zweidimensionalen Raum.

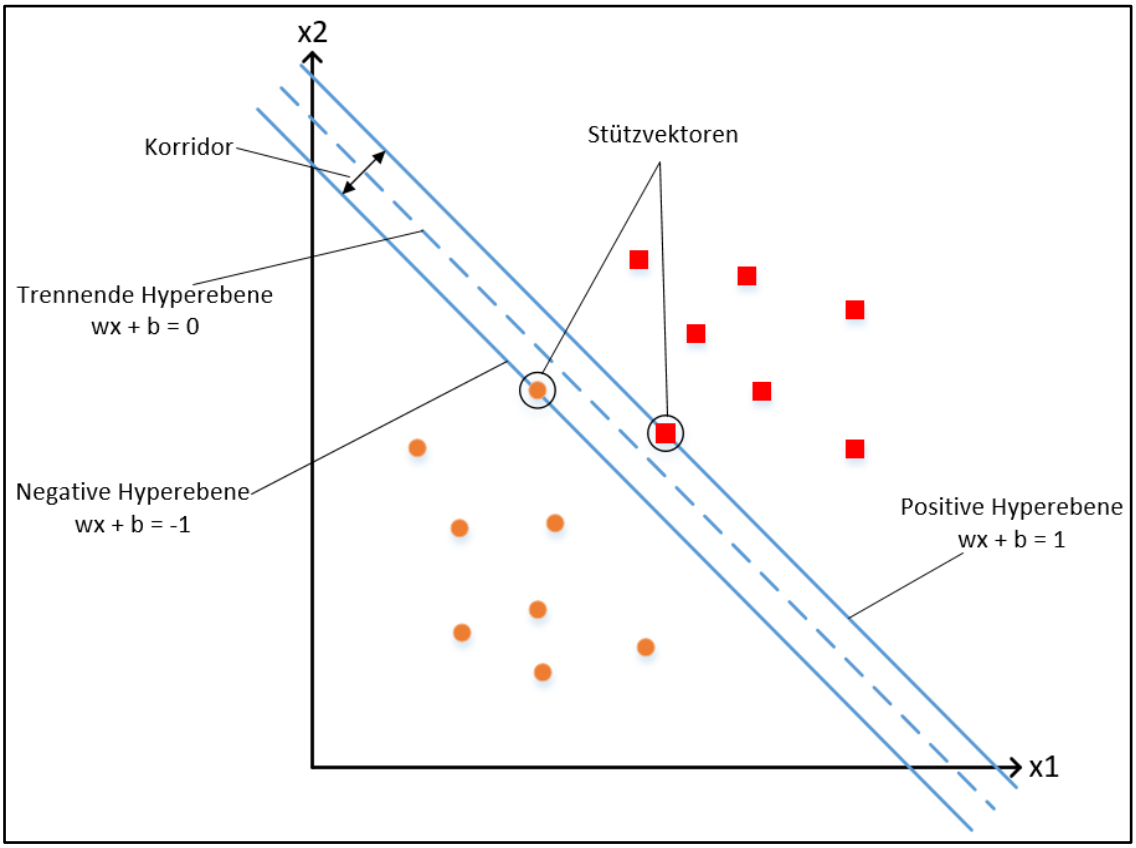


Bild 8: Klassifikation mittels SVC - Eigene Darstellung - Anlehnung an (Raschka, 2016, S. 69)

Der Grund für einen möglichst großen Korridor ist, dass solche Modelle tendenziell einen geringeren Generalisierungsfehler aufweisen, während Modelle mit kleinen Spielräumen anfälliger für eine Überanpassung sind (Raschka, 2016, S. 70).

Der Abstand zwischen der trennenden Hyperebene und den Stützvektoren errechnet sich durch:

$$\frac{|\omega x + b|}{\|\omega\|} = \frac{1}{\|\omega\|} \qquad (6.5.3.3)$$

Dies entspricht einem Korridor von $\frac{2}{|\omega|}$. Daraus folgt, dass $|\omega|$ minimiert werden muss, um den Korridor zu maximieren. Aufgrund der Tatsache, dass die Stützvektoren die am nächsten liegenden Punkte der Entscheidungshyperebene sind, wird die Bedingung hinzugefügt, dass kein Datenpunkt zwischen die positive und negative Hyperebene fällt:

$$\omega_0 + \omega^T x^{(i)} \geq 1, wenn\ y^{(i)} = 1 \quad (6.5.3.4)$$

$$\omega_0 + \omega^T x^{(i)} < -1, wenn\ y^{(i)} = -1 \quad (6.5.3.5)$$

Dies kann weiter zusammengefasst werden. Der Allquantor (∀) bezeichnet, dass eine Anwendung für jedes i erfolgt:

$$y^{(i)} \left(\omega_0 + \omega^T x^{(i)}\right) \geq 1\ \forall_i \quad (6.5.3.6)$$

Daraus folgt, dass ω und b, die die SVM-Entscheidungshyperebene bestimmen, durch das folgende Optimierungsproblem gelöst werden:

- Minimieren von $|\omega|$
- unter der Bedingung: $y^{(i)} \left(\omega_0 + \omega^T x^{(i)}\right) \geq 1$, für ein Trainingsset von $(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)$

Nach Raschka (2016, S. 70) ist es einfacher, den wechselseitigen Term $\frac{1}{2}\|\omega\|^2$ zu berechnen. Allerdings werden, um dieses Problem zu lösen, Methoden der quadratischen Optimierung benötigt, deren Beschreibung den Rahmen dieser Arbeit übersteigt.

Die Parameter ω und b werden verwendet, um einen Datensatz x‘ aufgrund der folgenden Bedingung zu klassifizieren:

$$\hat{y} \begin{cases} 1, & wenn\ \omega x' + b > 0 \\ -1, & wenn\ \omega x' + b < 0 \end{cases} \quad (6.5.3.7)$$

In der bisherigen Betrachtung erfolgte die Bestimmung des maximalen Korridors ohne die Betrachtung möglicher Ausreißer oder Fehler im Datensatz. Dies kann dazu führen, dass die Hyperebene zu sehr an die Testdaten angepasst wird und spätere Punkte dadurch schlechter klassifiziert werden. Es existiert eine Überanpassung.

Die Idee des „Soft Margins“ ist, eine gewisse Anzahl von Fehlern zu erlauben und den Spielraum so groß wie möglich zu halten, damit andere Punkte richtig klassifiziert werden können. Bild 9 beschreibt diesen Sachverhalt. Dargestellt wird ein Datensatz mit zugehöriger Hyperebene, links Hard- und rechts Soft Margin.

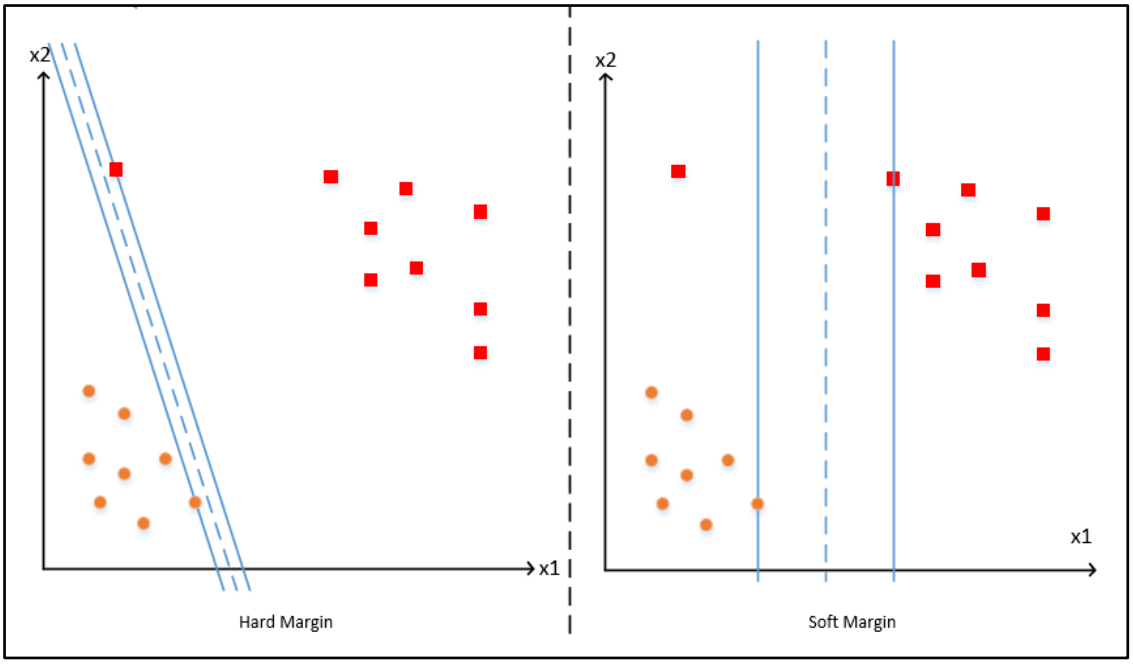


Bild 9: Soft and Hard Margin SVC - Eigene Darstellung – Anlehnung an (Raschka, 2016, S. 72)

Beim Soft Margin wird das Modell um die von Vladimir Vapnik 1955 eingeführte Schlupfvariable (ξ) erweitert (Raschka, 2016, S. 71). Diese positive Variable wird den linearen Beschränkungen hinzugefügt:

$$\omega_0 + \omega^T x^{(i)} \geq 1 - \xi^{(i)}, wenn\ y^{(i)} = 1 \tag{6.5.3.8}$$

$$\omega_0 + \omega^T x^{(i)} < -1 + \xi^{(i)}, wenn\ y^{(i)} = -1 \tag{6.5.3.9}$$

Der neue Term, der minimiert werden soll, lautet demnach:

$$\frac{1}{2}\|\omega\|^2 + \left(\sum_i \xi^{(i)}\right) \tag{6.5.3.10}$$

Der Hyperparameter `C` kontrolliert das Modell nach Liu (2019, S. 138) in zwei Dingen:

- Ein großer Wert von `C` bedeutet eine hohe Strafe für Fehlerklassifikation. Das bedeutet, dass das Modell anfällig für Überanpassung sein könnte, da während des Trainings nur wenige Fehler erlaubt sind. Ein SVM-Modell mit einem großen `C` hat eine geringe Verzerrung, aber es könnte eine hohe Varianz aufweisen.

- Wenn der Wert von C ausreichend klein ist, wird der Einfluss einer Fehlklassifikation relativ gering. Das Modell lässt mehr falsch klassifizierte Datenpunkte zu als das Modell mit großem C. Dadurch wird die Datentrennung weniger streng. Ein solches Modell hat eine geringe Varianz, aber es könnte durch eine hohe Verzerrung beeinträchtigt werden.

Für die meisten Datensätze lässt sich die Hyperebene nicht gut durch ein lineares Modell approximieren. In der SVM wird die Erweiterung auf den nichtlinearen Fall mit Hilfe der Kernel-Theorie behandelt (Igual & Seguí, 2017, S. 90). Kernfunktionen sind verallgemeinerte Funktionen, die zwei Vektoren (beliebiger Dimension) als Eingabe nehmen und eine Punktzahl ausgeben, die angibt, wie ähnlich die Eingabevektoren sind (ebd.). Der Kernel muss dabei eine positive semidefinite Funktion sein (ebd.). Die folgenden Kernel werden von scikit-learn unterstützt (scikit-learn developers, 2020i):

- Linearer Kernel:

$$k(x_i, x_j) = x_i^T x_j \tag{6.5.3.11}$$

- Polynomiale Kernel:

$$k(x_i, x_j) = \left(1 + x_i^T x_j\right)^p \tag{6.5.3.12}$$

- Radical Basis Function (kurz: RBF) Kernel:

$$k(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \tag{6.5.3.13}$$

Grundsätzlich kann jede Kernel-Funktion verwendet werden, die in der Lage ist, die Ähnlichkeit in höheren Dimensionen zu messen. Dabei werden die hochdimensionalen Ähnlichkeiten berechnet, ohne die Daten selbst in diese Dimension umzuwandeln. Dies erhöht dabei unwesentlich die Rechenkosten und wird deshalb Kerneltrick genannt (Misra, 2019). In der Regel ist es nicht möglich zu sagen, welcher Kernel optimal auf den Datensatz passt. In der Praxis erfolgt

das Auswählen durch ein Ausprobieren der Möglichkeiten. Ein zugehöriges Verfahren wird in Kapitel 6.8.3 vorgestellt.

6.6. Bias und Varianz Dilemma

Verzerrung[14] ist die Differenz zwischen der durchschnittlichen Vorhersage eines Modells und dem korrekten Wert der Vorhersage. Ein Modell mit hohem Bias passt sich wenig an die Trainingsdaten an und vereinfacht das Modell stark. Dies führt immer zu hohen Fehlern bei den Trainings- und Testdaten (Singh, 2018). Die Varianz beschreibt als Streuungsmaß die mittlere quadratische Abweichung der Vorhersagen zu den jeweiligen Datenpunkten. Ein Modell mit hoher Varianz pass sich den Trainingsdaten stark an und verallgemeinert nicht auf Daten, die es noch nicht gesehen hat (ebd.). Infolgedessen schneiden solche Modelle bei Trainingsdaten sehr gut ab, weisen aber bei Testdaten hohe Fehlerraten auf (ebd.).

Von einer Überanpassung spricht man, wenn das Modell das Rauschen zusammen mit dem Muster der Daten erfasst. Dies ist das Resultat einer Extraktion von zu vielen Informationen aus dem Datensatz. Die Überanpassung wird durch ein geringes Bias und eine hohe Varianz definiert, wodurch sie eine schlechte Leistung bei neuen Daten besitzen. Im Gegensatz dazu spricht man von einer Unteranpassung, wenn das Modell nicht in der Lage ist, das den Daten zugrundeliegende Muster zu erfassen. Gründe hierfür können das Training mit zu wenigen Daten oder die Auswahl eines falschen Modells sein, wie beispielsweise ein lineares Modell bei nicht linearen Daten. Sie besitzen ein hohes Bias und eine geringe Varianz. (Liu Y. H., 2019, S. 19)

Die folgende Abbildung beschreibt drei Modelle: Eine Überanpassung, eine Unteranpassung und ein ausbalanciertes Modell.

[14] Wird im Jargon des maschinellen Lernens auch Bias, abgeleitet vom englischen Begriff, genannt. Da dieser Ausdruck üblich ist, wird er im Folgenden verwendet.

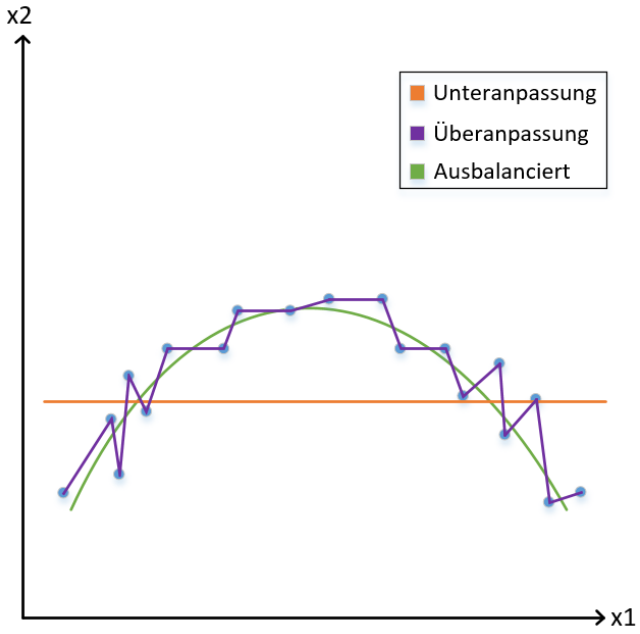


Bild 10: Über- und Unteranpassung - Eigene Darstellung – Anlehnung an (Singh, 2018)

Bias und Varianz stehen nach Brownlee (2019) dabei in einer direkten Beziehung zueinander:

- Eine Erhöhung des Bias wird die Varianz verringern.
- Eine Erhöhung der Varianz wird das Bias verringern.

Besitzt ein Modell sehr wenige Parameter, kann es ein hohes Bias und eine geringe Varianz aufweisen. Bei einer großen Anzahl an Parametern wird es hingegen eine hohe Varianz und ein geringes Bias aufweisen. Diese wechselseitige Beziehung führt dazu, dass ein optimales Ergebnis mit geringem Bias und einer kleinen Varianz nicht existiert. Die optimale Lösung liegt in der Mitte dieser Werte. (Singh, 2018)

6.7. Klassifikationsmetriken

Nach Bonnin (2017, S. 54) impliziert die Aufgabe der Klassifikation unterschiedliche Regeln für die Fehlerabschätzung. Der Vorteil besteht darin, dass die Anzahl der Ausgaben diskret ist, so dass auf binärem Wege genau bestimmt werden kann, ob eine Vorhersage fehlgeschlagen ist oder nicht.

Die Konfusionsmatrix ist eine Methode, um die Leistung verschiedener Klassifikationsalgorithmen zu verstehen. Das Grundprinzip besteht darin, in Form einer Matrix die Unterschiede zwischen den wahren und den vorhergesagten Klassen darzustellen. In einer binären Klassifikation besitzt die Matrix zwei Zeilen und Spalten. Bei einer Multi-Label-Klassifikation „k“ Zeilen und Spalten, wobei k der Anzahl der Klassen entspricht. Tabelle 12 beschreibt die Konfusionsmatrix einer binären Klassifikation. Jede Zeile entspricht einer tatsächlichen Klasse, während jede Spalte eine vorhergesagte Klasse darstellt. Eine jede Zelle wird dann verwendet, um die Anzahl der Stichproben in der Schnittmenge für diese Klassen zu zählen. (Liu & Jiang, 2013, S. 296)

Tabelle 12: Konfusionsmatrix binäre Klassifikation - Eigene Darstellung – Anlehnung an (Liu & Jiang, 2013, S. 296)

		Vorhergesagte Klasse	
		Positiv	Negativ
Tatsächliche Klasse	Positiv	richtig-positiv (rp)	falsch-negativ (fn)
	Negativ	falsch-positiv (fp)	richtig-negativ (rn)

Nach Bonaccorso (2017, S. 112) sind die einzelnen Schnittmengen dabei folgendermaßen definiert:

- Richtig-positiv: Eine positive Probe richtig klassifiziert.
- Falsch-positiv: Eine als positiv klassifizierte negative Probe.
- Richtig-negativ: Eine negative Probe richtig klassifiziert.
- Falsch-negativ: Eine als negativ klassifizierte positive Probe.

Bei einer Multi-Label-Klassifikation werden die tatsächlichen und vorhergesagten Klassen gegenübergestellt. Die Größe der Matrix beträgt somit k*k (vgl. Tabelle 13). Daraus folgt, dass die Schnittmengen für jede vorhergesagte Klasse einzeln bestimmt werden müssen. Tabelle 14 stellt die Schnittmengen für die vorhergesagte Klasse 1 dar.

Tabelle 13: Konfusionsmatrix drei Klassen - Eigene Darstellung – Anlehnung an (Liu & Jiang, 2013, S. 297)

		Vorhergesagte Klasse		
		Klasse 1	Klasse 2	Klasse 3
Tatsächliche Klasse	Klasse 1	Richtig	Falsch	Falsch
	Klasse 2	Falsch	Richtig	Falsch
	Klasse 3	Falsch	Falsch	Richtig

Tabelle 14: Schnittmengen Klasse 1 - Eigene Darstellung – Anlehnung an (Liu & Jiang, 2013, S. 296)

		Vorhergesagte Klasse		
		Klasse 1	Klasse 2	Klasse 3
Tatsächliche Klasse	Klasse 1	richtig-positiv (rp)	falsch-positiv (fp)	falsch-positiv (fp)
	Klasse 2	falsch-negativ (fn)	richtig-negativ (rn)	richtig-negativ (rn)
	Klasse 3	falsch-negativ (fn)	richtig-negativ (rn)	richtig-negativ (rn)

Die beiden Standardterminologien, die zur Analyse von Verwirrungsmatrixdaten verwendet werden sind Genauigkeit (engl. precision, P) und Trefferquote (engl. recall, R). Bonnin, (2017, S. 54) definiert die Genauigkeit als Fähigkeit des Klassifikators, eine negative Probe nicht als positiv zu markieren. Der beste Wert ist eins und der schlechteste Wert ist null. Die Trefferquote beschreibt er als die

Fähigkeit des Klassifikators, alle positiven Proben zu finden. Seine Werte reichen vom besten Wert eins bis null. Sie werden folgendermaßen berechnet:

$$P = \frac{tp}{tp + fp} \tag{6.7.1}$$

$$R = \frac{tp}{tp + fn} \tag{6.7.2}$$

Der Beta-Wert beschreibt einen gewichteten harmonischen Mittelwert zwischen Genauigkeit und Trefferquote. Ein Beta-Wert von eins bestimmt den so genannten F1-Wert, der ein perfektes Gleichgewicht zwischen den beiden Maßen darstellt und oftmals in der Praxis verwendet wird. Ein Beta-Wert kleiner als eins gibt der Genauigkeit mehr Bedeutung und ein Wert größer als eins gibt der Trefferquote mehr Bedeutung. Er wird nach Bonaccorso (2017, S. 114) folgendermaßen definiert:

$$F_{Beta} = (\beta^2 + 1)\frac{P * R}{\beta^2 * P + R} \tag{6.7.3}$$

Bei einer Gleichverteilung der Klassen kann die Metrik Präzision (engl. accuracy, A) verwendet werden. Sie beschreibt den Anteil der korrekten Vorhersagen des Modells und wird demnach wie folgt berechnet (Huilgol, 2019):

$$A = \frac{tp + tn}{(tp + fp + tn + fn)} \tag{6.7.4}$$

Sie wird ebenfalls für jede Klasse einzeln berechnet. Für eine allgemeine Präzision des Modells kann der Mittelwert dieser Präzision bestimmt werden (ebd.).

Nach (ebd.) wird Genauigkeit dann verwendet, wenn die richtig-positiven und richtig-negativen wichtiger sind, während die F1-Bewertung verwendet wird, wenn die falsch- negativen und falsch-positiven entscheidend sind.

6.8. Kreuzvalidierungsverfahren

Wie aus dem vorherigen Kapitel hervorgeht, wird immer eine Gewissheit benötigt, dass ein Modell die meisten Muster aus den Daten korrekt wiedergibt und dass es dabei nicht zu viel Rauschen aufnimmt beziehungsweise, dass das Modell wenig Bias und Varianz aufweist. Das Prüfen, ob die numerischen Ergebnisse die hypothetischen Beziehungen zwischen Variablen quantifizieren, wird als Validierung bezeichnet. Dieser Prozess beinhaltet eine Schätzung des Unterschieds zwischen den vorhergesagten und den korrekten Antworten. Durch das Auswerten der Trainingsfehler kann eine Aussage darüber getroffen werden, wie gut das Modell mit den Daten abschneidet, mit denen es trainiert wurde. Hier kann allerdings keine Aussage darüber getroffen werden, ob ein Modell über- oder unterangepasst ist. Hierfür muss eine Prüfung auf „neuen Daten“ erfolgen, welche im Training nicht verwendet wurden. (Gupta, 2017)

In einer Runde der Kreuzvalidierung (engl. Cross-Validation) werden die Originaldaten in zwei Untergruppen unterteilt, jeweils für Training und Test. Durch einen Test kann dann die Leistung des Modells auf Daten geprüft werden, welche nicht im Training verwendet wurden. Auf diese Weise können mehrere Runden der Kreuzvalidierung durchgeführt werden. Die Testergebnisse aus allen Runden werden schließlich gemittelt, um eine zuverlässigere Schätzung der Modellvorhersageleistung zu erhalten (Liu Y. H., 2019, S. 23f.). Kreuzvalidierung ist demnach eine statistische Methode, die zur Schätzung der Leistung von Modellen des maschinellen Lernens verwendet wird (Brownlee, 2020). Grundsätzlich ist es immer vorteilhaft, eine Kreuzvalidierung durchzuführen. Die einzige Ausnahme besteht, wenn die Rechenkosten als zu hoch angesehen werden.

Es existieren hauptsächlich zwei Kreuzvalidierungsschemata, erschöpfende und nicht erschöpfende. Erschöpfende verwenden in jeder Runde eine festgelegte Anzahl von Beobachtungen als Test- (oder Validierungs-) Stichproben und die übrigen Beobachtungen als Trainingsstichproben. Dieser Prozess wird so lange wiederholt, bis alle möglichen verschiedenen Untergruppen von Stichproben einmal für den Test verwendet wurden. Im Gegensatz dazu verwenden nicht

erschöpfende Methoden nicht alle möglichen Kombinationen (Liu Y. H., 2019, S. 24). Im Folgenden werden einige Methoden sowie ein spezieller Anwendungsfall der Kreuzvalidierungsverfahren zur Parameteroptimierung aufgezeigt.

6.8.1. Holdout Methode

In diesem einfachen Ansatz wird das Modell auf Trainingsdaten trainiert und auf Testdaten ausgewertet. Hierbei werden keine Iterationen durchlaufen. Normalerweise ist der Umfang der Trainingsdaten mehr als doppelt so groß wie der der Testdaten, übliche Verhältnisse sind 60:40, 70:30 oder 80:20 (Mutjaba, 2020). Vor einem Aufteilen der Daten wird der Datensatz zufällig gemischt. Dies hat zur Folge, dass das Modell bei jedem Training andere Ergebnisse liefert. Die Kritik daran besteht darin, dass der zufällig ausgewählte Trainingssatz unter Umständen repräsentativ für den gesamten Datensatz ist und Informationen durch das fehlende Trainieren auf den Testdaten verloren gehen. (ebd.)

6.8.2. K-fache Kreuzvalidierung

Die K-fache Kreuzvalidierung (engl. k-fold cross-validation) ist eine nicht erschöpfende Methode, welche den Datensatz in „K“ gleichmäßig große Teile teilt. Bei jeder Iteration wird eine dieser Gruppen zum Testsatz und alle anderen zum Trainingssatz. Dies wird K mal wiederholt, bis jede Gruppe einmal der Testdatensatz war. Für die Auswertung werden die Ergebnisse der einzelnen Runden gemittelt. (Liu Y. H., 2019, S. 24) Bild 11 stellt die Methode graphisch dar.

Bild 11: Kreuzvalidierungsverfahren - Eigene Darstellung – Anlehnung an (Mutjaba, 2020)

Nach Brownlee (2020) ist das Definieren von K ein wichtiger Schritt. Ein schlecht gewählter Wert für K kann zu einer falschen repräsentativen Vorstellung von der Leistung des Modells führen. Für die richtigen Wert für K spricht er drei Strategien an:

- Der Wert für K wird so gewählt, dass jeder Zug/jede Testgruppe von Datenproben groß genug ist, um für den breiteren Datensatz statistisch repräsentativ zu sein.
- Der Wert für K wird auf 10 festgelegt, ein Wert, der durch Experimente gefunden wurde, um im Allgemeinen zu einer Modellfertigkeitsschätzung mit geringerem Bias und bescheidener Varianz zu führen.
- K=n: Der Wert für K ist auf n festgelegt, wobei n die Größe des Datensatzes ist, um jeder Teststichprobe die Möglichkeit zu geben, im Hold Out-Datensatz verwendet zu werden. Dieser Ansatz wird als „Leave-one-out-Kreuzvalidierung“ bezeichnet und beschreibt eine erschöpfende Methode.

6.8.3. Hyperparameteroptimierung

In der Hyperparameteroptimierung wird ebenfalls auf die Kreuzvalidierung zurückgegriffen. Der Ansatz besteht in einem erschöpfenden Suchparadigma, bei dem eine Liste von Werten für verschiedene Hyperparameter spezifiziert werden und die Modellleistung für jede Kombination dieser Werte ausgewertet wird, um die optimale Menge zu erhalten (Igual & Seguí, 2017, S. 91). Scikit-learn besitzt hierfür eine eigene Klasse GridSearchCV, welche im Rahmen dieser Arbeit verwendet wird (scikit-learn developers, 2020c).

6.9. Label-Kodierung

Einige Modelle des maschinellen Lernens funktionieren nur mit numerischen Werten (Liu Y. H., 2019, S. 31). Für Sie ist es nicht möglich, Zeichenketten wie Namen zu interpretieren. Des Weiteren handelt es sich bei den Namen der Autoren um personenbezogene Daten (intersoft consulting, 2020). Zwar wurde die Zustimmung des Chefredakteures Thoralf Koß eingeholt, dennoch gilt es, die informationelle Selbstbestimmung der einzelnen Personen zu wahren (vgl. Anhang A1). Aus diesen Gründen wird jedem Autor der Datengrundlage eine Zahl, bzw. ein Label zugeordnet, durch welches er im Verlauf der Arbeit angesprochen wird. Dieser Prozess wird als Label-Kodierung bezeichnet.

6.10. Skalierung der Daten

Ein generischer Datensatz besteht aus verschiedenen Werten, die aus verschiedenen Verteilungen mit unterschiedlichen Skalen gewonnen werden können (Bonaccorso, 2017, S. 51). Ein Algorithmus für maschinelles Lernen ist nicht in der Lage, zwischen diesen verschiedenen Situationen zu unterscheiden, wodurch größere Werte die kleineren dominieren und dadurch das Ergebnis beeinträchtigen (Liu Y. H., 2019, S. 32). Aus diesem Grund müssen Datensätze vor ihrer Verarbeitung standardisiert bzw. auf einen einheitlichen Wertebereich zu skaliert werden (ebd.). Beispielsweise ist die Wortanzahl in einem Text immer um ein Vielfaches größer, als die durchschnittliche Wortlänge.

Für das Skalieren existieren zwei geläufige Techniken: Normalisieren und Standardisieren. Normalisieren bildet einen Datensatz auf einen definierten Zahlenbereich, beispielsweise von null bis eins, ab. Die Berechnung erfolgt dabei folgendermaßen (scikit-learn developers, 2020f):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{6.10.1}$$

Entspricht der Wert von X dem Minimum der Reihe, ergibt der Zähler und damit X‘ null. Nimmt X den maximalen Wert an, entsprechen Nenner und Zähler demselben Wert und das Ergebnis ist dadurch eins. Liegt der Wert von X zwischen dem Minimum und Maximum, liegt der Wert von X‘ zwischen null und eins. (Jaju, 2017)

Die Standardisierung subtrahiert den Mittelwert μ der Zahlenreihe vom Wert und dividiert das Ergebnis durch die Standardabweichung σ:

$$X' = \frac{X - \mu}{\sigma} \tag{6.10.2}$$

Als Ergebnis bewegen sich die Werte um null als Zentrum. Dabei gilt zu beachten, dass die Werte auf keinen Zahlenbereich limitiert sind, wodurch Ausreißer nicht beeinflusst werden. Die Standardisierung ist eine Anforderung von vielen Modellen des maschinellen Lernens. Sie funktionieren teilweise nur sehr schlecht, wenn die Daten nicht oder nur kaum einer Normalverteilung gleichen. Beispielsweise gehen SVM davon aus, dass alle Daten um null zentriert sind. (scikit-learn developers, 2020g)

Welche der beiden Techniken anzuwenden ist, hängt vom Anwendungsfall und dem verwendeten Algorithmus ab. Folgen die Daten keiner Gaußschen Verteilung und ist dies auch keine Voraussetzung für die Daten, ist die Normalverteilung vielversprechend. Beispiele für Algorithmen sind KNN und neuronale Netze. Wie bereits erwähnt kann die Standardisierung in Fällen hilfreich sein, in denen die Daten einer Gaußschen Verteilung folgen. Dies stellt aber keine notwendige Bedingung dar. Im Zweifelsfall können die Daten an die einzelnen Techniken angepasst und das Ergebnis des Modells geprüft werden. (Bhandari, 2020)

6.11. Verringern der Dimensionalität

Nach Azevedo (2019) sind zwei der größten Herausforderungen des maschinellen Lernens die Überanpassung (vgl. Kapitel 6.6) und der „Fluch der Dimensionalität“. Der "Fluch der Dimensionalität" bezieht sich auf das Problem Struktur in Daten zu finden, die in einen hochdimensionalen Raum eingebettet sind. Je mehr Merkmale existieren, desto mehr Datenpunkte werden benötigt, um den Raum zu füllen (Rojas, 2015). In der Wissenschaft wird hier oftmals „Ockhams Rasiermesser“ zitiert. Ockhams Rasiermesser spiegelt sowohl eine allgemeine Regel der Vorsicht als auch eine konservative Richtlinie für die Untersuchung wieder, die darin besteht, die einfachsten unter den möglichen Erklärungen für das untersuchte Phänomen oder Modellansätze anzustreben (Dresp-Langley, et al., 2019). Eine hohe Anzahl von Merkmalen bzw. eine hohe Komplexität hat eine allgemein schlechtere Leistung zur Folge. Des Weiteren erhöht die Anzahl der Merkmale die benötige Rechenleistung enorm. Die Anzahl der notwendigen Berechnungen ist abhängig von der Anzahl der Datensätze und Merkmale (vgl. Kapitel 6.5). Ein Verringern der Dimensionalität kann grundsätzlich auf zwei Wegen erfolgen: Merkmalsauswahl und -reduktion.

Es existieren nach Azevedo (2019) drei Hauptgruppen von Methoden zur Merkmalsauswahl, welche in den folgenden Abschnitten erläutert werden:

- Filter
- Wrapper
- Eingebettete Methoden (engl. embedded methods)

Darauffolgend wird die Hauptkomponentenanalyse als eine Methode der Merkmalsreduktion vorgestellt.

6.11.1. Filter

Filterbasierte Modelle errechnen in einem Datensatz festgelegte statistische Werte und suchen anhand dieser redundante oder irrelevante Informationen. Im Falle einer Redundanz wird ein Merkmal, bei Irrelevanz kein Merkmal

beibehalten (Coelho & Richert, 2015, S. 242). Die gängigsten statistischen Werte werden im Folgenden kurz anlehnend an Kaushik (2016) erläutert.

Der Varianz-Schwellwert ist ein einfacher Basisansatz für die Merkmalsauswahl. Er entfernt alle Merkmale, deren Varianz einen bestimmten Schwellenwert nicht erreichen. Standardmäßig werden alle Merkmale mit Null-Varianz entfernt, d.h. Merkmale, die in allen Stichproben denselben Wert haben. (scikit-learn developers, 2020a)

Der Korrelationskoeffizient ist ein Maß zum Berechnen der Abhängigkeit zwischen zweier stetigen Variablen X und Y. Dieser kann Werte zwischen -1 und +1 annehmen. Er wird definiert durch:

$$\rho_{x,y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \qquad (6.11.1.1)$$

Die Diskriminanzanalyse versucht lineare Kombinationen von Merkmalen zu finden und separiert sie in eine oder mehrere Klassen einer kategorischen Variablen. Es wird vorausgesetzt, dass die Daten innerhalb der Klassen normalverteilt sind.

Die Varianzanalyse ähnelt der linearen Diskriminanzanalyse mit der Ausnahme, dass sie mit einem oder mehreren kategorialen unabhängigen Merkmalen und einem kontinuierlichen abhängigen Merkmal durchgeführt wird. Sie bietet einen statistischen Test, ob die Mittelwerte mehrerer Gruppen gleich sind oder nicht.

Der Chi-Quadrat Test ist ein statistischer Test, der auf die Gruppen kategorialer Merkmale angewandt wird, um die Wahrscheinlichkeit einer Korrelation oder Assoziation zwischen ihnen anhand ihrer Häufigkeitsverteilung zu bewerten. Die Methode erwartet dabei ausschließlich positive Merkmale (Azevedo, 2019).

Eine weitere Möglichkeit sind Filter basierend auf dem Informationsgewinn (engl. information gain) von Variablen. Die grundlegende Idee der Informationstheorie ist zu quantifizieren, wie viel Information in einer Nachricht enthalten ist. Allgemeiner ausgedrückt kann damit die Information in einem Ereignis und einer Zufallsvariable quantifiziert werden, was als Entropie bezeichnet wird. Dies wird mithilfe der Wahrscheinlichkeitstheorie berechnet. Der Informationsgewinn für

jede Variable kann ausgewertet werden und die Variable ausgewählt werden, welche den Informationsgewinn maximiert, was wiederum die Entropie minimiert. In der Merkmalsauswahl wird der Gewinn jeder Variablen im Zusammenhang mit der Zielvariablen bewertet. In dieser Verwendung wird die Berechnung als gegenseitige Information (engl. mutual information) zwischen den beiden Zufallsvariablen bezeichnet. (Brownlee, 2020)

6.11.2. Wrapper

Filter sind effizient im Eliminieren nutzloser Merkmale. Sie sind jedoch in ihrer Bewertung limitiert. So existieren Merkmale, die voneinander unabhängig sind, eine Abhängigkeit von der Ereignisvariablen aufweisen, aber aus Sicht des Modells keine Information liefern. Aus diesem Grund verwenden Wrapper das Modell als Bewertungsfunktion, indem sie den Mehrwehrt eines Merkmals für dieses ermitteln. (Coelho & Richert, 2015, S. 251)

Für den Ablauf eines Wrappers gibt es unterschiedliche Möglichkeiten. Deterministische Wrapper beginnen mit keinen (Vorwärts-Auswahl) oder den im Modell enthaltenen Merkmalen (Rückwärts-Auswahl). Gemäß dem Modell werden die Merkmale dann iterativ verkleinert oder vergrößert. Das Hinzufügen oder Entfernen von Merkmalen kann dabei auf statistischen Variablen basieren. Bild 12 beschreibt den Ablauf einer Rückwärts-Auswahl. Stochastische Wrapper-Merkmalsauswahlverfahren verwenden einen binären Merkmalsindizierungsvektor, der angibt, ob jede Variable in die Liste der hervorstechenden Merkmale aufgenommen werden soll oder nicht. Nach jeder Iteration wird der Indikatorvektor manipuliert und ein Vergleich der Indexierungsänderung vor und nach der zufälligen Veränderung durchgeführt. Es wird der Indexierungsvektor ausgewählt, welcher die beste Leistung aufgrund einer bestimmten Metrik entspricht. Dieser Prozess wird solange wiederholt, bis keine Verbesserung mehr beobachtet wird. (Dinov, 2018, S. 558)

In der Praxis zeigen sich Wrapper in der Regel mit guten Resultaten. Kritisiert wird allerdings oftmals die benötigte Rechenleistung. Gerade bei mehreren

hundert Merkmalen kann der Prozess sehr rechenintensiv werden. (Azevedo, 2019)

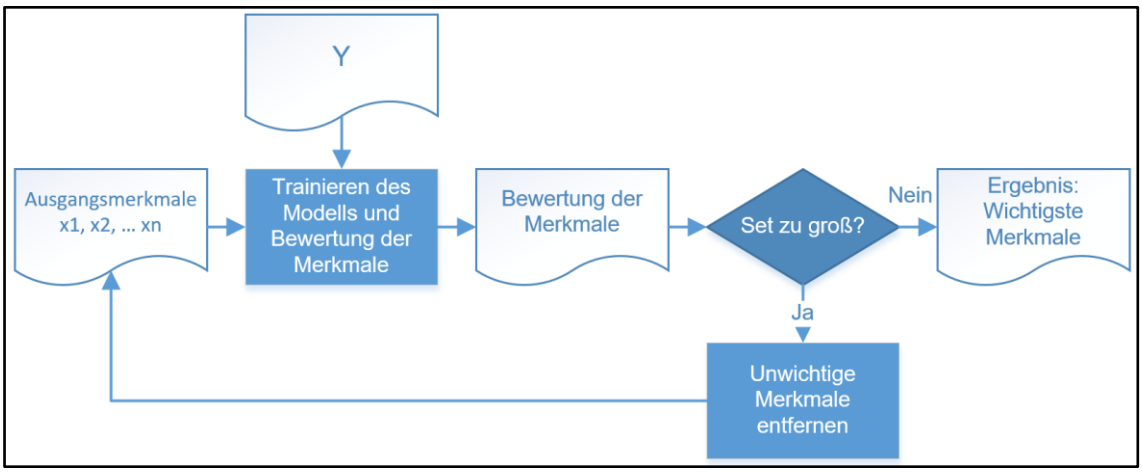


Bild 12: Ablauf Wrapper - Eigene Darstellung – Anlehnung an (Coelho & Richert, 2015, S. 251)

6.11.3. Eingebettete Methoden

Eingebettete Methoden sind bereits in den Algorithmen implementiert, welche ihre eigene Merkmalsauswahl besitzen. Das Training des Klassifikators beschreibt ein Optimierungsproblem, bei dem die Parameter (θ) einer Funktion minimiert werden sollen. Diese Funktion ist auch als Verlustfunktion bekannt (notiert als L(θ)). Allgemein beschrieben soll eine objektive Funktion die Komplexität des Modells minimieren, welche eine Verlustfunktion als auch Strafe ($\Omega(\theta)$) berücksichtigt. (Azevedo, 2019)

$$obj(\theta) = L(\theta) + \Omega(\theta) \qquad (6.11.36.11.1.1)$$

Die Verlustfunktion und Strafe werden dabei abhängig vom Modell definiert. Daraus folgt die Abhängigkeit der eingebetteten Methoden zum jeweiligen Modell. Besitzt der Klassifikator keine Methode, kann auf diesem Weg keine Merkmalsauswahl erfolgen. (ebd.)

6.11.4. Hauptkomponentenanalyse

Die Hauptkomponentenanalyse (engl. Principal component analysis, kurz: PCA) ist eine Technik, die für Anwendungen, wie Merkmalsreduktion, verlustbehaftete

Datenkompression, Merkmalsextraktion und Datenvisualisierung weit verbreitet ist (Bishop, 2006, S. 561). Anders als bei Wrappern oder eingebetteten Methoden handelt es sich hierbei um ein Verfahren des unsupervised learnings. Ebenfalls werden in diesem Verfahren die Merkmale nicht selektiert, sondern reduziert. Nach Liu (2019, S. 326) besitzt das Verfahren folgende Vorteile:

- Reduzierung der Trainingszeit von Vorhersagemodellen, da redundante oder korrelierende Features zu neuen Features zusammengeführt werden
- Verringerung der Überanpassung durch die Reduktion des beschriebenen „Rauschens“
- Wahrscheinlich verbesserte Leistung, da die Vorhersagemodelle aus Daten mit weniger redundanten oder korrelierten Features lernen

Im Folgenden werden die einzelnen Schritte der PCA nach Raschka (2016, S. 128ff.) und Bonaccorso (2017, S. 56ff.) aufgezeigt. Das Vorgehen wird hier kurz als das Finden der maximalen Varianz in hochdimensionalen Daten und das Projizieren dieser auf einen neuen Unterraum beschrieben. Dieser Unterraum besitzt genauso viele oder weniger Dimensionen. Die Hauptkomponenten (engl. principal components, kurz: PC) können als die Richtungen der maximalen Varianz interpretiert werden, wenn die neuen Merkmalsachsen orthogonal zueinander (nicht korreliert) sind. Dies wird in Bild 13 für den zweidimensionalen Raum veranschaulicht. Die Merkmale werden auf den Achsen durch $x1$ und $x2$ dargestellt, die Hauptkomponenten durch $PC1$ und $PC2$.

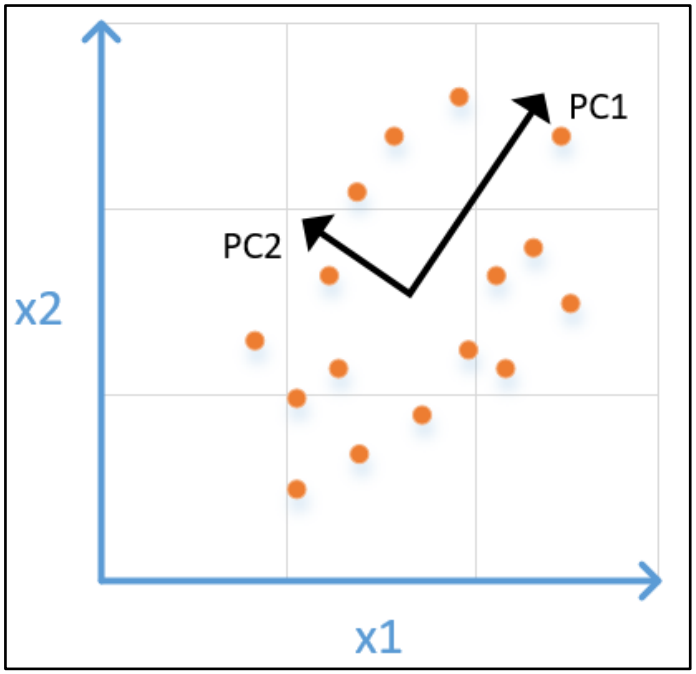


Bild 13: Hauptkomponenten im zweidimensionalen Raum - Eigene Darstellung – Anlehnung an (Raschka, 2016, S. 128)

Die wesentlichen Schritte der Hauptkomponentenanalyse umfassen:

1. Standardisieren des Datensatzes.
2. Konstruieren der Kovarianzmatrix.
3. Zerlegen der Kovarianzmatrix in ihre Eigenvektoren und Eigenwerte.
4. Wählen von k Eigenvektoren, die den k größten Eigenwerten entsprechen, wobei k die Dimensionalität des neuen Merkmalsunterraums ($k \leq d$) ist.
5. Konstruieren der Projektionsmatrix W aus den gewählten Eigenvektoren.
6. Transformieren des d-dimensionalen Eingabedatensatz X mit Hilfe der Projektionsmatrix W, um den neuen k-dimensionalen Merkmalsunterraum zu erhalten.

Die PCA reagiert sensibel auf unterschiedliche Skalierungen der Daten. Deshalb umfasst der erste Schritt immer eine Standardisierung (vgl. Kapitel 6.10).

Für eine mögliche Beurteilung der Information eines Merkmales und der zwischen ihnen herrschenden Korrelation wird eine Kovarianzmatrix (C) erstellt. In der symmetrischen $d*d$-dimensionalen Matrix werden die paarweisen

Kovarianzen zwischen den einzelnen Merkmalen gespeichert. Diese ist definiert durch:

$$\sigma_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - E[x_i])(x_{kj} - E[x_j]) \qquad (6.11.46.11.1.1)$$

Der Term $E[x]$ entspricht dem Erwartungswert des Merkmales x. Dieser entspricht hier dem Wert, welchen die diskrete Zufallsvariable durchschnittlich annimmt, demnach dem Mittelwert. Der Erwartungswert entspricht durch die Standardisierung hier dem Wert null. Eine positive Kovarianz zwischen zwei Merkmalen zeigt an, dass die Merkmale gemeinsam zunehmen oder abnehmen, während eine negative Kovarianz anzeigt, dass die Merkmale in entgegengesetzter Richtung variieren.

Die Kovarianzmatrix ist positiv semidefinit, was bedeutet, dass alle Eigenwerte der Matrix größer oder gleich null sind. Sie wird definiert durch:

$$c = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{pmatrix} \qquad (6.11.46.11.1.2)$$

Die Eigenvektoren der Kovarianzmatrix entsprechen den Hauptkomponenten, während die Eigenwerte deren Größe definieren. Die Definition dieser lautet nach Schneider (2013) wie folgt: „Ein Eigenvektor $\vec{x}$ einer Matrix ist ein vom Nullvektor verschiedener Vektor, dessen Richtung durch Multiplikation mit der Matrix nicht verändert wird. Der Eigenvektor wird also nur gestreckt. Der Streckungsfaktor λ heißt Eigenwert der Matrix.“ Das bedeutet, dass ein Vielfaches der Matrix sich durch eine Multiplikation des Eigenvektors und des Eigenwertes abbilden lässt:

$$c * \vec{x} = \lambda * \vec{x} \qquad (6.11.46.11.1.3)$$

Ein manuelles Berechnen der Eigenvektoren und Eigenwerte ist sehr umfangreich und wird in dieser Arbeit nicht behandelt. In der Praxis werden hierfür vorgefertigte Bibliotheken verwendet.

Die Anzahl der Eigenvektoren und zugehörigen Eigenwerten entspricht der Anzahl der Dimensionen der Matrix. Demzufolge ergeben sich aus der Eigenwertzerlegung d Eigenwerte mit den entsprechenden Eigenvektoren,

welche in einer $d * d$-dimensionalen Matrix als Spalten gespeichert sind. Diese Matrix entspricht der Projektionsmatrix W.

Nach Cheplyaka (2017) ist die Gesamtvarianz definiert als die Summe der Varianzen aller einzelnen Hauptkomponenten. Der Anteil der durch eine Hauptkomponente erklärten Varianz ist das Verhältnis zwischen der Varianz dieser Hauptkomponente und der Gesamtvarianz. Demnach wird die erklärte Varianz einer Hauptkomponente λ_j ermittelt durch:

$$\frac{\lambda_j}{\sum_{j=1}^{d} \lambda_j} \qquad (6.11.46.11.1.3)$$

Um die Dimensionalität des Datensatzes zu reduzieren, soll dieser auf einen neuen Merkmalsraum komprimiert werden. Zu diesem Zweck wird eine Teilmenge der Eigenvektoren ausgewählt, welche den größten Teil der Information enthalten. Hierbei definieren die Eigenwerte die Größe der Eigenvektoren, weshalb sie nach abnehmender Größe sortiert werden. Es werden also die k informativsten Eigenvektoren selektiert.

Üblicherweise wird der größte Teil der Varianz durch die erste Komponente erklärt und die weitere individuell erklärte Varianz nimmt mit jeder weiteren Komponente ab. Dadurch ist es möglich, die ursprüngliche Dimensionalität zu reduzieren, ohne einen inakzeptablen Informationsverlust zu erleiden, der einen Algorithmus dazu bringen könnte, falsche Klassen zu lernen. Selbstverständlich ist dies aber von den jeweiligen Merkmalen abhängig Bonaccorso (2017, S. 58). Bild 14 zeigt einen üblichen Verlauf.

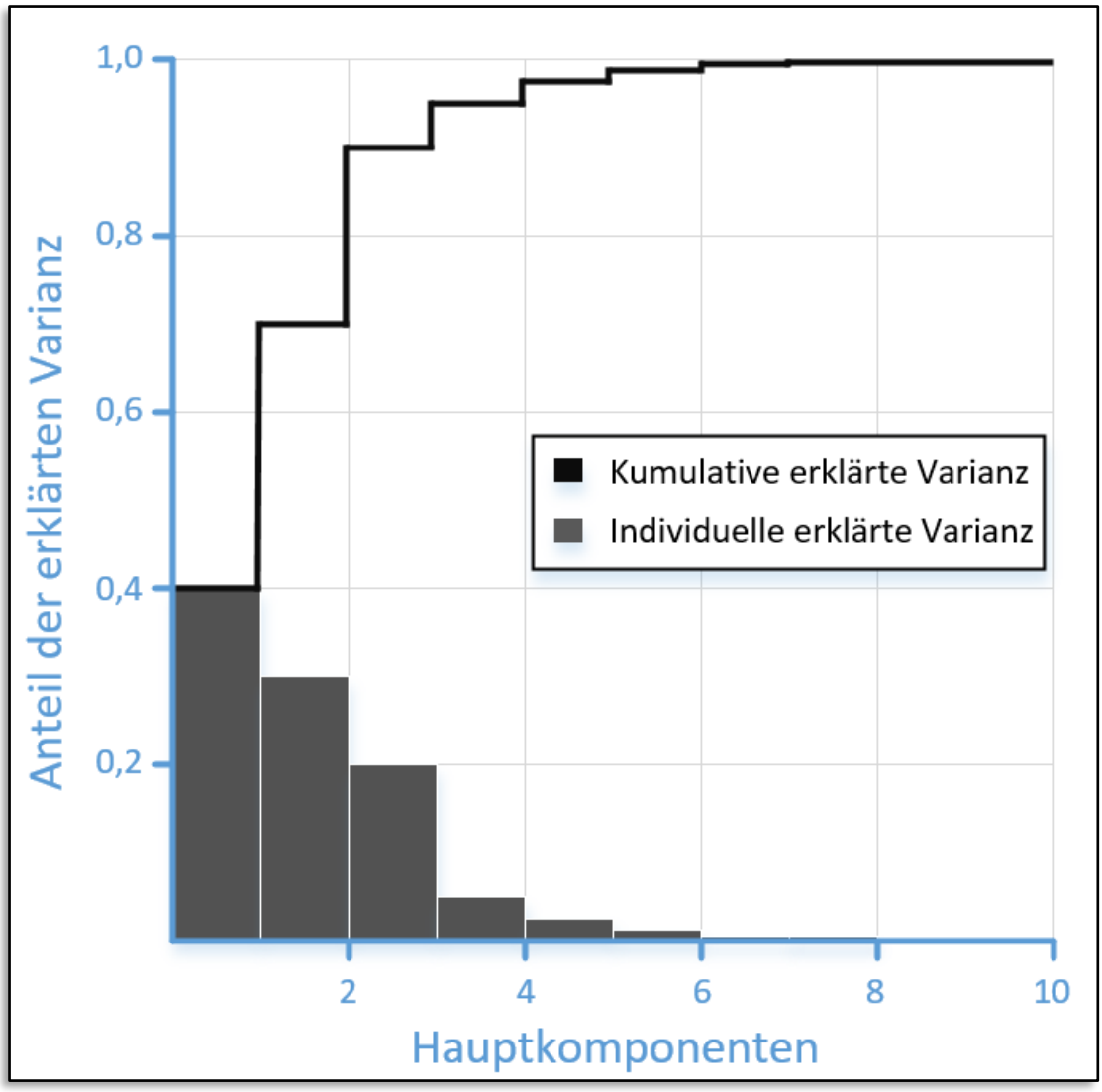


Bild 14 Hauptkomponentenanalyse erklärte Varianz - Eigene Darstellung – Anlehnung an (Raschka, 2016, S. 132)

Ein Nachteil der PCA ist, dass nach dem Verfahren nicht aufgeführt werden kann, welchen Informationsgehalt die einzelnen Merkmale besitzen. Dies kann nur durch die entstandenen Hauptkomponenten erfolgen. Des Weiteren ist nicht immer ein Verlauf wie in Bild 14 zu erkennen. Je weniger Abhängigkeiten zwischen den Merkmalen existieren, desto weniger Informationen können durch eine Hauptkomponente abgebildet werden. Ist dies der Fall, kann keine Reduktion der Dimensionalität ohne inakzeptablen Informationsverlust erfolgen.

Die Ergebnisse der PCA stehen in großer Abhängigkeit zur Anzahl der Stichproben (n) und Variablen (p). Beim Versuch eine allgemeine Faustregel zu definieren gibt es verschiedene Empfehlungen für ein Verhältnis zwischen den

Stichproben und Variablen (n:p). Nach Cattell (1978) sollte dieses zwischen 3 und 6 liegen. Gorsuch (1983) sprach sich für ein Minimum von 5 und Everitt (1975) für ein Minimum von 10 aus. MacCallum, Widaman, Zhang, & Hong (1999) zeigten in ihren Studien, dass die allgemeinen Faustregeln nicht gültig sind und dass die niedrigste Stichprobengröße von anderen Aspekten des Stichprobendesigns abhängt. Einen großen Einfluss besitzt die Korrelationsstruktur. Ist diese stark ausgeprägt, können bessere Ergebnisse mit einer kleineren Stichprobenanzahl erzielt werden (Rao, Shaukat, & Khan, 2016, S. 176). Aus diesem Grund gilt es im späteren Verlauf zu prüfen, ob das Verhältnis der Stichproben und Variablen in Abhängigkeit zur Korrelationsstruktur in dieser Arbeit ausreicht, um die Merkmale sinnvoll zu reduzieren.

7. Datenvorbereitung

Die Phase Datenvorbereitung umfasst alle Aktivitäten zur Erstellung der Daten, welche schlussendlich im Modell verwendet werden (vgl. Kapitel 2.3)

In der Vorverarbeitung werden Schritte zur Bereinigung der Daten durchgeführt. Des Weiteren erfolgt ein Aufteilen in Test- und Trainingsdaten. Ein Filtern der Daten, beziehungsweise die Merkmalsauswahl, wird anhand eines vorgestellten iterativen Verfahrens durchgeführt. Eine weitere Selektion erfolgt auf Basis der verschiedenen Varianten einiger Merkmale. Zu diesem Zweck werden die verschiedenen Varianten aufgezeigt und bewertet.

7.1. Vorverarbeitung

Im Rahmen der Vorverarbeitung werden alle grundsätzlichen Schritte durchgeführt, welche eine weitere Verarbeitung der Daten vereinfachen oder ermöglichen.

Merkmale mit einer zu geringen oder zu hohen Häufigkeit können eliminiert werden, bevor die Daten in Test- und Trainingsdatensätze aufgeteilt werden. Dies ist nach Forman (2002, S. 152) unproblematisch, solange der Filter nicht auf die Klassenbezeichnungen des Testsatzes zugreift und dadurch keine Informationen des Testsatzes an den Trainingssatz gegeben werden.

Da weitere Methoden statistische Maße oder Algorithmen anwenden, welche auf die Klassenbezeichnungen zugreifen, muss der Datensatz vor der weiteren Verarbeitung in einen Trainings- und Testdatensatz unterteilt werden.

Es besteht die Vermutung einer Korrelation zwischen durch Häufigkeiten bestimmten Merkmalen und der Textlänge. Merkmale mit hoher Korrelation sind stärker linear abhängig voneinander und haben daher fast die gleiche Wirkung auf die abhängige Variable. Es gilt diese Korrelation zu analysieren und gegebenenfalls aufzulösen.

Ein gängiger Schritt in der Vorverarbeitung der Daten ist deren Normalisierung bzw. Standardisierung. An dieser Stelle ist noch unklar, welche Methode ein besseres Ergebnis erzielt. Um eine Basis zum Vergleich dieser Methoden zu schaffen, erfolgt dieser Schritt erst während der Modellierung. Dies ist allerdings nur unter der Bedingung anwendbar, dass in den Folgeschritten alle Merkmale individuell betrachtet werden und keine Klassifikation von Kombinationen mehrerer Merkmale erfolgt. Hierdurch wird eine Dominanz von Werten mit einem höheren Wertebereich ausgeschlossen (vgl. Kapitel 6.10).

7.1.1. Merkmalsausprägungen mit geringer Häufigkeit

In Kapitel 5.6.4 wurden Satz-Adjektiv-Komposita wie "unsere-Musik-klingt-wie-diese-majestetisch-furchteinflößende-Skyline-aussieht" aufgezeigt. Da diese nur einmalig in der Gesamtheit der Artikel vorkommen, kann das Wort als solches nicht für die Klassifikation verwendet werden. Vielmehr besteht die Vermutung, dass die Tendenz eines Autors, sehr lange Worte zu verwenden, dessen Stil beschreibt (vgl. Kapitel 5.6.4).

Diese geringen Häufigkeiten sind nicht nur bei BoW zu beobachten, sondern bei allen Merkmalen mit einer hohen Anzahl von Dimensionen: BoW, Wort-N-Gramme, Zeichen-N-Gramme, Funktionswörter und Wortarten-N-Gramme. Die Folge der geringen Häufigkeiten sind lange Nullvektoren innerhalb des Datensatzes, welche sich nur unter Umständen von den Algorithmen interpretieren lassen und deshalb entfernt werden müssen (vgl. Kapitel 6.5).

In der Literatur existieren hierfür verschiedene Ansätze: Forman (2002, S. 152) schlägt vor, alle Merkmale mit einer Häufigkeit von kleiner gleich zwei zu entfernen. Grieve (2007, S. 11) hingegen verfolgt den Ansatz alle Merkmale zu eliminieren, welche nicht mindestens in zwei von vierzig möglichen Texten eines Autors vorkommen. Dies würde allerdings bedeuten, dass vor dem Aufteilen der Datensätze in Test- und Trainingssets die Klassifikationslabel verwendet werden. Einen weiteren Ansatz verfolgt Gamon (2004, S. 615) mit der Verwendung von Häufigkeitsgrenzen von fünf bis fünfhundert, wobei dessen Datensatz aus Sätzen und nicht Artikeln besteht, was zu höheren Häufigkeiten führt. Für Zeichen-N-

Gramme werden von Sapkota, Bethard, Montes, & Solorio (2015, S. 96) und Kruczek, Kruczek, & Kuta (2020, S. 5) alle N-Gramme gefiltert, welche weniger als fünfmal in den Artikeln vorhanden sind.

Für Zeichen-N-Gramme wird in dieser Arbeit ebenfalls ein Filter angewandt, der alle Merkmale extrahiert, welche fünfmal oder öfter in allen Artikeln vorhanden sind. Falls dies bei keinen Merkmalen der Fall ist, wird die Zahl so lange um eins reduziert, bis Merkmale zurückgegeben werden. Dies kann vor allem bei einer kleinen Anzahl von Texten für wenige Autoren vorkommen. Alle anderen der erwähnten Merkmale werden dahingehend reduziert, dass alle Ausprägungen herausgenommen werden, welche nicht in mindestens zwei Artikeln im Text vorhanden sind. Im Vergleich zum Zählen der gesamten Vorkommnisse wird sich dadurch versprochen, zusätzlich Eigennamen zu filtern oder Fehler der Textnormalisierung zu umgehen. Es wird angenommen, dass Eigennamen, wie der Titel eines Albums oder Liedtextes, über welches geschrieben wird, keine Aussage über den Autoren geben. Gamon (2004, S. 614) normalisiert hierzu alle Eigennamen zu „Name“. Ausnahmen bestehen hierbei, wenn ein Autor bestimmte Eigennamen vergleicht oder darauf verweist. Beispiele könnten das mehrfache Zitieren einer Person sein. Das Häufige Vorkommen des Jazz Gitarristen „Al Di Meola“ in den Credits eines Textes ist ein typischer Fehler in der Textnormalisierung, welcher auch zu langen Eigennamenketten in den Wortarten-N-Grammen führt (Vgl. Kapitel 5.6.11). Auch Fehler der Normalisierung, wie beispielsweise nicht erkannte Sonderzeichen, können so entfernt werden.

Tabelle 15 zeigt die Veränderung der Anzahl an Merkmalen nach dem Reduzieren der Merkmale mit geringen Häufigkeiten. Ein jedes Merkmal wird um ein Vielfaches reduziert. Besonders auffällig sind hierbei die Wort-N-Gramme, welche weniger als ein Prozent der ursprünglichen Dimensionen besitzen. Das BoW, Zeichen und Wortarten-N-Gramme besitzen insbesondere bei einem großen Wert für „N“ noch eine zu große Anzahl an Merkmalsausprägungen.

Tabelle 15: Dimensionalität Filter geringe Häufigkeit - Eigene Darstellung

	BoW	Funktions-wörter	Wort-N-Gramme				
			n=2	n=3	n=4	n=5	n=6
Vor Filter	70.885	587	399.649	466.107	470.928	470.136	468.275
Nach Filter	25.402	375	28.863	3.400	780	422	328
	Zeichen-Affix-N-Gramme				Zeichen-Wort-N-Gramme			
	n=2	n=3	n=4	n=5	n=2	n=3	n=4	n=5
Vor Filter	1.881	11.019	39.592	91.968	2.053	14.766	50.764	144.305
Nach Filter	1.139	6.234	17.934	31.791	1.191	8.692	14.049	25.571
	Wortarten-N-Gramme				Zeichen-Interpunktion-N-Gramme			
	n=2	n=3	n=4	n=5	n=2	n=3	n=4	n=5
Vor Filter	191	2.158	18.070	88.839	1.142	11.123	70.498	203.003
Nach Filter	184	2.003	14.728	54.980	660	4.321	31.327	64.660

7.1.2. Merkmalsausprägungen mit hoher Häufigkeit

Neben Merkmalsausprägungen mit geringer Häufigkeit existieren auch solche mit einer sehr hohen Häufigkeit. Ohne die vorhandene Varianz der Ausprägungen für die einzelnen Autoren lässt sich keine Individualität anhand dieses Merkmals feststellen. Dies zeigt sich beispielsweise in themenspezifischen Worten, welche in nahezu jedem Text vorkommen, wie „Album“, oder Artikeln[15] (vgl. Kapitel 5.6.1). Aus diesem Grund werden üblicherweise Stoppwörter aus diesen Modellen ausgeschlossen. Im Rahmen der Normalisierung wurden bereits für BoW allgemeine Stoppwörter der deutschen Sprache gefiltert, jedoch existiert für jede Textsammlung eine eigene Liste an Stoppwörtern (vgl. Kapitel 4.3.3). Dies hat zur Folge, dass dennoch Stoppwörter wie beispielsweise englische Artikel („the“) enthalten sind. Dies betrifft alle wortbasierten Merkmale: BoW, Wort-N-Gramme und Funktionswörter. Bei Zeichen oder Wortarten-N-Gramme würden zu viele Merkmale gefiltert werden. Sie besitzen eine deutlich geringere Anzahl von Nullwerten, da meist jede Wortart und eine Vielzahl von Zeichenkombinationen in einem Artikel vorhanden sind.

[15] Unter „Artikel“ ist hier das Geschlechtswort oder der Begleiter zu verstehen.

Ein Maß, um die Häufigkeit eines Merkmals über alle Dokumente beziehungsweise Artikel zu berechnen, ist die Dokumentenhäufigkeit (engl. document frequency, df). Abgeleitet von der inversen Dokumentenhäufigkeit ist sie nach Baeza-Yates & Riberio-Neto (1999) folgendermaßen definiert, wobei N_t für die Anzahl der Dokumente steht, in denen das jeweilige Wort vorkommt und N für die Gesamtheit der Dokumente:

$$df = \frac{N_t}{N}$$

Forman (2002, S. 152) schlägt vor, alle Merkmale mit einer Dokumentenhäufigkeit von größer 0,5 herauszunehmen, da diese so häufig in den Texten vorkommen, dass sie nicht dazu beitragen eine Klasse zu bestimmen. In andere Arbeiten der Autorenschafts-Attribution werden meist nur allgemeine Stoppwörter gefiltert und es existiert kein Verweis auf die Dokumentenhäufigkeit, weshalb in dieser Arbeit ein Schwellenwert der Dokumentenhäufigkeit von 0,5 verwendet wird, um Stoppwörter zu identifizieren.

Oftmals wird in der Literatur zum Ermitteln der Stoppwörter das TD-IDF-Maß verwendet. Für dieses Maß wird die Häufigkeit eines Terms (engl. term frequency) in Relation zur Häufigkeit des Terms in allen anderen Dokumenten gesetzt (engl. inverse document frequency) (Baeza-Yates & Riberio-Neto, 1999). Das bedeutet, dass Wörter, welche mit größerer Häufigkeit, aber in wenigen Dokumenten vorkommen, höher gewichtet werden. Da dies dazu führen würde, dass beispielsweise Eigennamen von Liedern oder Bands sehr hoch gewichtet werden, wird von diesem Maß hier explizit abgesehen.

Tabelle 16 zeigt die Veränderung der Merkmale nach dem Anwenden des Filters. Hierbei ist offensichtlich, dass kaum eine Reduktion der Anzahl der Dimensionen existiert. Nicht ein Wort-N-Gramm wird anhand des Filters selektiert. BoW hingegen verliert sechs Wörter: „the“, „band“, „song“, „fazit“, „album“, und „geben“. Diese Wörter wurden bereits in Kapitel Bag of Words5.6.1 aufgezeigt. Die größte Reduktion erfahren die Funktionswörter. Dies liegt daran, dass hier die allgemeinen Stoppwörter nicht bereits gefiltert wurden. Insgesamt ist durch

die geringe Verminderung der Dimensionen offensichtlich, weshalb der Filter in der meisten Literatur keine Beachtung findet.

Tabelle 16: Dimensionalität Filter hohe Häufigkeit - Eigene Darstellung

	BoW	Funktions-wörter	Wort-N-Gramme				
			n=2	n=3	n=4	n=5	n=6
Vor Filter	25.402	375	28.863	3.400	780	422	328
Nach Filter	25.396	340	28.863	3.400	780	422	328

7.1.3. Aufteilung in Test- und Trainingsdatensätze

Din Datensatz muss auf Überanpassung oder Unteranpassung geprüft werden (vgl. Kapitel 6.8). Hierzu wird ein Modell auf Daten angewandt, welche es noch nie zuvor gesehen hat, den Testdatensatz. Aus diesem Grund werden die Artikel mit einer üblichen 60:40 Verteilung in einen Trainings- und Testdatensatz aufgeteilt. Alle folgenden Filter und Modelle werden auf den Trainingsdatensatz angewandt und mit dem Testdatensatz validiert. Hierzu wird die „train_test_split“ Methode von scikit-learn verwendet (scikit-learn developers, 2020d).

Nach Liu (2019, S. 125) sollten während der Experimente und Exploration immer die gleichen Trainings- und Testdatensätze verwendet werden. Dies wird erreicht, indem ein fester Wert Zufallswert („random_state“) beim Generieren der Daten zugewiesen wird. Auf diese Weise wird sichergestellt, dass der Klassifikator auf einem festen Datensatz funktioniert und eine gute Leistung erbringt, bevor die Zufälligkeit mit einbezogen wird. In dieser Arbeit wird ein Zufallswert von 42 gewählt.

7.1.4. Individuelle relative Häufigkeit

Merkmale, welche durch ihre Häufigkeit gemessen werden, weisen eine starke Abhängigkeit zur Textlänge auf, da in einem längeren Text eine größere Anzahl von Worten, Zeichen oder Sätzen besteht. Dies führt zu einer Korrelation dieser zum Merkmal „Anzahl der Wörter“. Merkmale mit hoher Korrelation sind stärker linear voneinander abhängig und haben daher fast die gleiche Wirkung auf die abhängige Variable. Da sie dadurch keinen Mehrwert aufweisen, werden sie

üblicherweise selektiert, um die Dimensionalität des Modells zu verringern (vgl. Kapitel 6.11). Des Weiteren besitzt die Textlänge eine Abhängigkeit zum Themengebiet des Textes. Eine Korrelation anderer Merkmale zur Textlänge würde folglich in einer Themenabhängigkeit dieser resultieren. Um eine Unabhängigkeit der häufigkeitsbasierten Merkmale zu erreichen, wird anstatt der absoluten Häufigkeit die relative Häufigkeit dieser bestimmt.

Hierbei ist es relevant, in welcher Relation die relative Häufigkeit gebildet wird. Als Beispiel dient hier die relative Häufigkeit der Sonderzeichen, welche in Relation zu drei verschiedenen Größen ermittelt werden kann: Die Häufigkeit im Verhältnis zur Anzahl der gesamten Zeichen, Wörter und Sonderzeichen. Nach Grieve (2007, S. 12) entspricht diese Reihenfolge der Rangfolge der Präzision. Er machte Tests mit vierzig, zwanzig, zehn, fünf, vier, drei und zwei möglichen Autoren, wobei der Rang der Präzision immer gleichblieb. Die einzige Ausnahme bestand darin, dass an einigen Stellen die Präzision der relativen Häufigkeit gesehen zu den Wörtern und Zeichen gleich hoch gemessen wurde. Aus diesem Grund wird für jedes Merkmal eine individuelle relative Häufigkeit gebildet.

Tabelle 17 zeigt die Korrelation zwischen der Textlänge und der absoluten sowie individuellen relativen Häufigkeit der Merkmale gerundet auf sechs Nachkommastellen. Die Korrelation wird durch den durchschnittlichen Korrelationskoeffizienten aller Merkmale zur Textlänge dargestellt. Des Weiteren wird aufgezeigt, im Verhältnis zu welcher Größe die relative Häufigkeit bestimmt wird.

Die geringen Korrelationskoeffizienten könnten als nicht existente Korrelation zwischen den Merkmalen und der Textlänge interpretiert werden. Sie sind aber vielmehr auf die hohe Anzahl an Nullwerten bei den Merkmalen zurückzuführen. So kann der geringe Korrelationskoeffizient des BoW dadurch erklärt werden, dass viele Worte in einem Großteil der Texte nicht vorkommen. Bei diesen Nullwerten errechnet sich ein Korrelationskoeffizient von null. Der Durchschnitt wird dadurch um ein Vielfaches geringer. Deutlich erkennbar ist dies beispielsweise an den Wortlängen mit und ohne Intervalle. Diese unterscheiden sich dadurch, dass durch das Festlegen von Intervallen deutlich weniger

Nullwerte in den Daten existieren. Bei Merkmalen, welche keine Ausprägungen mit Nullwerten besitzen ergibt sich dadurch ein größerer Korrelationskoeffizient. Beispiele sind hier Wortarten, Ziffern oder Wortlängen mit Intervallen.

Tabelle 17: Korrelationskoeffizienten relative und absolute Häufigkeit - Eigene Darstellung

Merkmal	n	Korr. abs. Häufigkeit	relativ zu	Korr. rel. Häufigkeit
Bag-of-Words	-	0,024017	Wörter im Text	0,006914
Wort-N-Gramm	2	0,013158	Wörter im Text	0,007160
	3	0,013453		0,009765
	4	0,021214		0,019265
	5	0,026615		0,025629
	6	0,027679		0,027373
Sonderzeichen	-	0,235031	Zeichen im Text	0,038921
Zeichen Affix N-Gramm	2	0,169051	Zeichen im Text	0,004613
	3	0,086297		0,004558
	4	0,055543		0,004277
	5	0,044095		0,004636
Zeichen-Wort-N-Gramm	2	0,179449	Zeichen im Text	0,005455
	3	0,087735		0,003402
	4	0,054362		0,003984
	5	0,041917		0,004213
Z.-Interpunktion-N-Gramm	2	0,094730	Zeichen im Text	0,004506
	3	0,056235		0,006555
	4	0,038746		0,004809
	5	0,031682		0,004397
Ziffern	-	0,290216	Zeichen im Text	0,074218
Funktionswörter	-	0,108312	Wörter im Text	0,005086
Wortarten	-	0,745296	Wörter im Text	0,005627
Wortarten-N-Gramm	2	0,338783	Wörter im Text	-0,001091
	3	0,115100		0,001481
	4	0,045585		0,003963
	5	0,026509		0,005966
Satzanfänge und -endungen	-	0,263779	Sätze im Text	-0,010207
Wortlänge ohne Intervalle	-	0,298254	Wörter im Text	0,001155
Wortlänge mit Intervallen für Zahlen größer 20	-	0,638194	Wörter im Text	-0,003325

Dennoch zeigt sich ausnahmslos ein geringerer Korrelationskoeffizient nach der Berechnung der individuellen relativen Häufigkeit. Auch bei den vorher erwähnten größeren Korrelationskoeffizienten nähert sich dieser der Null an. Folglich wird die Korrelation zwischen den Merkmalen und der Textlänge durch das Bestimmen der individuellen relativen Häufigkeit verringert oder sogar aufgehoben.

7.2. Merkmalsauswahl

Im Datenverständnis zeigten sich einige Merkmale mit einer sehr großen Anzahl an Dimensionen. Nach der Vorverarbeitung der Daten existieren noch immer Merkmale mit 10.000 bis 65.000 Dimensionen innerhalb der verwendeten Textsammlung. Im Vergleich dazu führen Abbasi & Chen (2008, S. 9) eine Anzahl von weniger als 500 Merkmalen für einige Arbeiten an. Es existieren allerdings auch Modelle, welche mehr als tausend Merkmale verwenden, wie beispielsweise Narayanan, et al (2012, S. 306) entwickelte Verfahren. Dabei gilt, die Anzahl der Merkmale immer in Relation zur Anzahl der Texte zu sehen. Aus einer größeren Anzahl von Texten folgt grundsätzlich eine größere Anzahl an Merkmalsausprägungen. Wie bereits im vorherigen Kapitel beschrieben wurde, führt eine größere Anzahl an Merkmalen nicht immer zu einem besseren Ergebnis. Unter Umständen kann sich dieses sogar verschlechtern (vgl. Kapitel 6.11). Aus diesem Grund gilt es, die Merkmale mithilfe der in (ebd.) vorgestellten Methoden zu reduzieren. Dabei sollen irrelevante Ausprägungen, welche die Komplexität des Modells ohne einen erkennbaren Mehrwert erhöhen, gefiltert werden.

In Kapitel 6.11 wird der wesentliche Unterschied zwischen Filtern und Wrappern verdeutlicht. Filter basieren auf statistischen Maßen, sind deutlich schneller in der Berechnung und haben eine geringe Neigung zu einer Überanpassung. Außerdem existiert kein allgemeines Optimum für die Anzahl der selektierten Merkmale, wodurch unklar ist, wie viele Merkmale ausgewählt werden sollen. Wrapper hingegen liefern für einen festgelegten Algorithmus die Merkmale für ein optimales Ergebnis, sind aber sehr rechenintensiv und haben eine größere

Neigung zu einer Überanpassung. Zudem ermöglichen bekannte Wrapper-Bibliotheken nur unter Umständen eine Hyperparameteroptimierung, wodurch das Ergebnis für SVC oder KNN zweifelhaft wäre. Werte, welche für einen SVC mit linearem Kernel gute Ergebnisse liefern, müssen das nicht für einen SVC mit polynomialen Kernel ebenso tun und umgekehrt. Um die positiven Eigenschaften der Verfahren zu kombinieren, wurde in dieser Arbeit eine Kombination beider Verfahren aufgestellt, welche im folgenden Kapitel näher erläutert wird.

Eine weitere Reduktion der Merkmale wäre mit dem Durchführen einer Hauptkomponentenanalyse möglich. Da diese auf der Korrelation der einzelnen Merkmale basiert, müssten die Ausgangsmerkmale die Gesamtheit der im Modell verwendeten Daten umfassen. Selbst nach der folgenden Reduktion der Variablen das Verhältnis der Stichproben zu den Variablen im späteren Modell ungefähr 1:6. Dieses ist um ein Vielfaches geringer als beschriebene Faustregeln von 3-6:1, 5:1 oder 10:1 (vgl. Kapitel 6.11.4). Tests ergaben, dass die Korrelationsstruktur nicht ausreicht, um dieses geringe Verhältnis auszugleichen und die Ergebnisse der PCA deshalb nicht zielführend sind. Aufgrund der eindeutigen Ergebnisse wird davon abgesehen, diese im Folgenden aufzuzeigen und es wird entsprechend auf ein Anwenden der PCA im Rahmen dieser Arbeit verzichtet.

7.2.1. Merkmalsauswahl durch einen iterativen Filter

Der in dieser Arbeit entwickelte iterative Filter ist eine Kombination aus Filtern und Wrappern. Im ersten Schritt erfolgt eine Einstufung der Merkmale anhand eines Filters. Durch eine Vorwärts-Auswahl wird diese Einstufung, beginnend mit den besten Merkmalen, durchlaufen. Nach einer optionalen Hyperparameteroptimierung wird eine Bewertung für die jeweiligen Verfahren durchgeführt. Anhand dieser Bewertung wird der optimale Punkt für die jeweiligen Klassifikatoren ausgewählt und die Merkmale anhand dieses Punktes selektiert. Dieser Ablauf wird für jedes in Kapitel 7.1.1 genannte Merkmal mit einer großen Anzahl von Ausprägungen durchgeführt. Die einzelnen Schritte werden in einem Ablaufdiagram in Bild 15 aufgezeigt und im Folgenden näher beschrieben.

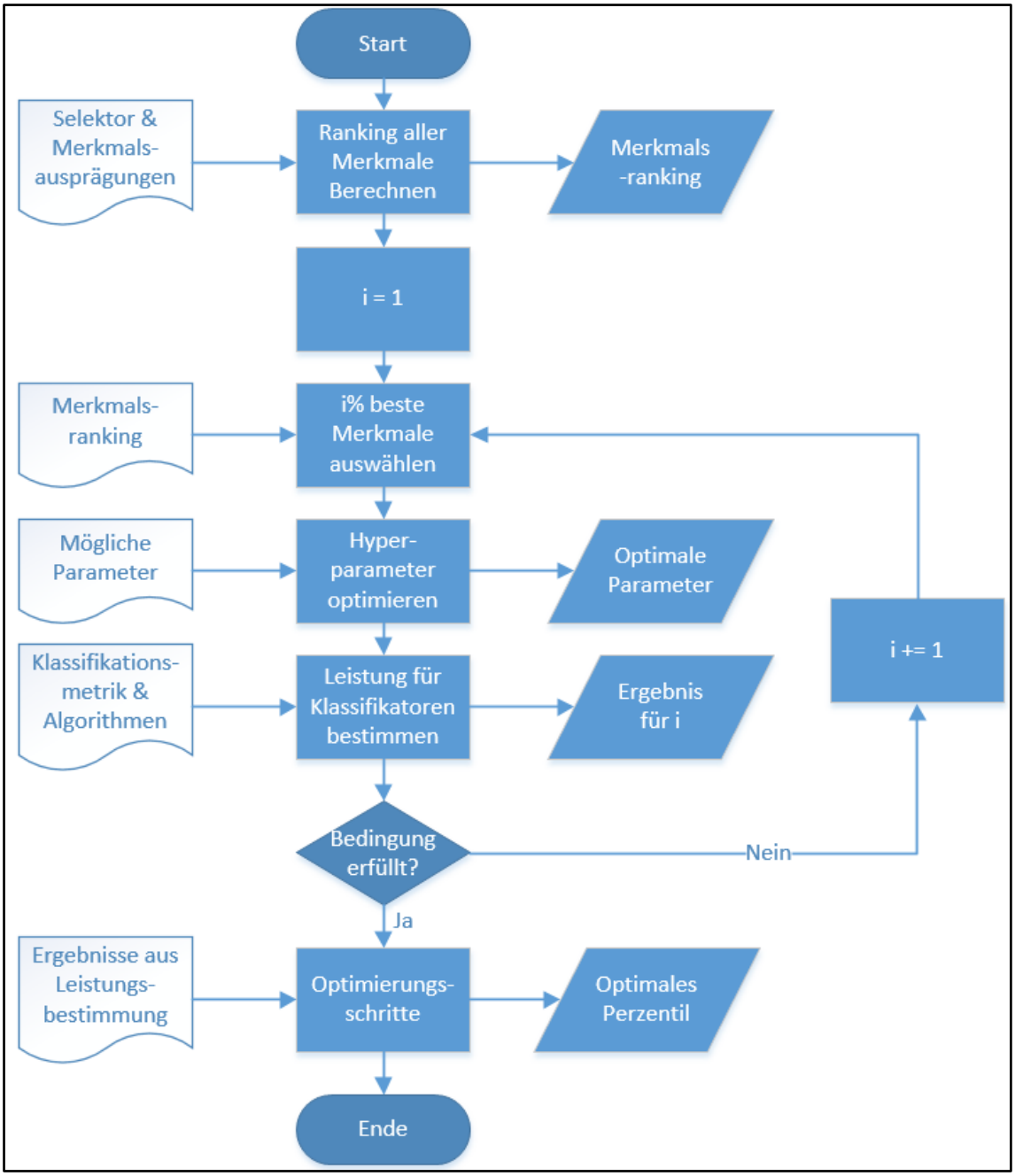


Bild 15: Ablaufdiagram iterativer Filter - Eigene Darstellung

Im ersten Schritt wird ein Filter auf die Daten angewandt, bei dem alle Merkmale beibehalten werden, wodurch eine Rangfolge dieser folgt. Hierfür wird die Bibliothek „SelectKBest“ von scikit-learn verwendet (scikit-learn developers, 2020b). Als Selektionsmetrik wird der Informationsgewinn verwendet (vgl. Kapitel 6.11.1). Dieser zeigte sich in verschiedenen Studien als effiziente Methode. Beispiele hierfür sind Forschingsarbeiten von Forman (2002), Efron, Marchionini,

& Zhiang (2003) und Koppel & Schler (2003). Insbesondere wird hier die Effizienz im Vergleich zu suchbasierten Methoden (Liu & Dash, 1997) und seine Stärke in Multi-Label-Klassifikationen (Yang & Pedersen, 1997) hervorgehoben. (Abbasi & Chen, 2008, S. 14)

Für jeden Klassifikator erfolgt nun eine Vorwärts-Auswahl über die ermittelte Rangfolge. Als Schrittgröße wird ein Prozent der Merkmale festgelegt. Diese Größe bietet einen guten Kompromiss aus benötigter Rechenleistung und Genauigkeit.

Innerhalb eines jeden Iterationszyklus wird eine Hyperparameteroptimierung mit einer zehnfachen Kreuzvalidierung auf die Trainingsdaten durchgeführt. Ziel ist es, für jeden Merkmalsvektor ein bestmögliches Ergebnis zu erzielen. Da auch in der späteren Modellierung eine Hyperparameteroptimierung erfolgt, wäre das Ergebnis sonst nicht aussagekräftig. GNB (Gaussian Naïve Bayes) besitzt keine Hyperparameter, welche optimiert werden können.

Anschließend an die Optimierung der Parameter können die Werte für die einzelnen Klassifikatoren berechnet werden. Wurde eine Hyperparameteroptimierung angewandt, entspricht dies dem dort errechneten Wert. Im Falle von GNB wird anlehnend dazu der durchschnittliche Wert der zehnfachen Kreuzvalidierung ermittelt. Als Klassifikationsmetrik wird in dieser Auswertung die Genauigkeit[16] als Metrik ausgewählt. Sie beschreibt die allgemeine Leistung des Verfahrens und ermöglicht eine einfache Interpretation der Kurven. Die Anforderung einer Balance der Klassen ist ebenfalls erfüllt.

Für den Verlauf der Genauigkeitskurven über die Iterationen existieren drei gängige Muster, welche hier anhand der Zeichen-Affix-3-Gramme (vgl. Bild 16) und Funktionswörter (vgl. Bild 17) aufgezeigt werden. Der erste Kurvenverlauf wird durch einen anfänglich streng monoton steigenden oder monoton fallenden Verlauf in Folge einer Konvergenz gegen einen Genauigkeitswert beschrieben. Diese Verläufe sind üblicherweise bei den Klassifikationen mittels SVC und KNN

[16] In diesem speziellen Fall mit einer Balance der Klassen entsprechen die Werte der Genauigkeit dem F1-Wert.

zu erkennen und werden anhand deren Kurven in Bild 16 und Bild 17 dargestellt. Der zweite Kurvenverlauf wird durch eine anfangs streng monoton steigende Kurve bis zu einem abflachenden Höhepunkt beschrieben, woraufhin die Kurve monoton abfällt. Der dritte Verlauf unterscheidet sich vom zweiten dahingehend, dass dieser nach dem Erreichen des Höhepunktes weiterhin schwankt und für einzelne Prozentwerte weiterhin kleinere Hochpunkte besitzt. Der zweite und dritte Verlauf lässt sich meist dem GNB zuordnen und kann durch die beim Hinzukommen von Nullwerten sinkende Wahrscheinlichkeit erklärt werden. Für die einzelnen Hochpunkte besteht die Vermutung, dass hier Merkmale ausgewählt werden, die eher einer Normalverteilung ähneln (vgl. Kapitel 6.5.2). Verlauf Zwei wird durch die GNB-Kurve der Zeichen-Affix-3-Gramme und Verlauf Drei durch die GNB-Kurve der Funktionswörter dargestellt.

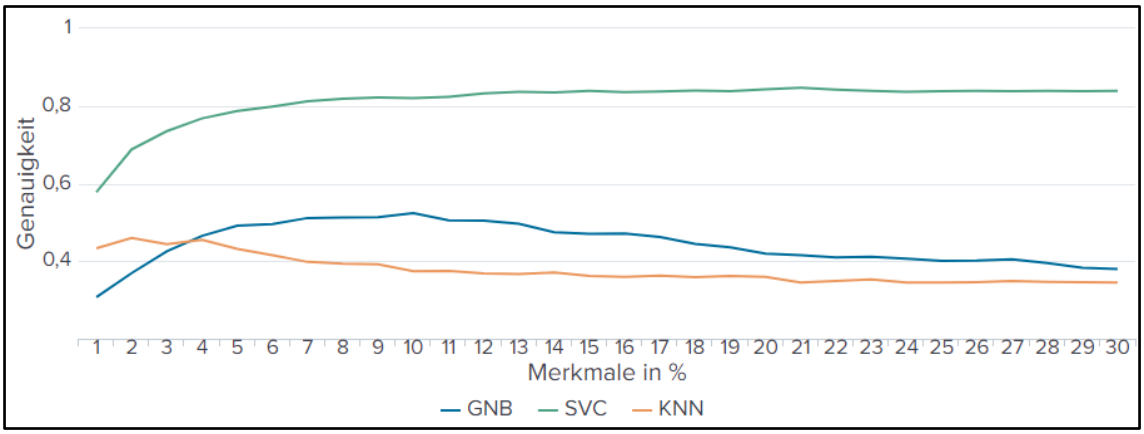


Bild 16: Genauigkeitskurve Zeichen-Affix-3-Gramme - Eigene Darstellung

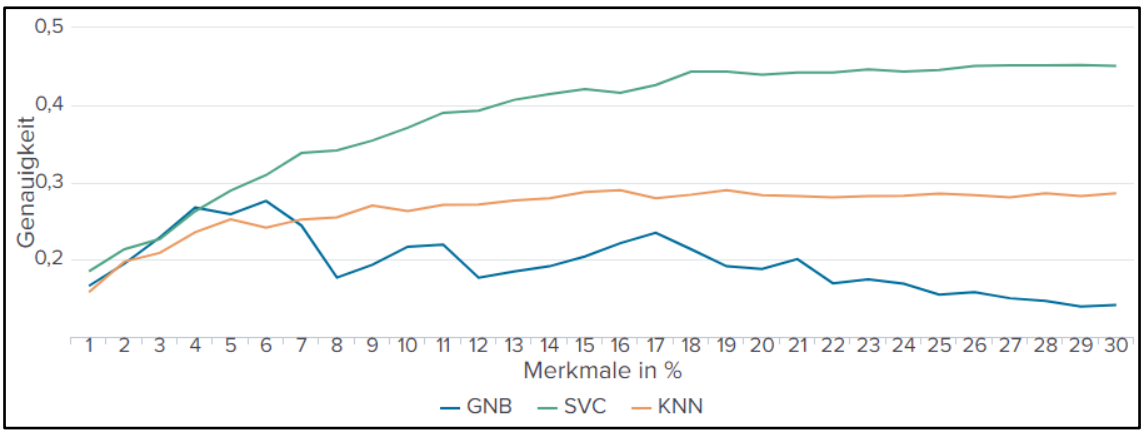


Bild 17: Genauigkeitskurve Funktionswörter - Eigene Darstellung

Aus diesen Kurven ergeben sich entsprechend drei Abbruchbedingungen der Iteration:

1. Die Kurve konvergiert gegen einen Wert und alle fünf Folgepunkte besitzen eine Abweichung von 0.5% oder weniger Genauigkeit zu einem Punkt.
2. Der Verlauf der Kurve ist über alle fünf Folgepunkte hinweg monoton fallend.
3. Alle fünf Folgepunkte eines Punktes besitzen mindestens 2% Genauigkeit weniger.

Das Optimum wird als der Punkt bestimmt, für welchen eine dieser Bedingungen eintritt. Falls keine dieser Bedingungen eintritt, wird die maximale Anzahl der Iterationen auf 50% der Werte festgelegt. Dieser Wert entspricht der minimalen Reduktion der Merkmale, welche erreicht werden soll.

Das bestimmte Optimum entspricht nicht immer dem tatsächlich optimalen Perzentil der Merkmale, weshalb das Verfahren um zwei Optimierungsschritte ergänzt wird. Im ersten Schritt wird das neue Optimum als das Maximum der Kurve definiert, wenn dieses einen geringeren Anteil der Merkmale als das bisherige Optimum verwendet. Im zweiten Schritt wird ein Kompromiss zwischen der Genauigkeit und der Anzahl verwendeter Merkmale ermittelt. Der Punkt mit den wenigsten verwendeten Merkmalen, welcher mindestens 2% weniger Merkmale und maximal weniger als 1% Genauigkeitsverlust aufweist, wird als das neues Optimum festgelegt. Existiert ein solcher Punkt nicht, bleibt das Optimum unverändert.

Von einer weiteren Iteration für ein genaueres Ergebnis wird abgesehen. Dies würde zu einer Reduktion von maximal einem Prozent der Merkmale führen, aber dabei bis zu einem Prozent allein durch den zweiten Optimierungsschritt verlieren. Des Weiteren ist anzumerken, dass durch das Verwenden der zehnfachen Kreuzvalidierung ein gewisser Zufall bei der Auswahl der Artikel besteht. Dieser Zufall führt zu grob abweichenden Ergebnissen bei mehrfachen Durchläufen des Verfahrens.

7.2.2. Bewertung der Ergebnisse des iterativen Filters

Nachdem dem Abschluss des Verfahrens erfolgt eine Auswertung der Daten. Hierzu werden in Tabelle 18 die Anzahl der Dimensionen als auch die Genauigkeitswerte der Klassifikatoren, vor und der Filterung der Daten, aufgezeigt. Für eine erleichterte Übersicht werden Genauigkeitswerte nach dem Filtern farblich hinterlegt. Dabei wird folgendes Schema verwendet:

- Grün: Gleichbleibende oder verbesserte Genauigkeit
- Orange: Maximal 5% Verlust an Genauigkeit
- Rot: Mehr als 5% Verlust an Genauigkeit

Dabei ist zu beachten, dass die farbliche Hinterlegung lediglich zur Orientierung dient und eine Bewertung immer in Abhängigkeit zu der Anzahl der reduzierten Dimensionen erfolgt. Grüne Werte sind als positiv zu erachten, da die Genauigkeit beim Verlust von Dimensionen gesteigert werden konnte, orange steht für einen moderaten Verlust, welcher bei der Merkmalsreduktion als Kompromiss zwischen Genauigkeit und Dimensionalität grundsätzlich annehmbar ist. Rote Werte müssen individuell betrachtet werden. Der große Genauigkeitsverlust kann nur durch das Verwenden eines Bruchteils der Werte gerechtfertigt werden. Weiterhin gilt anzumerken, dass eine Verminderung der Genauigkeit eines Merkmals nicht gleichzusetzen ist mit einer Verminderung der Genauigkeit des gesamten Modells. Durch redundante Informationen in den einzelnen Merkmalen kann der Verlust durchaus geringer ausfallen.

Die ersten Auffälligkeiten sind die schlechten Genauigkeitswerte des BoW-Modells und der Wort-N-Gramme. BoW besitzt in etwa 25.000 Dimensionen. Ein Verlust von maximal 6% bei dem Verwenden von einem Fünftel der Merkmale bedeutet in anderen Worten, dass 6% der Genauigkeit in 83% der Merkmale stecken. Ein Verzicht auf diese Anzahl von Merkmalen mit diesem geringen Informationsgehalt wird als angemessen betrachtet. Hervorzuheben sind hier die Ergebnisse des SVC, welcher nur 12% der ursprünglichen Ausprägungen verwendet und dabei nur 1,5% an Genauigkeit verliert. Eine weniger deutliche Reduktion ist beim KNN zu beobachten, jedoch werden hier in etwa ein Drittel

der Merkmale bei einem Verlust von etwa 3% Genauigkeit verwendet. Insgesamt gilt hier das Ergebnis trotz der Genauigkeitsverluste als positiv zu bewerten.

Tabelle 18: Merkmale vor und nach dem iterativen Filter - Eigene Darstellung

	n	Vor Filter				Nach Filter					
		Dim.	Genauigkeit (%)			Dimensionen			Genauigkeit (%)		
			GNB	SVC	KNN	GNB	SVC	KNN	GNB	SVC	KNN
Bag-of-Words	-	25.396	56	72,3	36,3	4.318	3.048	8.891	50	70,8	33,2
Wort N-Gramm	2	28.863	46,3	25,6	6,3	9.525	7.793	866	34	20,5	7,6
	3	3.400	21,8	12,3	10,3	1.122	272	714	13,8	5,1	9
	4	780	11	6,7	9,2	8	8	8	3,1	3,1	4,3
	5	422	7,1	4,6	6,4	4	4	4	3	3	4,2
	6	328	3,4	3,4	4,6	3	3	3	3	3	4,2
Zeichen-Affix-N-Gramm	2	1.139	26,4	73,4	35,5	194	239	68	54,4	74,5	52,8
	3	6.234	27,3	84,4	35,5	623	810	125	46,5	83,6	47,1
	4	17.934	36,3	86,3	45,4	538	1.793	179	59,7	84	45,1
	5	31.791	44,8	88,4	60,4	636	2.225	8266	57	86,7	55,6
Zeichen-Wort-N-Gramm	2	1.191	22,1	61,8	20,3	274	524	60	39,5	61,1	20,8
	3	8.692	25,3	82,3	28,2	782	869	87	53	77,9	39
	4	14.049	35,1	82,5	45,3	627	3.446	11278	55,5	80,1	38
	5	25.571	35,2	74,4	52,6	647	6.466	9052	42,6	77,6	34,3
Z.-Interpunkt.-N-Gramm	2	660	23,8	56,3	33,1	20	152	40	25	53,5	43,7
	3	4.321	30,3	66	18	346	519	86	28,6	61,7	30,6
	4	31.327	38,1	63,5	11,5	2.810	2.388	140	34,3	62,2	25
	5	64.660	39,4	61,8	10,4	2.557	2.557	256	30,6	60,4	24,6
Funktionsw.	-	340	22	47,6	28,3	41	79	52	32,2	44,1	27,6
Wortarten N-Gramm	2	184	14,5	44,8	26,4	79	57	39	38,1	35,5	23,3
	3	2.003	9,2	45,8	27,1	260	361	721	35	40,7	25,5
	4	14.728	5,7	39,4	25,5	147	1.473	3.387	23,9	36,8	26,2
	5	54.980	6,8	25,4	21,2	1.649	3.849	2.199	8,9	24,1	9,2

Negativ hingegen erweisen sich die Ergebnisse der Wort-N-Gramme mit einem Wert von N größer als drei für GNB. Hier sinkt die Genauigkeit um 7,9%. Insgesamt werden aus den Wort-N-Grammen mit einem N größer als drei nur sehr wenige Merkmale mit sehr geringen Genauigkeitswerten selektiert. Durch die sehr geringen Genauigkeitswerte, welche sich auch bei einer anderen Anzahl an Ausprägungen nicht verändern, greift die Abbruchbedingung des Filters

bereits bei einem Prozent der Merkmale. Aus diesem Grund besteht die Annahme, dass diese Merkmale insgesamt keinen Mehrwehrt für die Autorschaftsattribution bieten. Deshalb werden Wort-4 bis Wort-6-Gramme an dieser Stelle nicht weiter betrachtet und im späteren Verlauf geprüft, ob ein Verwenden dieser Merkmale als sinnvoll erachtet wird.

Bei Wort-N-Grammen, mit einem N kleiner gleich drei wird das Verhältnis der verlorenen Dimensionalität im Vergleich zur Genauigkeit akzeptiert. Eine Ausnahme stellen hierbei Wort 3-Gramme unter der Klassifikation von GNB dar. Der Verlust von 8% Genauigkeit wird als zu groß gesehen. Aus diesem Grund wird hier der Filter für dieses Modell nicht angewandt und in folgenden Kapiteln der gesamte Datensatz verwendet.

Ebenfalls stechen die hohen Genauigkeitsverluste des SVC bei den Wortarten 2-Grammen und 3-Grammen hervor. Ein Genauigkeitsverlust von 9% beim Verzicht auf 127 Merkmale erscheint bei den Wortarten 2-Grammen nicht angemessen. Aus diesem Grund wird auch hier der Filter für SVC nicht angewandt. Bei den 3-Grammen entspricht der Verlust von 5,1% Genauigkeit 1.642 Merkmalen. Dies wird hingegen als hinnehmbarer Verlust erachtet.

Zuletzt gilt es die hohen Genauigkeitsverluste des GNB bei Zeichen-Interpunktion-5-Grammen und KNN bei Zeichen-Wort-4-Grammen und 5-Grammen zu erwähnen. Beachtet man wiederum die dadurch reduzierten Dimensionen, werden diese ebenfalls als positive Ergebnisse interpretiert.

Insgesamt erreichte das Verfahren eine starke Reduktion der Dimensionen unter annehmbaren Verlusten der Genauigkeit. Teilweise konnte diese sogar verbessert werden.

7.3. Merkmalsvarianten

Im Datenverständnis haben sich bereits verschiedene Varianten für einige Merkmale gezeigt. Diese beschreiben, in welcher Form die gesuchte Information des Merkmals aus den Texten extrahiert werden können. Dazu gehören verschiedene Längen für N-Gramme, Durchschnittswerte, Klassenbildungen bei

Wortlängen oder das Zusammenfassen von einzelnen Ziffern auf ihre gesamtheitliche Häufigkeit.

In den folgenden Kapiteln werden die Varianten der Merkmale aufgezeigt und bewertet. Das Ziel ist es eine Merkmalsvariante für den weiteren Verlauf der Arbeit zu bestimmen.

Eine Bewertung der Methoden erfolgt dabei über das Ermitteln des F1-Wertes als Gleichgewicht zwischen Genauigkeit und Trefferquote über eine verschiedene Anzahl von Texten für jeden der 25 Autoren. Von einer Variation der Autorenanzahl wird abgesehen, da davon auszugehen ist, dass die Bewertung im Allgemeinen für alle Verfahren besser ausfällt, dabei aber die Gefahr einer nicht repräsentativen Menge an Variationen für Autoren existiert. Bei der Textanazahl handelt es sich um die Gesamtzahl der Texte, auf welche ein 60/40 Trainings-/Testsplit durchgeführt wird (vgl. Kapitel 6.8.1). Nach einem Training inklusive einer Hyperparameteroptimierung wird der F1-Wert anhand der Testdaten für die einzelnen Klassifikatoren bestimmt.

Für eine valide Datengrundlage müssen bei allen Daten die vorherigen Schritte der Vorverarbeitung erfolgen. Dies umfasst das Durchlaufen des iterativen Filters für die in Kapitel 7.2 erwähnten hochdimensionalen Merkmale.

7.3.1. Extraktionsmethode Wortlänge

In Kapitel 5.6.4 wurde das grundlegende Merkmal „Wortlänge“ beschrieben. Dabei wurden unterschiedliche Methoden aufgezeigt, um die Eigenschaft der Wortlänge eines Textes abzubilden beziehungsweise zu extrahieren: Das Zählen aller Wortlängen und die durchschnittliche Länge der Worte. Im Vergleich dieser Methoden wurden die folgenden, sich ausschließenden, Thesen aufgestellt: Der Durchschnitt

- liefert durch den geringeren Informationsgehalt schlechtere Ergebnisse.
- verringert das Bias und dadurch bessere Ergebnisse.
- liefert ungefähr gleiche Ergebnisse mit einem geringeren Rechenaufwand.

Weiterhin besteht die Vermutung, dass die Tendenz von Autoren, lange Worte zu verwenden, mehr Informationsgehalt besitzt, als das Zählen einzelner Wortlängen. Um dies abzubilden, wurde die Einführung von Intervallen vorgeschlagen. In der Fachsprache wird dies auch als Klasseneinteilung (engl. binning) bezeichnet. Üblicherweise wird eine Klasseneinteilung verwendet, um durch eine Verallgemeinerung die Chance auf eine Überanpassung zu verringern (Liu Y. H., 2019, S. 34).

Für die Bewertung werden fünf verschiedene Verfahren definiert:

1. Durchschnittliche Wortlänge
2. Alle Längen einzeln
3. Längen von 1-20 einzeln. Eine Klasse für >20
4. Längen von 1-10 einzeln. Klassen für 10-20 und >20
5. Längen von 1-20 einzeln. Klassen für 20-30 und >30

Vorangestellt an die Bewertung erfolgt hier eine Standardisierung der Daten, da sonst durch die geringen Unterschiede zwischen den durchschnittlichen Wortlängen eine Hyperparameteroptimierung zu einem sehr hohen Rechenaufwand führen kann. Eine Dominanz kleinerer Wortlängen mit einer hohen Anzahl von Werten wird dadurch ebenfalls vermieden.

Die Ergebnisse der einzelnen Klassifikatoren werden in Bild 18 aufgezeigt. Anhand der Kurven lässt sich kein eindeutig bestes Verfahren ableiten. Die F1-Werte liegen bei allen Verfahren auf die Gesamtheit betrachtet sehr nah beieinander. Die durchschnittliche Wortlänge zeigt aber grundsätzlich die schlechtesten und das Ermitteln aller Wortlängen die zweitschlechtesten Ergebnisse. Die einzige Ausnahme besteht in der Klassifikation mit GNB, in der die durchschnittliche Wortlänge besser zu bewerten ist als das 2. Verfahren. Dies ist wahrscheinlich auf die hohe Empfindlichkeit des GNB gegenüber von Nullreihen zurückzuführen. Verfahren 3 zeigt im Vergleich meist sehr gute Ergebnisse, welche bei einer Anzahl von nur 10 Texten pro Autor stark abfallen.

Verfahren 4 und 5 liefern ebenfalls gute Ergebnisse. Diese erscheinen allerdings über die Klassifikatoren hinweg inkonsistent. Verfahren 4 besitzt die besten

Ergebnisse für KNN und GNB, aber nur sehr schlechte Werte für wenige Texte bei SVC.

Aufgrund seiner Konsistenz gilt Verfahren 3 als bestes zu bewerten. Die abfallenden Werte bei einer sehr geringen Anzahl an Texten wird dadurch erklärt, dass diese nicht als Trainingsmenge genügen, um zwischen der Vielzahl an Autoren zu unterscheiden.

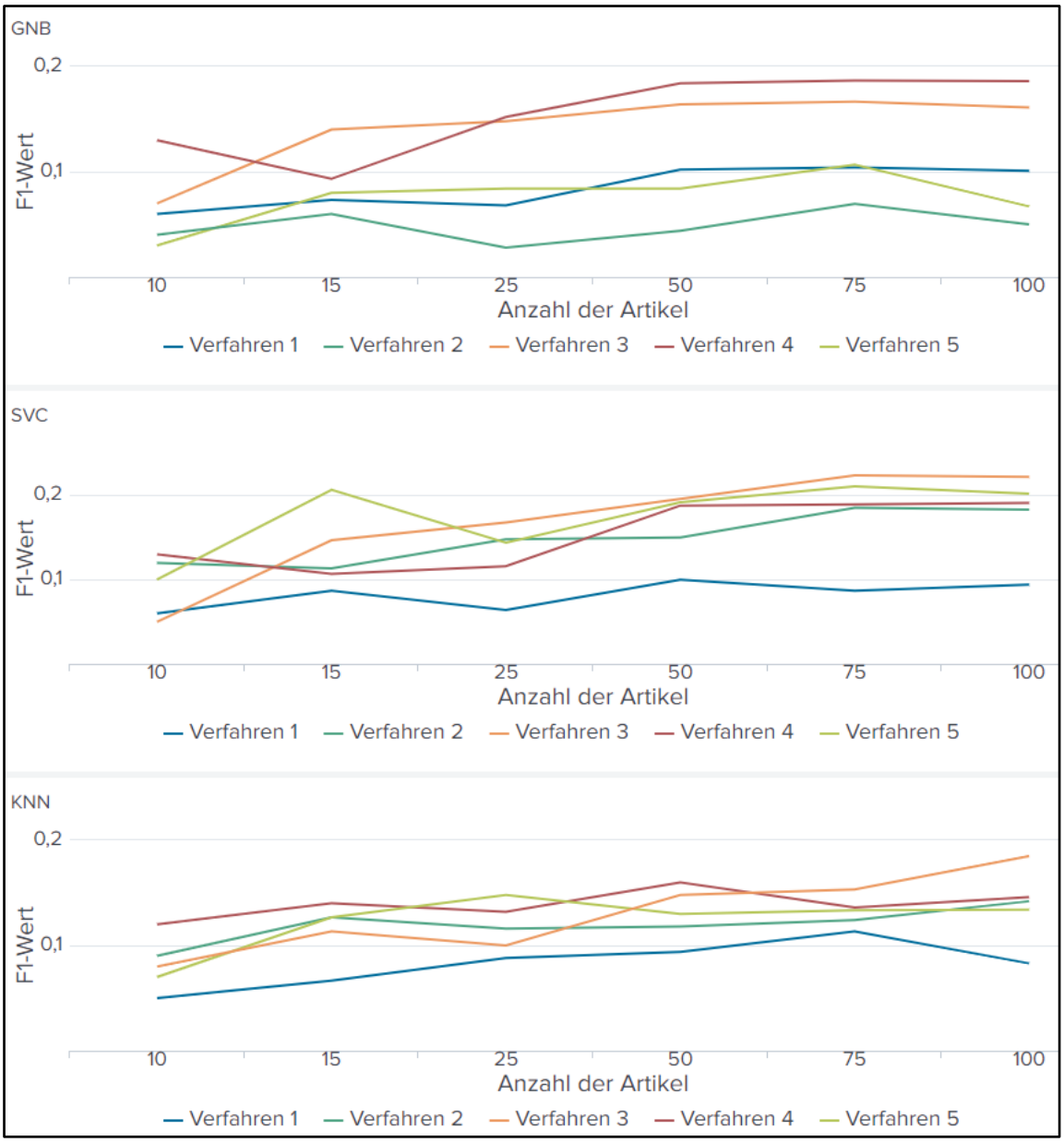


Bild 18: Bewertung Extraktionsmethode Wortlänge - Eigene Darstellung

7.3.2. Extraktionsmethode Ziffern

In Kapitel 5.6.8 wurde bereits aufgezeigt, dass in der Summe aller Artikel ein Unterschied zwischen den verwendeten Ziffern der Autoren existiert. Hierfür wurde eine Hitzekarte aufgezeigt, welche die 25 Autoren den Ziffern von 0-9 gegenüberstellt und die Häufigkeit ihrer Verwendung aufzeigt. Dabei wurde die Vermutung aufgestellt, dass Autoren eher alle Ziffern besonders häufig oder selten verwenden. Ein häufiges Verwenden einer und seltenes Verwenden einer anderen Ziffer wurde nur in geringen Fällen erkannt. Aus diesem Grund gilt es zu prüfen, ob ein informationeller Mehrwert dadurch entsteht, alle Ziffern auf ihre jeweilige Häufigkeit zu prüfen oder ob es ausreicht, lediglich die Häufigkeit aller Ziffern zu bestimmen.

Bild 19 zeigt die Ergebnisse für die einzelnen Klassifikatoren anhand eines Liniendiagrams. Außer im Falle von GNB scheinen die einzelnen Ziffern einen erhöhten Informationsgehalt zu besitzen und bessere Ergebnisse zu liefern. Allerdings zeigen sich nur geringe Unterschiede zwischen den F1-Werten. Mit einer maximalen Abweichung von 0,05 werden die neun zusätzlichen Merkmale nicht gerechtfertigt. Aus diesem Grund werden die Ziffern im weiteren Verlauf anhand ihrer gesamten und nicht der jeweiligen Häufigkeit gemessen.

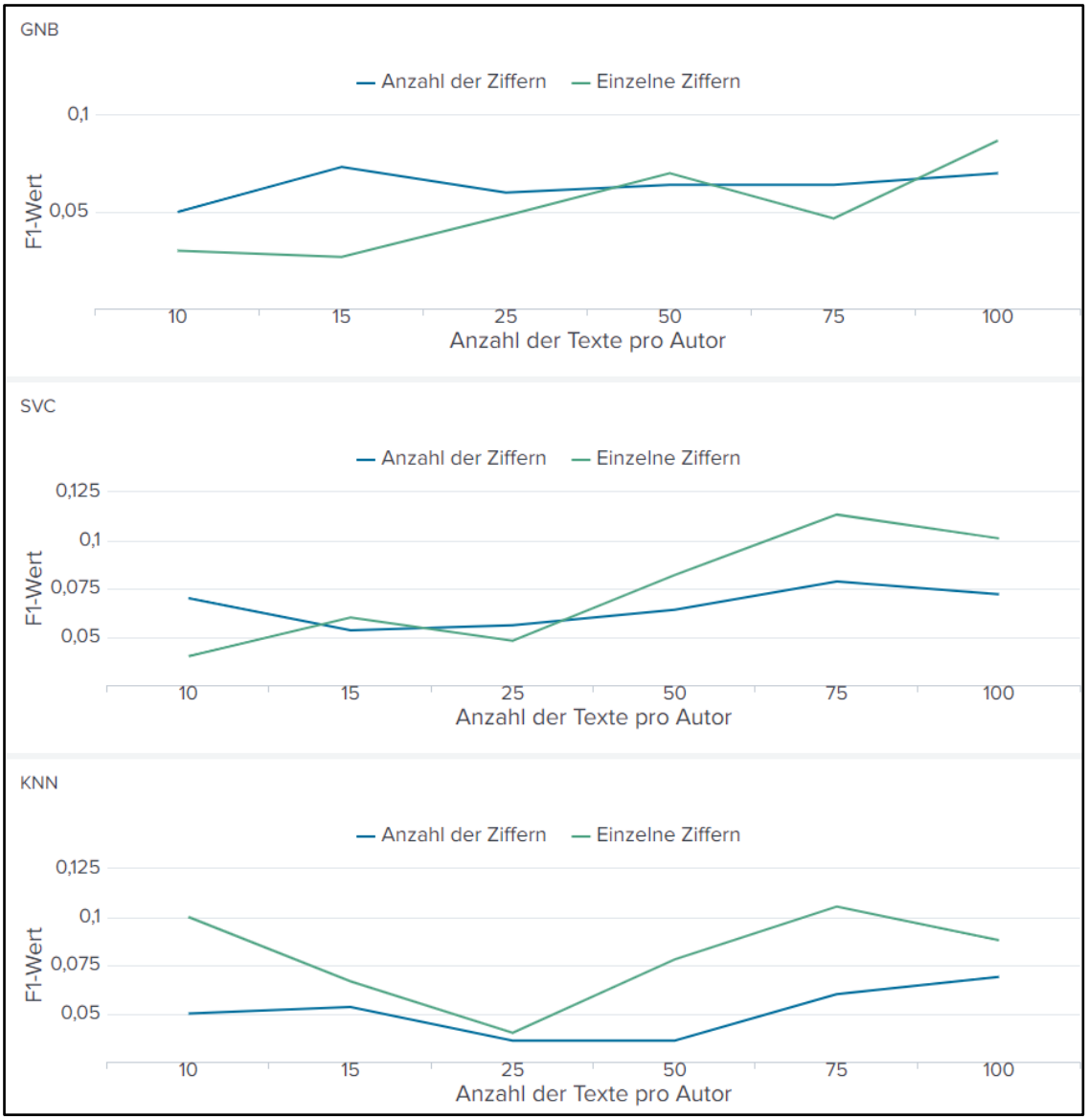


Bild 19: Bewertung Extraktionsmethode Ziffern - Eigene Darstellung

7.3.3. Bestimmen der Länge von Wort-N-Grammen

Um die Reihenfolge und den Kontext der verwendeten Wörter eines Autors zu erfassen, wurden in Kapitel 5.6.11 Wort-N-Gramme vorgeschlagen. Dabei gilt es zu prüfen, welche Länge von n die besten Ergebnisse verspricht.

In Kapitel 7.2.2 äußerte sich bereits die Vermutung, dass die Genauigkeitswerte für große N-Gramme (N größer als drei) sehr ernüchternd sind und ein Filtern dieser Merkmale nicht sinnvoll erscheint. Die Datengrundlage entspricht demnach den ungefilterten Daten mit individueller relativer Häufigkeit. Um die

Wort-N-Gramme mit einem großen N direkt mit den Wort-N-Grammen mit einem kleinen N vergleichen zu können, umfasst der erste Schritt ein Bewerten dieser. Tabelle 19 zeigt die Ergebnisse der Bewertung.

Tabelle 19: Bewertung Wort-N-Gramme mit N von 4-6 - Eigene Darstellung

Art.	GNB			SVC			KNN		
	4-Gramm	5-Gramm	6-Gramm	4-Gramm	5-Gramm	6-Gramm	4-Gramm	5-Gramm	6-Gramm
10	0,0700	0,0700	0,0700	0,0700	0,0700	0,0700	0,0700	0,0700	0,0700
15	0,0800	0,0667	0,0667	0,0533	0,0467	0,0467	0,0733	0,0600	0,0600
25	0,0720	0,0680	0,0520	0,0800	0,0640	0,0520	0,0640	0,0680	0,0520
50	0,0760	0,0560	0,0460	0,0640	0,0560	0,0400	0,0720	0,0540	0,0460
75	0,0840	0,0573	0,0453	0,0667	0,0493	0,0440	0,0733	0,0573	0,0440
100	0,1040	0,0650	0,0500	0,1010	0,0560	0,0450	0,0860	0,0600	0,0490

Entsprechend der Vermutung ergeben sich als Ergebnis der Bewertung niedrige F1-Werte. Dabei ist zu beobachten, dass die Bewertung konstant bei allen Klassifikatoren für ein größeres N abnimmt. Der maximale F1-Wert liegt bei 0,1 für 100 Artikel bei 4-grammen unter der Verwendung des Klassifikators GNB oder SVC. Ähnliche Ergebnisse konnte die Anzahl der Ziffern mit nur einer Dimension erzielen.

Als Datengrundlage der Wort-2 und 3-gramme werden die jeweiligen Ergebnisse für die einzelnen Klassifikatoren des iterativen Filters verwendet (vgl. Kapitel 7.2.1). Eine Ausnahme besteht hier bei den Daten für den Klassifikator GNB bei Wort 3-grammen. Durch die unbefriedigenden Ergebnisse des iterativen Filters werden hier die ungefilterten Daten nach dem Bestimmen der relativen Häufigkeit verwendet (vgl. Kapitel 7.2.2).

Bild 20 zeigt die Ergebnisse getrennt nach den drei Klassifikatoren. Für SVC und KNN zeigen Wort 2-gramme für jede Anzahl von Texten bessere Ergebnisse als Wort 3-gramme. Weniger eindeutig zeigt sich der Verlauf bei GNB. Teilweise können hier bessere Ergebnisse durch Wort 2-gramme erzielt werden. Im Falle von KNN sind die Ergebnisse allerdings nur leicht besser als bei 4 bis 6-grammen. Insgesamt besteht die Vermutung, dass Wort-N-Gramme kein informatives Merkmal für die Klassifikation bei KNN und in Anbetracht der

Dimensionalität auch bei den anderen Klassifikatoren ungeeignet ist. Eine Aussage darüber kann aber erst mit einer Auswertung über alle Merkmale getroffen werden, da zu diesem Zeitpunkt die Korrelation der Wort-N-Gramme zu anderen Merkmalen unklar ist.

Aufgrund der konstant besseren Ergebnisse der Wort-2-Gramme bei den Klassifikatoren SVC und KNN und der geringen Differenz bei GNB wird im weiteren Verlauf der Arbeit N=2 für Wort-N-Gramme gewählt.

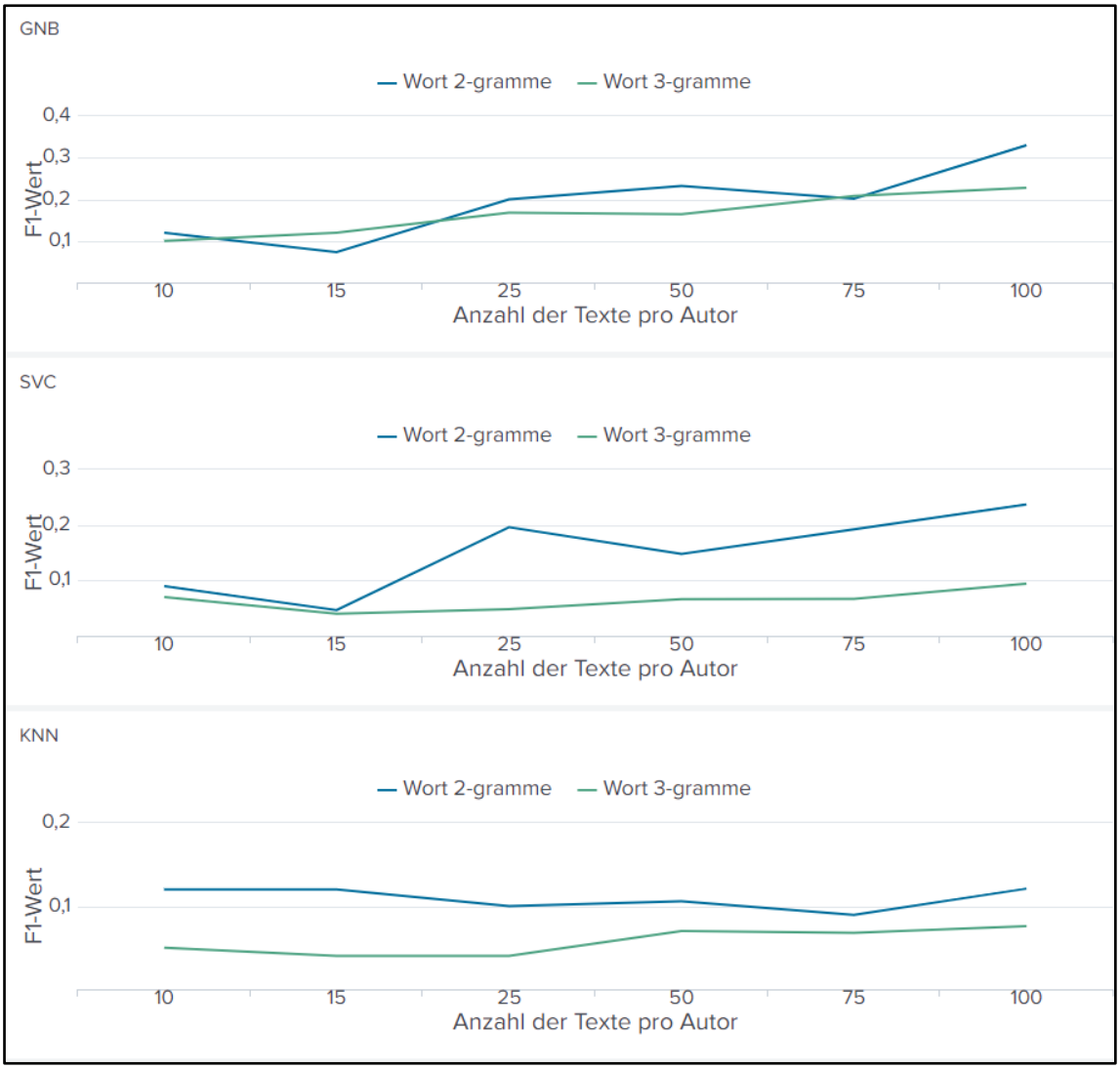


Bild 20: Bewertung von Wort-2- und Wort-3-Grammen - Eigene Darstellung

7.3.4. Bestimmen der Länge von Zeichen-N-Grammen

In Kapitel 0 wurden Zeichen-N-Gramme als eines der vielversprechendsten Merkmale aufgezeigt. Dabei wurde die Größe von N als eine der wichtigsten Eigenschaften dieser bezeichnet. Ein größeres N erfasst eher lexikalische und kontextbasierte Informationen, erhöht dabei aber auch die Dimensionalität des Merkmals. Ein kleines N hingegen ist für die Darstellung der Kontextinformationen nicht geeignet, steigert aber durch die verringerte Anzahl der Dimensionen die Geschwindigkeit des Modells. Ob ein Erfassen von Kontextinformationen für das Klassifikationsergebnis positiv oder negativ zu bewerten ist, gilt im Folgenden zu evaluieren. In der englischen Sprache wurden die besten Ergebnisse mit einer Länge von N gleich drei erzielt (Sapkota, Bethard, Montes, & Solorio, 2015, S. 96). Die Definition von N ist allerdings sprachenabhängig und ein größerer Wert für N erscheint in der deutschen Sprache, welche durchschnittlich längere Wörter als die englische Sprache besitzt, durchaus sinnvoll (vgl. Kapitel 5.6.4). Aus diesem Grund werden Längen der „Zeichen-N-Gramm“-Superklassen von zwei bis fünf für die verschiedenen Klassifikatoren geprüft.

Die Wahl eines optimalen N kann für „Zeichen-Affix-N-Gramme“ nicht pauschal für alle Klassifikatoren erfolgen. Abhängig vom Algorithmus zeigen sich teilweise stark differenzierte Ergebnisse. Basierend auf dem GNB zeigen sich für 2-Gramme die besten F1-Werte. Die Kurve verläuft gegen Ende abflachend, ist aber dennoch bis zu 75 Artikel pro Autor führend. Bei 100 Dokumenten zeigen sich bessere Ergebnisse für 4-Gramme. Diese zeigen aber ansonsten für weniger als 75 Artikel deutlich geringere Werte als die Kurve der 2-Gramme. 4- und 5-Gramme zeigen konstant geringere F1-Werte.

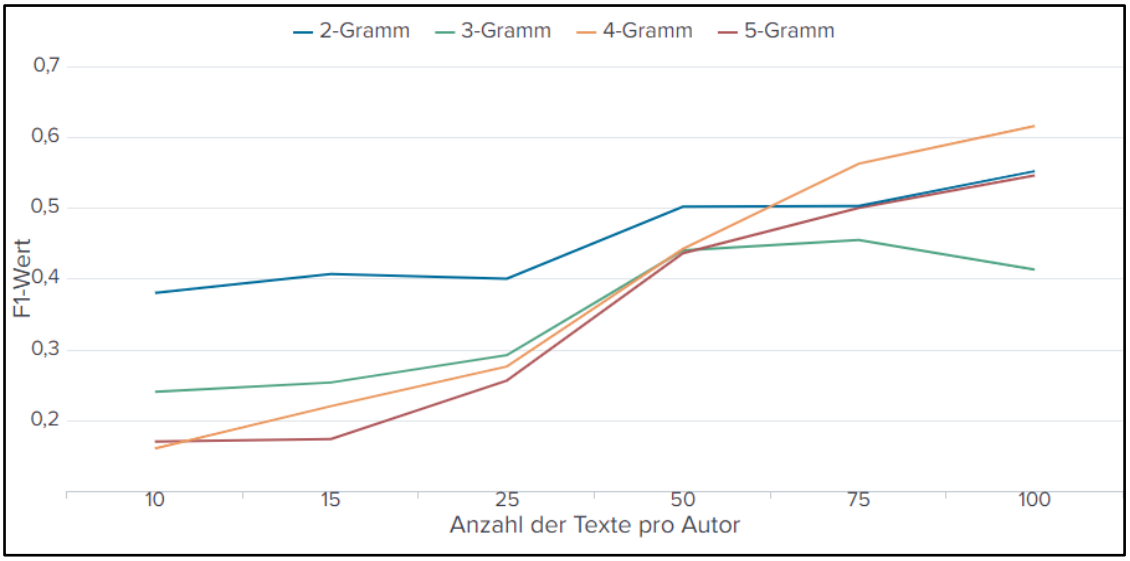


Bild 21: Bewertung Zeichen-Affix-N-Gramme GNB - Eigene Darstellung

Die zuvor schlecht bewerteten 4-Gramme zeigen sehr gute Ergebnisse für den SVC. Lediglich bei 10 oder 15 Artikeln können die anderen Kurven bessere Ergebnisse aufweisen. Allein 3-Gramme besitzen für 15 Artikel deutlich bessere Werte. Die konstant gute Bewertung der 2-Gramme unter Anwendung des GNB zeigen hier auch konstante, aber vergleichsweise die geringsten Werte.

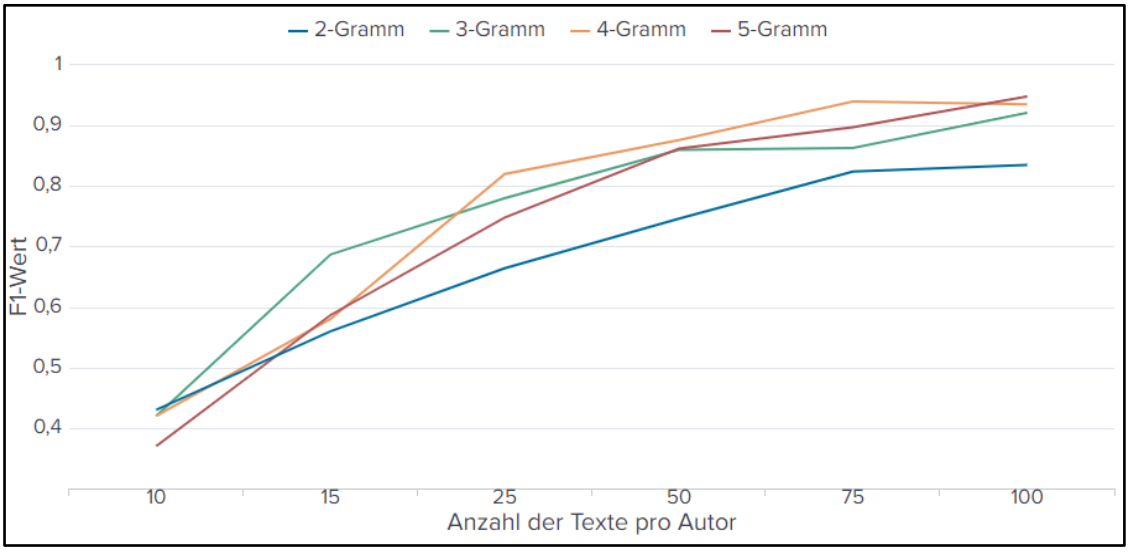


Bild 22: Bewertung Zeichen-Affix-N-Gramme SVC - Eigene Darstellung

Die Kurven der 4- und 5-Gramme unter Anwendung von KNN verlaufen schwankend zwischen den besten und schlechtesten F1-Werten. Da eine gewisse Konstanz der Ergebnisse für ein Verfahren als unabdingbar angesehen

wird, werden die Kurven der 2 und 3-Gramme insgesamt als günstiger erachtet. Hierbei werden im Vergleich von 2-Grammen durchgängig höhere Werte erzielt. Für GNB und KNN wird deshalb eine Länge von N=2 und für SVC eine Länge von N=4 im weiteren Verlauf ausgewählt.

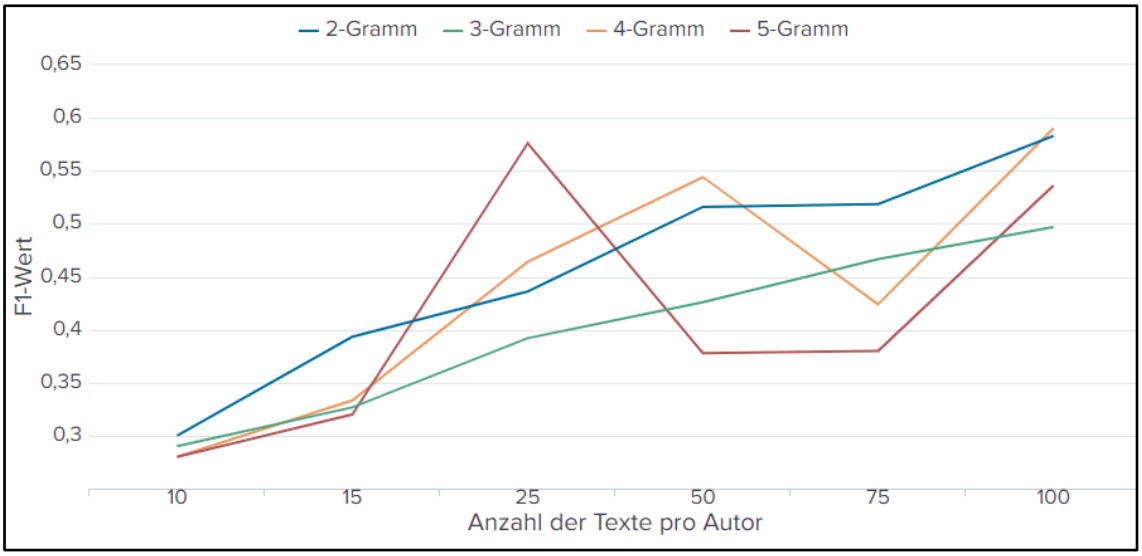


Bild 23: Bewertung Zeichen-Affix-N-Gramme KNN - Eigene Darstellung

Über alle Klassifikatoren der „wortbasierten Zeichen-N-Gramme“ besitzen die 2- und 5-gramme die geringsten F1-Werte. Aufgrund der Übersichtlichkeit werden aus diesem Grund nur die 3 und 4-Gramme in den folgenden Bildern aufgezeigt.

Insgesamt gleichen sich die Verläufe der Kurven für die Klassifikatoren. In der Bewertung durch GNB ergeben sich grundsätzlich bessere Werte für 3-Gramme. Die einzige Ausnahme besteht in einem 0,025 höheren F1-Wert bei 100 Artikeln. In der Klassifikation mittels SVC zeigen sich bessere Ergebnisse für 3-Gramme bei einer geringeren (<50) und für 4-Gramme bei einer höheren Anzahl (>=50) von Artikeln. Die Differenz der Kurven zeigt sich bei weniger Dokumenten allerdings höher, wodurch 3-Gramme insgesamt bessere Ergebnisse aufzeigen. Gegensätzlich verlaufen die Kurven bei der Bewertung durch KNN. Hier zeigen 4-Gramme bessere Ergebnisse bei einer kleineren (<75) und schlechtere bei einer höheren Anzahl von Dokumenten. Auch hier ergeben sich höhere Differenzen bei der kleineren Anzahl von Dokumenten, wodurch hier 4-Gramme besser zu bewerten sind. Folglich werden für GNB und SVC 3-Gramme und für KNN 4-Gramme verwendet.

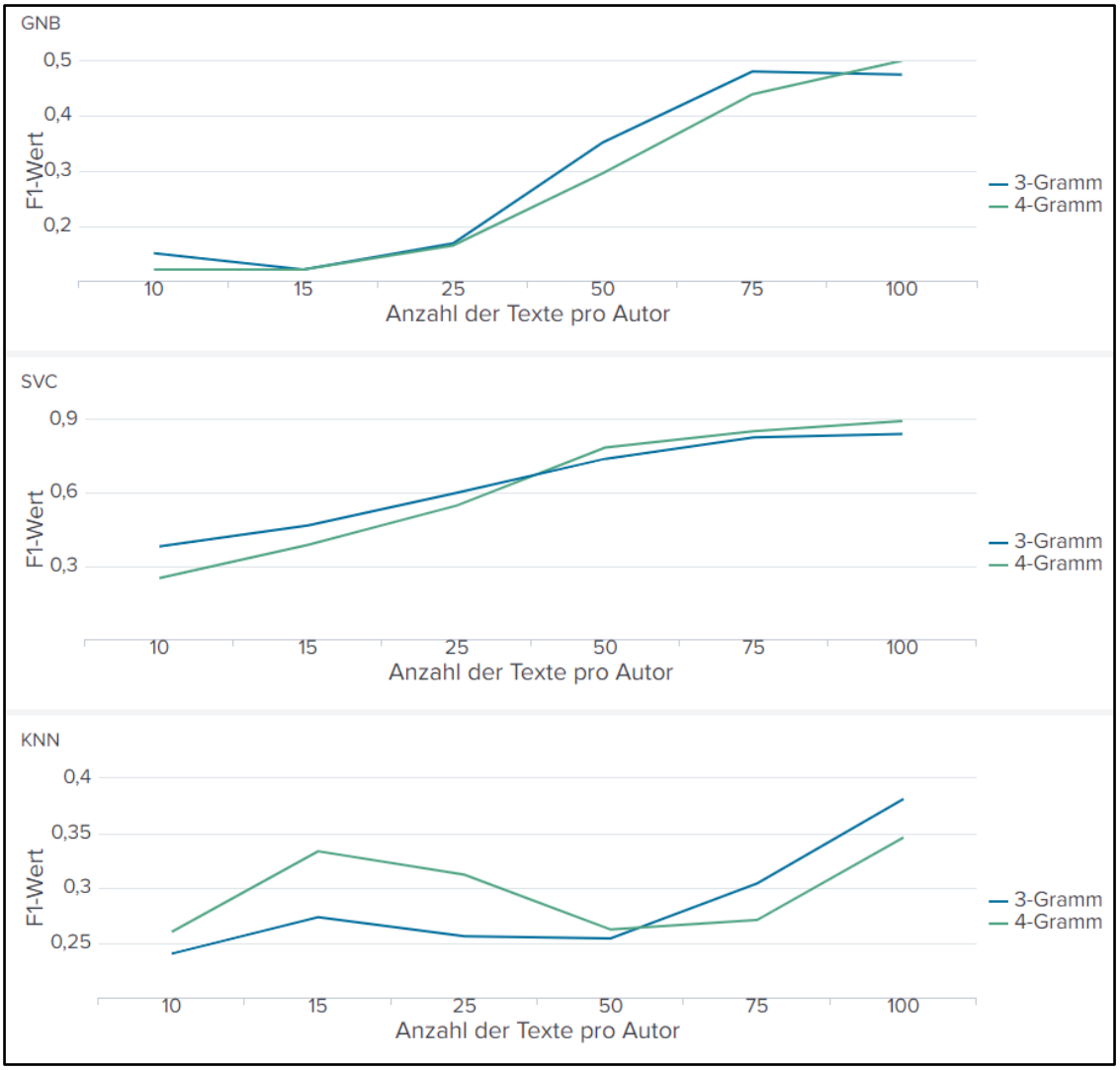


Bild 24: Bewertung Zeichen-Wort-N-Gramme - Eigene Darstellung

Auch bei den „Zeichen-Interpunktion-N-Grammen“ können zwei N-Gramme im Vorfeld ausgeschlossen werden. Bei allen Klassifikatoren ergeben sich für 4- und 5-Gramme die geringsten F1-Werte.

Für die Klassifikatoren GNB und KNN zeigen sich im direkten Vergleich der beiden Kurven eindeutige Ergebnisse. Die Graphen der 2-Gramme besitzen durchgehend höhere Werte als die der 3-Gramme. Bei der Klassifikation nach SVC zeigen 3-Gramme grundsätzlich höhere Werte. Die einzige Ausnahme besteht bei 25 Artikeln mit einem marginalen Unterschied von 0,16. Daraus folgt, dass im weiteren Verlauf für SVC 3-Gramme und für GNB als auch KNB 2-Gramme verwendet werden.

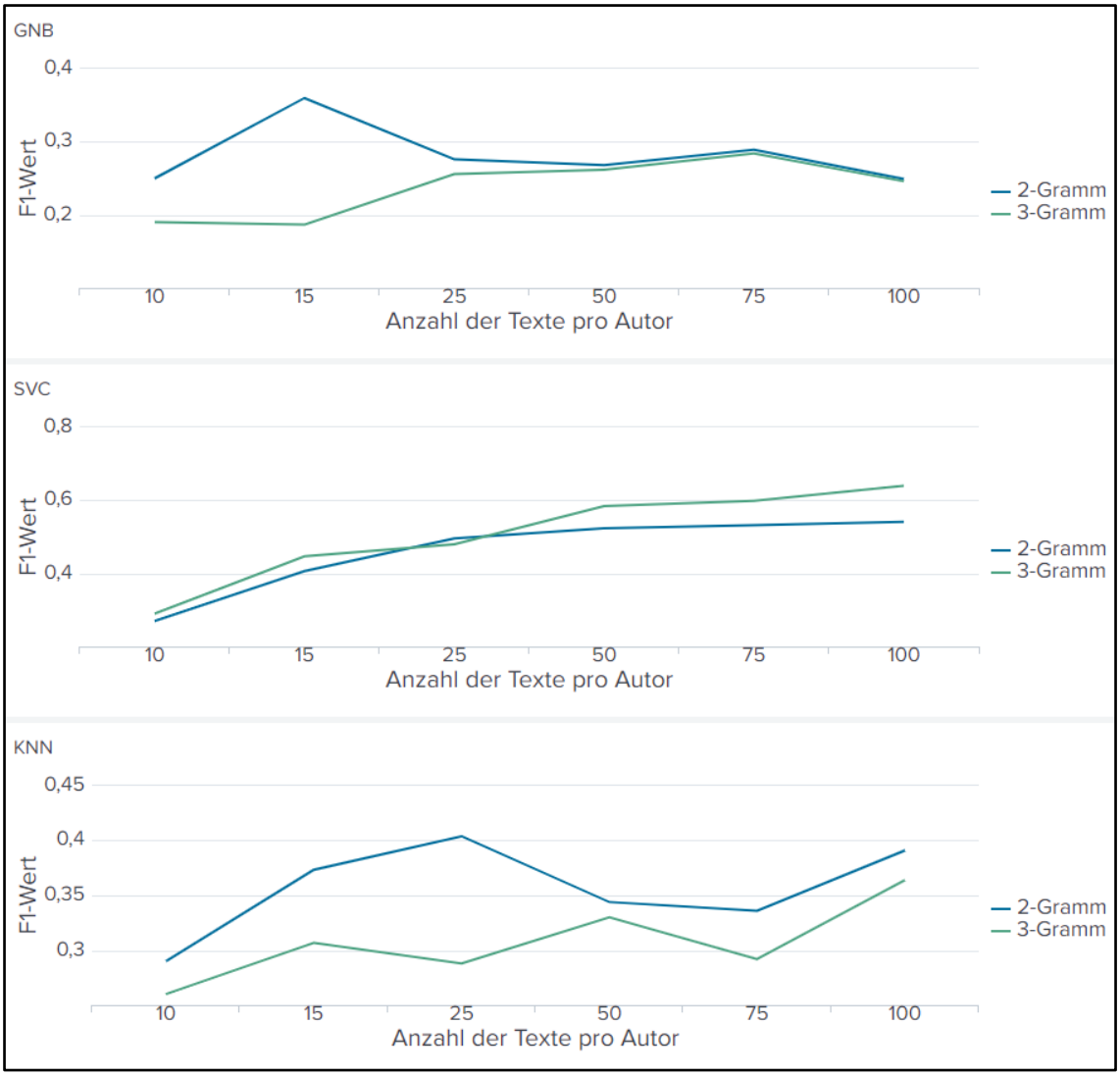


Bild 25: Bewertung Zeichen-Interpunktion-N-Gramme - Eigene Darstellung

Zusammenfassend zeigt sich eine starke Abhängigkeit der Größe von N zum Klassifikator und der Klasse der Zeichen-N-Gramme. Grundsätzlich bessere Ergebnisse unter der Verwendung von langen N-Grammen konnten nicht bestätigt werden. Unter der Verwendung der 5-Gramme sind meist geringere F1-Werte festzustellen. Zeichen-4-Gramme konnten teilweise höhere F1-Werte erzielen. Darunter die Klassifikation mittels Zeichen-Affix-N-Grammen durch den SVC und Zeichen-Wort-N-Gramme durch KNN. GNB und KNN erzielten die höchsten Werte bei Zeichen-Interpunktion-N-Grammen mit einem kleinen N von zwei. Für SVC zeigten sich hier 3-Gramme am effektivsten.

7.3.5. Wortarten-N-Gramme

In Kapitel 5.6.11 wurden „Wortarten-N-Gramme“ als Möglichkeit beschrieben, die syntaktische Struktur eines Textes zu erfassen. Dabei sollen die „Wortart-N-Gramme“ die verwendeten Phrasen eines Autors darstellen. Es gilt zu prüfen, ob durch ein größeres N längere Phrasen erfasst werden können und der dadurch gesteigerte Informationsgehalt die hohe Anzahl an Dimensionen längerer „Wortarten-N-Gramme“ rechtfertigt.

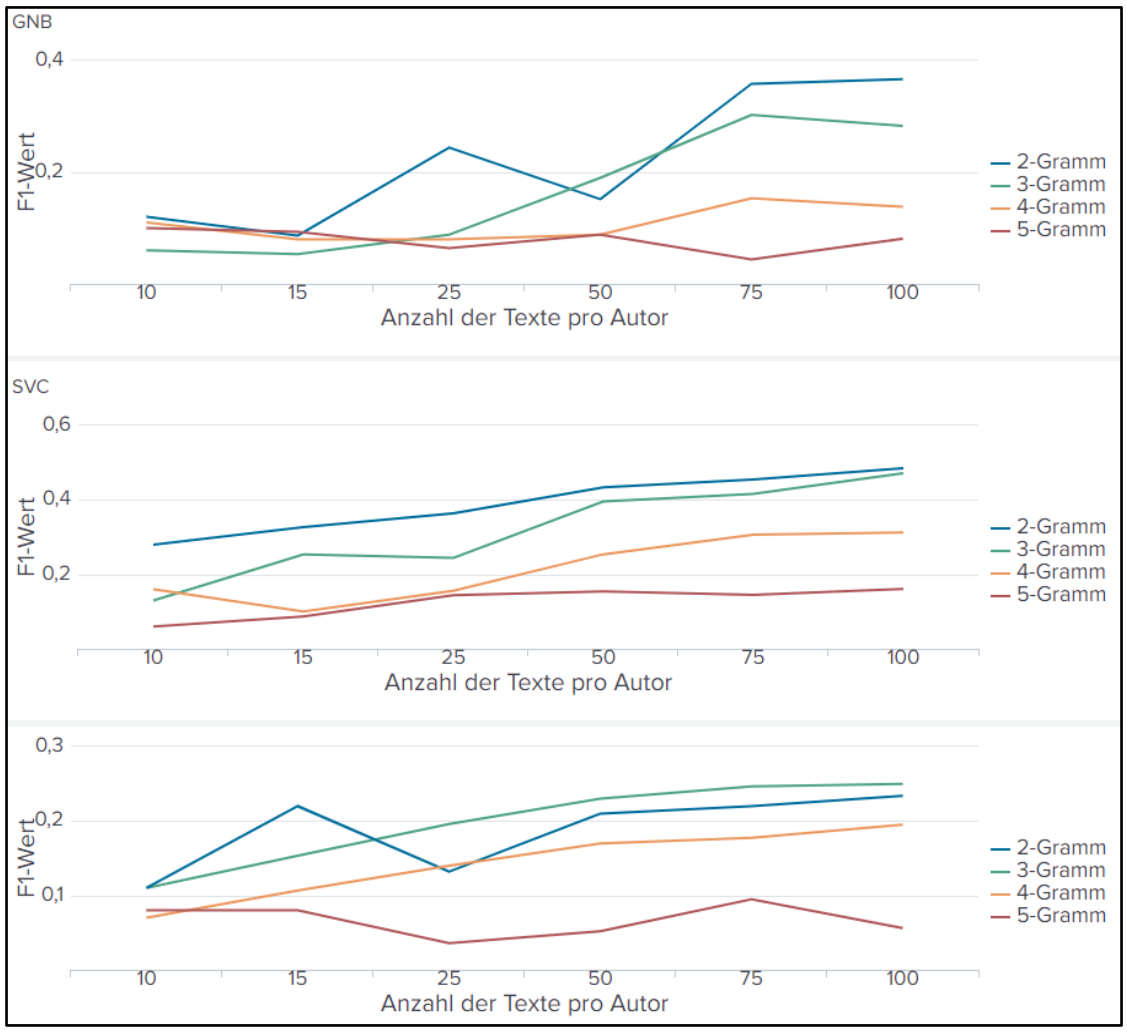


Bild 26: Bewertung Wortarten-N-Gramme - Eigene Darstellung

Grundsätzlich beschreiben die Ergebnisse aller Klassifikatoren einen Trend, bei welchem die F1-Werte mit zunehmendem N abnehmen. Eine Ausnahme besteht bei der Klassifikation mittels KNN. Hier ergeben sich bessere Werte für 3-Gramme ab 25 Texten pro Autor. Aus diesem Grund werden für GNB und SVC

2-Gramme und KNN 3-Gramme als bestmögliche Wortarten-N-Gramme gewählt. Die Bewertung anhand des F1-Wertes wird in Bild 26 dargestellt.

Anlehnend an bisherige Arbeiten wird von größeren Werten als drei für N abgesehen (Halvani, Steinebach, & Neitzel, 2014; Abbasi & Chen, 2008; Stamatatos, 2009). Die deutlich erhöhte Anzahl an Dimensionen wird keineswegs durch die Klassifikationsergebnisse gerechtfertigt. Die Information kann mit kleineren N-Grammen besser extrahiert werden.

8. Modellierung

Das Ziel dieser Arbeit ist die Aufstellung eines Modells der Autorschaftsattribution in der deutschen Sprache, welches die Identifikation des Autors eines anonymen Textes aufgrund eine Menge von Texten mit bekanntem Autor ermöglicht. (vgl. Kapitel 2.2). Weiterhin soll eine Bewertung der hierfür verwendeten Merkmale erfolgen und gegebenenfalls existierende Eigenheiten dieser aufgezeigt werden. Bei dem Modell handelt es sich dabei um ein Verfahren innerhalb eines „Closed-Set Problems“, dem instanzbasierten Paradigma folgend.

Zu diesem Zweck wurden in den vorherigen Kapiteln die Daten vorverarbeitet, Merkmale ausgewählt und Varianten der Merkmale bewertet. Auf dieser Basis kann ein Modell mit der Gesamtheit der Merkmale erstellt werden. Anschließend erfolgt eine Bewertung der verwendeten Merkmale.

8.1. Wahl des Klassifikators

Im Rahmen dieser Arbeit wurden im bisherigen Verlauf drei verschiedene Klassifikatoren betrachtet. GNB als probabilistisches Modell, KNN als parametrisches Modell und SVC als Breiter-Rand-Klassifikator (vgl. Kapitel 6.5). In den meisten bestehenden Arbeiten wird SVC aufgrund seiner guten Ergebnisse verwendet (Potha & Stamatatos, 2014, S. 315; Stolerman, Overdorf, Afroz, & Greenstadt, 2014, S. 189), welche andere Algorithmen deutlich übertreffen (Abbasi & Chen, 2008, S. 4; Sapkota, Bethard, Montes, & Solorio, 2015, S. 96). Diese Ergebnisse konnten in den bisherigen Kapiteln dieser Arbeit bestätigt werden.

In der Datenvorbereitung zeigte SVC grundsätzlich bessere Genauigkeits- und F1-Werte für die jeweiligen Merkmale. GNB zeigte zusätzlich große Probleme bei einem hohen Vorkommen an Nullwerten. Die Ergebnisse konnten durch die Merkmalsauswahl des iterativen Filters deutlich verbessert werden (vgl. Kapitel 7.2.2). Dennoch reichten diese nicht an die des SVC ran. Die Klassifikation mit

KNN zeigte oft schwankende Kurven, welche sich ebenfalls meist unter den Ergebnissen des SVC bewegten.

Diese Ergebnisse wurden durch weiterführende Tests mit Kombinationen von Merkmalen bestätigt. Die Klassifikation mittels SVC zeigt hier deutlich bessere Ergebnisse. Da diese lediglich die bestehende Literatur bestätigen und keinen neuen Aspekt aufzeigen, werden sie an dieser Stelle nicht dargestellt. Das Modell verwendet aufgrund der vielversprechenden Ergebnisse SVC als Klassifikator.

8.2. Zusammenfassung der verwendeten Merkmale

Das Modell basiert auf den in Kapitel 5.6 vorgestellten Merkmalen. Einige dieser Merkmale wurden in der Datenvorbereitung bezüglich ihrer Dimensionalität und ihres Wertebereiches angepasst. Die Schritte beinhalteten das Eliminieren von Merkmalsausprägungen mit geringer (vgl. Kapitel 7.1.1) und hoher (vgl. Kapitel 7.1.2) Häufigkeit, das Ermitteln der individuellen relativen Häufigkeit (vgl. Kapitel 7.1.4) und einer Merkmalsauswahl anhand des vorgestellten iterativen Filters (vgl. Kapitel 7.2.1). Weiterhin wurden verschiedene Varianten einiger Merkmale aufgezeigt und bewertet.

Da die jeweiligen Schritte nur auf einen spezifischen Teil der Merkmale angewandt wurden, werden diese in der folgenden Tabelle aufgezeigt. Dabei werden zur Übersichtlichkeit die Schritte durch die jeweilige Kapitelnummer bezeichnet, in welcher sie behandelt wurden. Des Weiteren werden die Dimensionalität bei der Gesamtzahl an 2.500 Artikeln der Textsammlung und der Wertebereich des Merkmals dargestellt. Die Merkmale „Anzahl der Wörter“, „Wortschatzreichtum“ und „Verständlichkeitsindizes“ können theoretisch unendlich viele Werte annehmen, weshalb hier ein Bereich zwischen dem Minimum und Maximum der Datengrundlage angegeben wird. Zuletzt wird die Tabelle um die anzunehmende Themenabhängigkeit, beschrieben durch „T“, erweitert. Es werden nur die bestbewerteten Varianten der Merkmale für den SVC im Modell verwendet.

Tabelle 20: Zusammenfassung der verwendeten Merkmale - Eigene Darstellung

Merkmal	Dim.	Wertebereich	Datenvorbereitung				T
			7.1.1	7.1.2	7.1.4	7.2.1	
Bag-of-Words	3.048	0 bis 1	X	X	X	X	X
Wort 2-Gramm	7.793	0 bis 1	X	X	X	X	X
Anzahl der Wörter	1	95 bis 1733					X
Wortlänge 1-20 und >20	21	0 bis 1			X		
Wortschatzreichtum	1	67 bis 782					
Sonderzeichen	28	0 bis 1			X		
Zeichen-Affix-4-Gramme	1.793	0 bis 1	X		X	X	
Zeichen-Wort-3-Gramme	869	0 bis 1	X		X	X	X
Zeichen-Interpunktion-3-Gramme	519	0 bis 1	X		X	X	
Ziffern gesamte Häufigkeit	1	0 bis 1			X		
Funktionswörter	79	0 bis 1	X	X	X	X	
Wortarten	14	0 bis 1			X		
Wortarten 2-gramme	184	0 bis 1	X		X		
Satzanfänge und –endungen	27	0 bis 1			X		
Verständlichkeitsindizes	1	-56 bis 71,9					

8.3. Optimieren der Parameter

SVC versucht, die beste Hyperebene zur Trennung der verschiedenen Klassen zu finden, indem der Abstand zwischen den Testpunkten und der Hyperebene maximiert wird (vgl. Kapitel 6.5.3). Dabei besitzt er einige Parameter, welche das Verhalten des Klassifikators beeinflussen. Für ein optimales Ergebnis müssen diese optimiert werden. Dies erfolgt mittels Kreuzvalidierung (vgl. Kapitel 6.8.3). Im Folgenden werden die Hyperparameter für SVC vorgestellt und die geprüften Parameter aufgelistet.

Bei einem der Parameter handelt es sich um den verwendeten Kernel. Die Hauptfunktion des Kernels besteht darin, den niedrigdimensionalen Eingaberaum in einen höherdimensionalen Raum zu transformieren (vgl. Kapitel 6.5.3). Als mögliche Parameter werden die vorgestellten Kernel verwendet. Darunter der lineare-, polynomiale- und RBF-Kernel (vgl. Kapitel 6.5.3).

Die in (ebd.) vorgestellten Schlupfvariablen werden über den Parameter C abgebildet. C ist der Bestrafungs-Parameter, der den Fehlklassifikations- oder

Fehlerterm darstellt. Dieser besagt, wie viele Fehler verkraftbar sind. Bei einem hohen C besteht die Gefahr einer Überanpassung. (Liu C. , 2020) Es werden deshalb die Werte „0.001, 0.01, 0.1, 1, 10“ verwendet.

Gamma ist ein Parameter für nicht lineare Hyperebenen. Je höher der Gamma-Wert ist, desto mehr wird versucht, den Trainingsdatensatz genau anzupassen (Fraj, 2018). Wenn Gamma höher ist, haben nahe gelegene Punkte einen hohen Einfluss; ein niedriges Gamma bedeutet, dass auch weit entfernte Punkte bei der Ermittlung der Entscheidungsgrenze berücksichtigt werden (Liu C. , 2020). Die Bibliothek besitzt zwei variable Attribute „scale“ [1 / (Anzahl der Merkmale * Varianz der Merkmale)] und „auto“ (1 / Anzahl der Merkmale) (scikit-learn developers, 2020i). Diese werden für die Hyperparameteroptimierung in diesem Modell verwendet.

8.4. Wahl der Skalierungsmethode

Tabelle 20 zeigt, dass innerhalb des Modells Merkmale mit unterschiedlichen Wertebereichen vorliegen. Die meisten Merkmale bewegen sich durch das Bestimmen der individuellen relativen Häufigkeit in einem Bereich von null bis eins. Dennoch existieren drei Merkmale mit einem sehr großen Wertebereich: „Anzahl der Wörter“, „Wortschatzreichtum“ und „Verständlichkeitsindizes“. Eine mögliche Folge wäre, dass andere Merkmale durch diese großen Merkmale dominiert werden und dadurch schlechtere Ergebnisse des Klassifikators entstehen (vgl. Kapitel 6.10). Dies kann durch eine Normalisierung oder Standardisierung der Daten behoben werden (ebd.). Weiterhin wird in (ebd.) angeführt, dass SVM davon ausgehen, dass alle Daten um null zentriert sind. Dies würde für eine Standardisierung der Daten sprechen. Dennoch sind die Ergebnisse dieser Methoden stark vom Anwendungsfall abhängig und im Zweifelsfall ist ein Vergleich der Ergebnisse unter der jeweiligen Anwendung zielführend (ebd.). Aus diesem Grund werden in Bild 27 die Ergebnisse des Verfahrens mit einem jeweiligen Graph für eine fehlende Skalierung, mit einer Normalisierung oder Standardisierung aufgezeigt. Auf der Ordinate werden verschiedene Kombinationen aus der Anzahl der Autoren und Texte aufgezeigt.

Dabei ist zu beachten, dass die Anzahl der Texte zur Ermittlung der Ergebnisse einer 60:40 Trainings-/Testaufteilung unterzogen werden.

Anlehnend an die Grundannahme der SVM ergeben sich durch eine Standardisierung konstant signifikant höhere F1-Werte. Eine Normalisierung verringert die Werte in der Regel. Aufgrund dieser eindeutigen Ergebnisse erfolgt im weiteren Verlauf eine Standardisierung der Werte. Demnach entspricht die Kurve der standardisierten Werte den vorläufigen Klassifikationsergebnissen des Verfahrens.

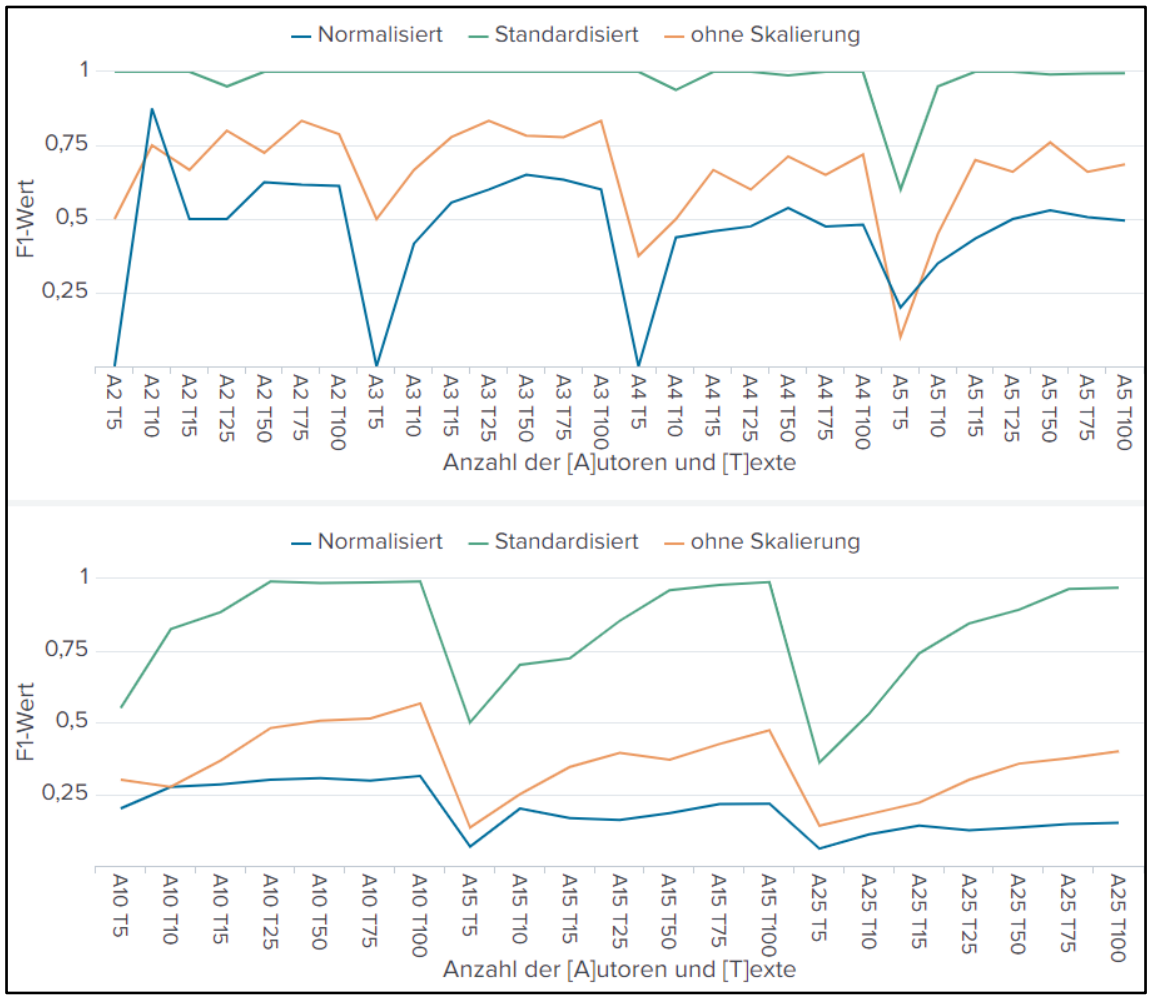


Bild 27: Bewertung des Modells unter verschiedenen Skalierungen - Eigene Darstellung

8.5. Bewertung der Merkmale

Nachdem in den vorherigen Kapiteln ein Modell entwickelt wurde, werden anhand dieses Modells nun die Merkmale bewertet. Um eine möglichst gesamtheitliche Bewertung der Merkmale durchzuführen, werden diese unter verschiedenen Szenarien betrachtet.

Im vorherigen Kapitel wurde durch die Kurve der standardisierten Werte die vorläufigen Klassifikationsergebnisse des Verfahrens unter Verwendung aller Merkmale aufgezeigt. Anhand dieser Kurve können zwei weitere Beobachtungen erfolgen. Zum einen die geringen Unterschiede der F1 Werte bei geringer Autoranzahl unter steigender Anzahl der Texte. Durch das Erreichen des Maximalwertes bei einer geringen Anzahl von Trainingsdaten können die Ergebnisse durch eine größere Anzahl von Artikeln nicht verbessert werden. Zum anderen ist ein ähnlicher Verlauf der Kurve bei steigender Textanzahl und gleichbleibender Autoranzahl zu beobachten. Eine größere Anzahl von Autoren führt grundsätzlich zu geringeren F1-Werten und folgt in einer Steigung durch eine erhöhte Anzahl an Artikeln.

Dies bestätigt die in Kapitel 7.3 getroffene Annahme, dass die Bewertung im Allgemeinen bei geringerer Autorenzahl für alle Verfahren besser ausfällt, dabei aber die Gefahr einer nicht repräsentativen Menge an Variationen für Autoren existiert. Aus diesem Grund wird auch an dieser Stelle ein Vergleich der Merkmale nur über die Gesamtzahl der Autoren bei variierender Textanzahl vorgenommen. Die Textanzahl entspricht dabei der gesamten Menge des Verfahrens, welche einer Unterteilung von 60:40 in Test- und Trainingsdaten unterzogen werden.

8.5.1. Bewertung der individuellen Merkmale

Im ersten Schritt erfolgt eine individuelle Bewertung der Merkmale. Das vorgestellte Modell wird zu diesem Zweck nur mit dem jeweiligen Merkmal durchgeführt und mittels F1-Wert bewertet. Tabelle 21 zeigt die Bewertungen für die variierende Textanzahl auf. Jede Spalte ist dabei einer Farbskalierung unterzogen, welche den höchsten Wert am dunkelsten und den niedrigsten Wert

weiß kennzeichnet. Die Merkmale werden anhand ihrer Merkmalsgruppen unterteilt.

Tabelle 21: Bewertung der individuellen Merkmale - Eigene Darstellung

		Anzahl der Texte						
		5	10	15	25	50	75	100
Lexikalisch	Bow	0,1000	0,2600	0,3333	0,5160	0,6000	0,6680	0,7070
	Wort-2-Gramme	0,0800	0,0700	0,0667	0,1400	0,1960	0,0947	0,1930
	Anzahl der Wörter	0,1200	0,0800	0,0600	0,1040	0,1140	0,1107	0,1240
	Wortlänge	0,1600	0,1300	0,1333	0,1560	0,1420	0,1907	0,1830
	Wortschatzreichtum	0,0600	0,1000	0,0667	0,0840	0,0760	0,0920	0,0930
Symbolisch	Z.-Affix-4-Gramme	0,1800	0,3800	0,6133	0,7720	0,8920	0,9333	0,9390
	Z.-Wort-3-Gramme	0,1600	0,3600	0,3667	0,6240	0,7620	0,8320	0,8520
	Z.-Interpunktion-3-Gramme	0,2000	0,2900	0,4533	0,4880	0,5880	0,6000	0,6220
	Sonderzeichen	0,2400	0,3200	0,4000	0,3880	0,3880	0,4133	0,4170
	Ziffern	0,1000	0,0700	0,0533	0,0560	0,0700	0,0813	0,0830
Syntaktisch	Funktionswörter	0,1000	0,1400	0,1467	0,1840	0,2400	0,2853	0,3120
	Wortarten	0,2200	0,2400	0,2333	0,3320	0,2980	0,3360	0,3260
	Wortarten 2-gramme	0,2400	0,2800	0,3267	0,3640	0,4340	0,4547	0,4850
	Satzanfänge & –endungen	0,0400	0,0600	0,1067	0,1080	0,1260	0,1480	0,1410
	Verständlichkeitsindizes	0,1000	0,0700	0,0733	0,0960	0,1160	0,1107	0,1080

Die höchsten F1-Werte sind bei Zeichen-Affix-4-Grammen, Zeichen-Wort-3-Grammen, BoW und Zeichen-Interpunktion-3-Grammen absteigend in dieser Reihenfolge zu erkennen. Unter der ausschließlichen Verwendung von Zeichen-Affix-4-Grammen werden etwa 94% der Testdaten korrekt klassifiziert. Augenscheinlich besitzen symbolische Merkmale insgesamt die höchsten F1-Werte. Mittelhohe F1-Werte zeigen sich bei Sonderzeichen, Wortarten und Wortarten-2-Grammen. Hierbei ist auffällig, dass diese bereits bei wenigen Texten vergleichsweise hohe F1-Werte erzielen. Dies ist ebenfalls bei Wortlängen, Anzahl der Wörter und Verständlichkeitsindizes erkennbar.

Im Vergleich zur Klassifikation mit der Gesamtheit der Merkmale, zeigen einige Merkmale sehr hohe F1-Werte auf. Daraus entsteht die Vermutung, dass eine starke Korrelation zwischen den Merkmalen existiert.

8.5.2. Differenzfunktion

Im vorherigen Kapitel äußerte sich die Vermutung einer starken Korrelation der verwendeten Merkmale. Diese Vermutung soll im Folgenden geprüft werden. Das einfachste Verfahren zum Prüfen der Korrelation ist das Bestimmen des Korrelationskoeffizienten. Mit diesem Ansatz müsste ein Koeffizient zwischen jeder Merkmalsausprägung ermittelt werden. Bei 100 Texten würde dies 206.755.641 Korrelationskoeffizienten entsprechen. Weiterhin würde dies nur eine Aussage über die Korrelation einzelner Ausprägungen der Merkmale geben. In Kapitel 7.1.4 hat sich bereits gezeigt, dass eine hohe Anzahl von Nullwerten den durchschnittlichen Korrelationskoeffizienten verringert und dadurch eine Betrachtung für das gesamte Merkmal erschwert. Aufgrund der Rechenintensivität und der schwierigen Interpretationsmöglichkeit wird ein anderer Ansatz verfolgt.

Dieser Ansatz besteht aus im Folgenden genannten Differenzfunktion. Für das Ermitteln der Differenzfunktion werden die F1-Werte des Modells ohne die Verwendung eines Merkmals ermittelt. Die Ergebnisse werden von den F1-Werten des Modells unter Verwendung aller Merkmale subtrahiert. Demnach entspricht es dem direkten Mehrwert eines Merkmals für das Klassifikationsergebnis in diesem Modell. Die Werte der Differenzfunktionen werden in Tabelle 22 dargestellt. Hohe Werte werden grün und tiefe Werte blau in einer Farbskala markiert.

Eine Interpretation dieser Werte muss immer im Verhältnis zu den ermittelten F1-Werten in 8.5.1 erfolgen. Ein Großteil der Werte geht gegen null. Gerade Merkmale mit bereits geringen F1-Werten in Tabelle 21 scheinen keinen Mehrwert für die Klassifizierung zu liefern. Aber auch Merkmale mit zuvor sehr hohen F1-Werten besitzen in der Differenz geringe Werte. Dies spricht für eine starke Korrelation der Daten. Negative Werte bedeuten, dass die Klassifikation ohne das Merkmal bessere Ergebnisse erzielt. Bei „Zeichen-Affix-4-Grammen“ zeigt sich ein hoher negativer Wert. Dies spricht für eine hohe Korrelation zu anderen Merkmalen, welche ein schlechteres Klassifikationsergebnis erzielen als solche, zu denen keine Korrelation besteht. „Wort-2-Gramme“ scheinen das

Klassifikationsergebnis eher negativ zu beeinflussen, dies ist wenig überraschend in Anbetracht der in Kapitel 7.3.3 aufgezeigten geringen Bewertung.

Tabelle 22: Werte der Differenzfunktionen - Eigene Darstellung

		Anzahl der Texte						
		5	10	15	25	50	75	100
Lexikalisch	Bow	0,0000	-0,0100	-0,0133	0,0440	0,0060	0,0027	0,0190
	Wort-2-Gramme	0,0000	0,0100	-0,0200	-0,0280	-0,0260	0,0013	-0,0080
	Anzahl der Wörter	0,0200	0,0200	0,0067	0,0000	0,0000	0,0000	0,0000
	Wortlänge	0,0200	0,0100	0,0067	0,0080	0,0000	0,0000	0,0020
	Wortschatzreichtum	0,0200	0,0100	0,0000	0,0000	0,0000	0,0000	0,0000
Symbolisch	Z.-Affix-4-Gramme	-0,0600	0,0700	0,0733	0,0280	0,0440	0,0173	0,0320
	Z.-Wort-3-Gramme	-0,0200	0,0400	-0,0133	-0,0080	0,0060	0,0000	0,0090
	Z.-Interpunktion-3-Gramme	0,0600	0,0100	0,0333	0,0320	0,0160	0,0107	0,0080
	Sonderzeichen	0,0400	0,0100	0,0200	0,0080	0,0000	0,0000	0,0030
	Ziffern	0,0000	0,0100	0,0067	0,0000	0,0000	0,0000	0,0000
Syntaktisch	Funktionswörter	-0,0400	-0,0100	0,0000	0,0080	-0,0020	-0,0013	0,0010
	Wortarten	-0,0200	0,0100	0,0067	0,0040	0,0020	-0,0013	0,0020
	Wortarten 2-gramme	0,0600	0,0400	0,0267	0,0120	0,0060	-0,0013	0,0040
	Satzanfänge & –endungen	0,0000	0,0000	0,0067	0,0040	0,0020	0,0000	0,0010
	Verständlichkeitsindizes	0,0000	0,0100	-0,0067	0,0000	0,0000	0,0000	0,0000

Insgesamt zeigen sich sehr geringe Werte für themenabhängige Merkmale. Diese hatten im vorherigen Kapitel noch hohe bis sehr hohe F1-Werte. (Sapkota, Bethard, Montes, & Solorio (2015, S. 99f.) konnten bereits anhand von Zeichen-N-Grammen nachweisen, dass ohne die themenabhängigen Zeichen-Wort-N-Gramme bessere Ergebnisse erzielt werden konnten. Sie führten die Vermutung an, dass die Informationen in Zeichen-Wort-N-Grammen enthaltenen Informationen auf eine andere Art in den anderen Zeichen-N-Grammen enthalten sein müssen (ebd.). Aus diesem Grund erfolgt im folgenden Kapitel eine Prüfung, ob ein gleiches oder besseres Ergebnis ohne die Verwendung der themenabhängigen Merkmale erzielt werden kann und wie sich dies auf die Werte der Differenzfunktionen auswirkt.

8.5.3. Bewertung themenabhängiger Merkmale

Im vorherigen Kapitel wurde die These aufgestellt, dass ein Entfernen der themenabhängigen Merkmale zu gleichbleibenden oder besseren Klassifikationsergebnissen führen kann. Hierbei ist anzumerken, dass keine genaue Aussage darüber getroffen werden kann, ob themenabhängige Unterschiede zwischen den Texten besteht (vgl. Kapitel 5.1). Es besteht allerdings die Annahme, dass Musikreviews verschiedener Genres eine große Schnittmenge an themenspezifischem Vokabular besitzen.

Für den Vergleich erfolgt die Bewertung zweier Modelle mittels F1-Wert. Ein Modell verwendet dabei alle Merkmale und das andere keine themenspezifischen Merkmale. Das zweite Modell verwendet dadurch ca. 18,5% der Merkmale des ersten Modells. Im Vergleich der Modelle wird im Gegensatz zur Bewertung von einzelnen Merkmalen eine Variation der Autorenanzahl vorgenommen. Die Klassifikationsergebnisse befinden sich in Bild 28.

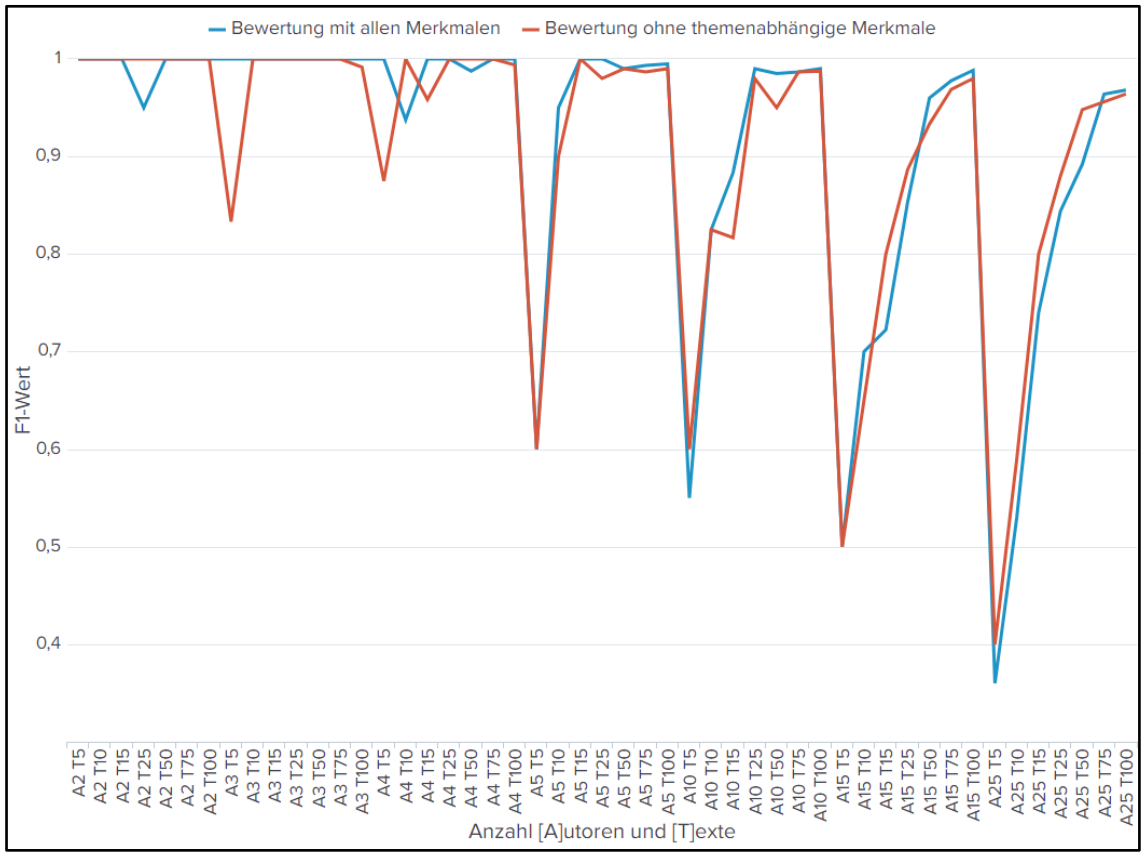


Bild 28: Vergleich der Modelle mit und ohne themenabhängiger Merkmale - Eigene Darstellung

Das Bild verdeutlicht einen sehr geringen Unterschied beider Kurven. Außer einigen Ausreißern erzielen beide Modelle grundsätzlich sehr ähnliche F1-Werte. In manchen Fällen kann das Modell ohne themenabhängige Merkmale sogar bessere Ergebnisse erzielen. Dies wird auf Vermutung zurückgeführt, dass die Informationen der themenabhängigen Merkmale bereits in den anderen Merkmalen enthalten sind (vgl. Kapitel 8.5.2). In Anbetracht dessen, dass das Modell diese Ergebnisse mit rund 18,5% der Merkmale bei 100 Texten erzielt, werden themenabhängige Merkmale im weiteren Verlauf nicht betrachtet. Weiterhin gilt es anzumerken, dass die Vermutung besteht, dass die Ergebnisse des ersten Modells weiterhin sinken, wenn Texte mit stark differierenden Themen als Grundlage der Klassifikation dienen.

Ein neues Aufstellen der Differenzfunktionen zeigt deutlich höhere Werte für „Zeichen-Affix-4-Gramme“. Auch bei den restlichen Merkmalen sind grundsätzlich höhere F1-Werte zu betrachten. In einer weiteren Analyse könnte versucht werden, weitere irrelevante Merkmale zu entfernen und Korrelationen aufzulösen. Da aber alle Merkmale insgesamt einen positiven Einfluss auf das Ergebnis besitzen, wird davon abgesehen. Die größte vermutete Korrelation existiert innerhalb der syntaktischen Merkmale. Die negativen Werte dieser sind im Vergleich zu den positiven Auswirkungen bei einer geringen Textanzahl zu vernachlässigen.

Tabelle 23: Werte der Differenzfunktionen themenunabhängiger Merkmale - Eigene Darstellung

		Anzahl der Texte						
		5	10	15	25	50	75	100
Lex.	Wortlänge	0,0800	0,0400	0,0067	0,0000	0,0040	0,0000	0,0010
	Wortschatzreichtum	0,1000	0,0200	0,0067	0,0040	0,0020	-0,0013	-0,0010
Symbolisch	Z.-Affix-4-Gramme	0,0800	0,1400	0,1600	0,1320	0,1620	0,1440	0,1290
	Z.-Interpunktion-3-Gramme	0,0400	0,0400	0,0733	0,0400	0,0200	0,0107	0,0070
	Sonderzeichen	0,0600	0,0500	0,0200	0,0040	0,0000	0,0067	0,0000
	Ziffern	0,1000	0,0100	0,0000	0,0000	0,0000	0,0000	-0,0010
Syntaktisch	Funktionswörter	0,1200	0,0300	-0,0067	0,0120	-0,0020	0,0053	0,0010
	Wortarten	0,0000	0,0500	0,0000	0,0000	-0,0040	-0,0027	0,0000
	Wortarten 2-gramme	0,0600	0,1000	-0,0067	0,0240	0,0060	0,0053	0,0030
	Satzanfänge & –endungen	0,0800	0,0000	0,0067	-0,0040	0,0000	0,0013	-0,0040
	Verständlichkeitsindizes	0,0000	0,0000	0,0067	0,0000	-0,0020	0,0013	0,0010

8.5.4. Zusammenfassende Bewertung der Merkmale

Zusammenfassend stellen sich „Zeichen-Affix-4-Gramme“ als effektivstes Merkmal dar. Mit einem F1-Wert von bis zu 0,939 bei 100 Texten erzielt es nahezu dieselben Werte wie die Kombination aller Merkmale. Es ist allerdings unklar, ob eine Themenabhängigkeit des Merkmals durch ein höheres N existiert. Dies kann in dieser Arbeit aufgrund der Datengrundlage nicht evaluiert werden (vgl. Kapitel 5.1).

Gefolgt wird das Merkmal von den „Zeichen-Interpunktion-3-Grammen“ und daraufhin „Sonderzeichen“. Beide erzielen gute Ergebnisse gerade bei einer geringeren Anzahl von Texten. Die niedrigen Werte der Differenzfunktion begründen sich wahrscheinlich durch eine Korrelation dieser beiden Merkmale.

Ebenfalls gute Ergebnisse zeigen die syntaktischen Merkmale „Wortarten-2-Gramme“, „Wortarten“ und „Funktionswörter“ absteigend in dieser Reihenfolge. Auch hier existiert vermutlich eine starke Korrelation innerhalb der Merkmalsgruppe.

Die restlichen Merkmale besitzen geringere Werte, können aber dennoch positiv zum Klassifikationsergebnis beitragen. Eine genaue Rangfolge erscheint hier allerdings nicht sinnvoll.

8.6. Klassifikationswerte des Modells

Bild 28 stellt bereits den Verlauf des schlussendlichen Modells ohne themenabhängige Merkmale dar. Zusammenfassend werden die F1-Werte des Modells gerundet auf vier Nachkommastellen in Tabelle 24 aufgezeigt.

Grundsätzlich sind bessere Klassifikationsergebnisse bei einer höheren Textanzahl und geringe F1-Werte bei einer größeren Autorenzahl zu beobachten. Geringere F1-Werte bei einer gleichbleibenden Autorenzahl und einer größeren Textanzahl kann durch die steigende Anzahl an Testdaten erklärt werden. Durch Ausreißer können sich hier leicht geringere F1-Werte zeigen.

Tabelle 24: F1-Werte des Modells - Eigene Darstellung

		Anzahl der Texte						
		5	10	15	25	50	75	100
Anzahl der Autoren	2	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	3	0,8333	1,0000	1,0000	1,0000	1,0000	1,0000	0,9917
	4	0,8750	1,0000	0,9583	1,0000	1,0000	1,0000	0,9938
	5	0,6000	0,9000	1,0000	0,9800	0,9900	0,9867	0,9900
	10	0,6000	0,8250	0,8167	0,9800	0,9500	0,9867	0,9875
	15	0,5000	0,6500	0,8000	0,8867	0,9333	0,9689	0,9800
	25	0,4000	0,5900	0,8000	0,8800	0,9480	0,9560	0,9640

Als Vergleichswert dient die viel zitierte Arbeit von Abbasi & Chen (2008). Bei der Unterscheidung von 25 Autoren erreichen sie Klassifikationsraten von 93,62% bei einem Email-Datensatz mit durchschnittlich 27.774 Wörter pro Autor und 100% für einen Datensatz aus Ebay-Kommentaren mit durchschnittlich 23.423 Wörter pro Autor (Abbasi & Chen, 2008, S. 26). Der Unterschied der beiden Datensätze bestand vorwiegend in der Menge an „Rauschen“ innerhalb der Daten. Als Rauschen werden beispielsweise nicht filterbarer Inhalt wie weitergeleitete oder rezitierte Abschnitte bezeichnet (Abbasi & Chen, 2008, S. 20).

Bei 60 Artikeln für das Training mit einer durchschnittlichen Artikellänge von 378,95 entspricht dies durchschnittlich 22.737 Wörtern pro Autor. Hier konnte eine Klassifikationsrate von 96,4% erreicht werden. Für ein Erreichen einer Klassifizierungsrate von etwa 94% würden nur 30 Texte als Trainingsdaten benötigt werden. Ein Rauschen durch das Einfügen von Credits und Songtexten innerhalb der Daten konnte ebenfalls festgestellt werden (vgl. Kapitel 5.6.2). Der Wert des erarbeiteten Modells fügt sich hier demnach etwa in der Mitte ein. Aus diesem Grund gilt die Entwicklung des Modells als Erfolg.

9. Evaluierung

9.1. Benötigte Rechenleistung

Zu Beginn der Arbeit wurden die notwendigen Berechnungen mit einem Lenovo ThinkPad T580 mit einem Intel Core i7-8550U Prozessor durchgeführt. Die Hardwarespezifikationen sind 32GB Arbeitsspeicher und 8 logische Prozessoren mit einer jeweiligen Geschwindigkeit von 2,0 GHz. Da die Berechnungen der scikit-learn Bibliothek nicht GPU-optimiert werden können, ist die GPU-Leistung irrelevant.

Die Hardware stieß dabei schnell an ihr Limit. Rechenintensive Operationen wie das Ermitteln aller Korrelationskoeffizienten in Kapitel 7.1.4 oder das Durchlaufen des iterativen Filters für hochdimensionale Merkmale in Kapitel 7.2.1 konnten nur mit einer Laufzeit von über einem Tag durchgeführt werden. Besonders rechenintensiv zeigte sich dabei die Hyperparameteroptimierung in den einzelnen Iterationen des iterativen Filters. Das Durchführen der Hyperparameteroptimierung und dem anschließenden Ermitteln einer Klassifikationsmetrik entspricht einem essentiellen Schritt der Arbeit, welcher in verschiedenen Szenarien durchgeführt wird. Um die Arbeit in der festgelegten Bearbeitungszeit von 26 Wochen durchführen zu können, musste die bestehende Hardware erweitert werden.

Aus diesem Grund wurden alle Berechnungen auf dem „IBM Power System AC922/LC922“ durchgeführt. Das System wurde als leistungsstarke Trainingsplattform entwickelt und bietet eine Infrastruktur, welche es ermöglicht, mit daten- und rechenintensiven Anwendungen umzugehen (IBM, 2020). Durch die Verwendung von nicht für das System optimierten Bibliotheken konnten allerdings zusätzliche Beschleunigungen durch das Design der Infrastruktur nicht verwendet werden. Dennoch besitzt das System sehr hohe Hardwarespezifikationen von 40 Prozessoren mit jeweils 2,7 GHz und einem Terrabyte Arbeitsspeicher (ebd.). Unter Verwendung dieser Hardware konnten die notwendigen Berechnungen in annehmbarer Zeit durchgeführt werden.

9.2. Bewertung der Ergebnisse

Das Ziel dieser Arbeit ist die Aufstellung eines Modells der Autorschaftsattribution in der deutschen Sprache, mit dem die Identifizierung des Autors eines anonymen Textes aufgrund einer gegebenen Menge von Texten mit bekanntem Autor möglich ist. Weiterhin soll eine Bewertung der hierfür verwendeten Merkmale erfolgen und gegebenenfalls existierende Eigenheiten dieser aufgezeigt werden.

Grundsätzlich kann die Modellierung des Verfahrens als Erfolg angesehen werden. Gemäß anderen Forschungen zeigte der SVC im Vergleich zu den anderen Klassifikatoren signifikant bessere Ergebnisse (vgl. Kapitel 8.1). Mit einer Klassifikationsrate von etwa 96,4% bei 25 Autoren und 60 Texten als Trainingsdaten bewegt sich das Verfahren auf einem ähnlichen Niveau wie bekannte Verfahren in der englischen Sprache (vgl. Kapitel 8.6). Sehr gute Ergebnisse können ebenfalls schon mit einer geringen Anzahl von Texten bei wenigen Autoren erzielt werden, so genügen drei Texte als Trainingsdaten, um eine Unterscheidung zwischen zwei Autoren herbeizuführen (ebd.). Weniger gute Ergebnisse zeigen sich bei einer Unterscheidung mehrerer Autoren mit einer geringen Anzahl von Texten. Bei 10 Autoren werden mindestens 15 Texte und bei 15 Autoren mindestens 30 Texte pro Person für ähnliche Ergebnisse benötigt. Für einen Anwendungsfall in der Praxis sind diese Ergebnisse deshalb mit Vorsicht zu betrachten. Zwar wurde bereits in Kapitel 3.1 angeführt, dass es einem gängigen Muster entspricht, dass sich ein Autor in einer eingrenzbaren Gruppe von Personen befindet, dennoch ist dabei unklar, wie groß diese Gruppen in der Praxis sind und wie viele Texte dieser Personen mit den in Kapitel 3.3 definierten Anforderungen als Datengrundlage existieren. Dies muss in einer gesonderten Betrachtung für den jeweiligen Fall erfolgen.

Weiterhin gestaltet sich eine Vergleichbarkeit verschiedener Verfahren aufgrund der stark voneinander abweichenden Datengrundlagen schwierig. Abgesehen von einem festgestellten „Rauschen“ in den Daten, welches sich durch einen schwer filterbaren Inhalt, welcher nicht von den Autoren stammt, äußert, wurden einige Anforderungen an die Datengrundlage im Rahmen dieser Arbeit definiert

(vgl. Kapitel 3.3). Diese Anforderungen sind zwar unabdingbar für einen Nachweis der grundsätzlichen Funktion der Merkmale, verleihen den Ergebnissen aber einen eher theoretischen Charakter. Eine Verifikation der gewonnenen Erkenntnisse an weiteren Datengrundlagen ist deshalb geboten.

Dieser Umstand soll allerdings keineswegs die Relevanz der gewonnenen Erkenntnisse über die untersuchten Merkmale im Rahmen dieser Arbeit schmälern. Bei der Untersuchung der „Wortlängen“ der Datengrundlage konnte eine Eigenheit der deutschen Sprache innerhalb eines Datensatzes von Musikreviews festgestellt werden. Ein Großteil der Autoren neigt dazu, sehr lange Komposita zu bilden, welche bis zu über 70 Zeichen betragen (vgl. Kapitel 5.6.4). Die daraus entstehende These, dass die Tendenz der Autoren zur Nutzung sehr langer Wörter mehr Aussagekraft als die reine Wortlänge besitzt, konnte in Kapitel 7.3.1 nachgewiesen werden. Als bestes Verfahren zeigte sich das Bilden einer Klasse für alle Wörter mit einer Länge von über 20 Zeichen (ebd.). In der englischen Sprache werden hingegen nur Worte mit einer Länge von 1-20 gezählt. Weiterhin konnten für den SVC bessere Klassifikationsergebnisse unter der Verwendung von „Zeichen-Affix-4-Grammen“, anstatt der im Englischen üblichen „3-Gramme“ erzielt werden (vgl. Kapitel 7.3.4). Dabei wird vermutet, dass längere „Zeichen-N-Gramme“ nötig sind, um die Informationen der durchschnittlich längeren deutschen Wörter zu erfassen (ebd.). Für alle anderen „Zeichen-N-Gramme“ und unter anderen Klassifikatoren konnte diese Vermutung allerdings nicht bestätigt werden. Insgesamt zeigten sich „Zeichen-Affix-4-Gramme“ als erfolgreichstes Merkmal in der Klassifikation. Unter der alleinigen Verwendung dieses Merkmals konnten bereits Genauigkeiten von 93,9% bei der Unterscheidung von 25 Autoren bei 100 Texten erreicht werden (vgl. Kapitel 8.5.4).

Anhand von Differenzfunktionen konnte nachgewiesen werden, dass die ohnehin stark kritisierten themenabhängigen Merkmale eine starke Korrelation zu den restlichen Merkmalen besitzen und deren Informationen auf andere Weise durch diese abgebildet werden (vgl. Kapitel 8.5.3). Dies hat zur Folge, dass ohne ein Verwenden aller themenabhängigen Merkmale nahezu identische Ergebnisse erzielt werden können (ebd.). Hierdurch konnte zum einen eine Reduktion von

rund 80% der Dimensionen des Modells erfolgen und zum anderen besteht die Vermutung, dass bei Texten mit einer eindeutigen Differenz der Thematik bessere Ergebnisse erzielt werden. Dies zeigt allerdings gleichermaßen einen Kritikpunkt an diesen Ergebnissen auf. Aufgrund der Datengrundlage kann keine eindeutige Aussage darüber getroffen werden, ob sich die verschiedenen Genres der Musikreviews unterscheiden (vgl. Kapitel 5.1). Eine starke Unterscheidung könnte eine schlechtere Leistung themenabhängiger Merkmale zur Folge haben. Weiterhin kann eine Themenabhängigkeit der „Zeichen-Affix-4-Gramme“ nicht ausgeschlossen werden. Wie in Kapitel 7.3.4 angeführt, wird mit einem größeren N mehr Kontext und dadurch Inhalt des Textes erfasst. Weiterhin wies das Merkmal die größte Korrelation zu den themenabhängigen Merkmalen auf (vgl. Kapitel 8.5.3).

Das Ziel der „Wortarten-N-Gramme“ ist das Erfassen der grammatikalischen Struktur eines Textes anhand von Phrasen (vgl. Kapitel 5.6.11). Insbesondere von langen „N-Grammen“ wurde sich ein Erfassen von autorenspezifischen Strukturen versprochen (ebd.). Anlehnend an existierende Arbeiten konnte allerdings nachgewiesen werden, dass Werte für ein N größer als zwei zunehmend geringere Klassifikationsergebnisse erzielen (vgl. Kapitel 7.3.3). Durch den starken Anstieg der Dimensionalität mit der Wahl größerer „N-Gramme“ ist deshalb von der Verwendung dieser abzusehen. Zwar konnten durch „Wortarten“ und „Wortarten-2-Gramme“ im Vergleich zu anderen Merkmalen gute Klassifikationsergebnisse erzielt werden, dennoch ist die ursprüngliche Annahme, dass die syntaktische Struktur eines Autors durch Wortarten-N-Gramme abgebildet werden kann, dadurch zu verneinen. Eine detaillierte Betrachtung der durch diese Annahme ausgeschlossenen Merkmale wie der „Satz- und Phasenstruktur“ oder der „Neuschreiberegeln“ erscheint daher vielversprechend (vgl. Kapitel 5.4).

Auch andere syntaktische Merkmale wie die in der Forschung oftmals verwendeten „Funktionswörter“ erwiesen nur mittelmäßige Ergebnisse (vgl. Kapitel 8.5.4). Als Begründung des Merkmals wird die starke Reflektion stilistischer Entscheidungen angeführt (vgl. Kapitel 5.6.9). Vielmehr besteht die Vermutung, dass diese stilistischen Entscheidungen besser durch „Zeichen-

Interpunktions-3-Gramme“ erfasst werden konnten. Diese lieferten nach den „Zeichen-Affix-4-Grammen“ die zweitbesten Ergebnisse (vgl. Kapitel 8.5.4).

Die Dimensionalität der Merkmale konnte bereits in der Vorverarbeitung deutlich reduziert werden. Ein Reduzieren der Merkmale um Ausprägungen mit einer geringen Häufigkeit erwies sich als essentielle Maßnahme (vgl. Kapitel 7.1.1). Das Reduzieren der Ausprägungen um Merkmale mit einem hohen Vorkommnis zeigte sich dagegen in Anbetracht der geringen Auswirkungen als irrelevant (vgl. Kapitel 7.1.2). Ferner konnte die Korrelation der Merkmale zur Textlänge durch das Bestimmen der individuellen relativen Häufigkeit signifikant reduziert werden (vgl. Kapitel 7.1.4). Weiterhin überzeugte die Merkmalsauswahl durch den in dieser Arbeit entwickelten iterativen Filter mit insgesamt sehr guten Ergebnissen. Hier konnte eine deutliche Reduktion der Merkmale unter einem in der Regel vernachlässigbaren Verlust an Genauigkeit erfolgen (vgl. Kapitel 7.2.2). Kritisch gilt hier die benötigte Rechenleistung des Verfahrens zu betrachten. Ein Durchführen der Methode war nur unter der Verwendung einer hochdimensionierten Hardware in annehmbarer Zeit durchführbar (vgl. Kapitel 9.1).

Zusammenfassend lässt sich die Forschungsfrage damit beantworten, dass das Unterscheiden von Autoren mithilfe von sprachlichen Merkmalen der existierenden Forschung in der deutschen Sprache möglich ist. Die Bewertung der genannten Merkmale konnte hierbei größtenteils bestätigt werden. Dennoch konnten diese Ergebnisse nur unter dem Ausschluss einiger Herausforderungen erzielt werden. Dabei erfolgte die Auswertung lediglich auf der Grundlage eines Datensatzes. Folglich können die Erkenntnisse dieser Arbeit als Grundlage für eine weitere Forschung betrachtet werden.

9.3. Ausblick

Wie bereits im vorherigen Kapitel dargelegt, sind die Ergebnisse dieser Arbeit als Grundlage eines weiten Forschungsgebietes zu verstehen. Dabei wird ein Verifizieren der gewonnenen Erkenntnisse an weiteren Datengrundlagen als unerlässlich angesehen. Weiterhin konnten die Ergebnisse dieser Arbeit nur unter dem Ausschluss einiger Herausforderungen erzielt werden, woraus sich weiterführende Forschungsfragen ergeben.

Insbesondere bei inkriminierten Texten kommt es häufig zu einer bewussten Verfälschung des Schreibstils, um die eigene Identität zu verschleiern oder den Verdacht von sich zu lenken (vgl. Kapitel 3.3). Eine Untersuchung des möglichen Erkennens der bewussten Verfälschung könnte aufschlussreiche Ergebnisse liefern und die Relevanz der automatisierten Autorschaftsanalyse im kriminalistischen Kontext steigern.

Sprachliche Differenzen innerhalb der Texte können zu Verfälschungen der Ergebnisse führen (vgl. Kapitel 3.3). Ein Betrachten dieser Auswirkungen auf die Merkmale und eine Analyse des Sprachstils könnte die Ergebnisse in multilingualen Datengrundlagen verbessern. Weiterhin könnte die Summe der Trainingsdaten eines Modells erhöht werden, falls eine Analyse über Dokumente unterschiedlicher Sprache eines Autors möglich ist.

Weitere Einschränkungen auf die Summe der vorhandenen Texte entstehen durch die Änderung des Schreibstils eines Autors über die Zeit. Demnach könnte ein Vergrößern der Datengrundlage in der Praxis durch eine Analyse der Änderung des Schreibstils über die Zeit erzielt werden. Mithilfe von stilistischen Merkmalen eines Autors, welche sich nicht über die Zeit ändern, könnte beispielsweise auf Daten eines Archivs zurückgegriffen werden.

Weiterhin werden die möglichen Anwendungsbereiche durch die Anforderung möglichst langer Texte eingeschränkt (Vgl. Kapitel 3.3). Beispielsweise finden sich in der modernen Kommunikation wie Direktnachrichten und E-Mail häufig eine Vielzahl von kurzen Nachrichten. Profilbasierte Modelle könnten hier durch das Zusammenfügen der Texte bessere Ergebnisse erzielen (vgl. Kapitel 3.2).

Weiterhin könnten die in Rahmen dieser Arbeit ausgeschlossenen strukturellen Merkmale zu annehmbaren Klassifikationsergebnissen beitragen.

Aufgrund der starken Korrelation zwischen themenabhängiger und anderer Merkmale konnten in dieser Arbeit ohne deren Verwendung ähnliche Ergebnisse erzielt werden. Hierbei kann aufgrund der Datengrundlage keine allgemeingültige Aussage über die Bewertung dieser getroffen werden. Ein Untersuchen dieser Merkmale anhand einer passenden Datengrundlage könnte hierbei abhelfen. Insgesamt haben sich im Rahmen dieser Arbeit Vermutungen über weitere vielversprechende Merkmale ergeben, welche positive Auswirkungen auf das Klassifikationsergebnis besitzen könnten.

Zusammenfassend stellt die Autorschaftsanalyse ein großes Forschungsgebiet dar, dessen Anwendungsgebiet insbesondere durch die hohen Anforderungen an die Daten eingeschränkt wird. Weitere Forschungen und Lösungen für das Minimieren dieser Anforderungen können dadurch die Relevanz in der Praxis steigern.

Literaturverzeichnis

Abbasi, A., & Chen, H.-c. (Oktober 2005). Chen, H.: Applying Authorship Analysis to Extremist-Group Web Forum Messages. (IEEE, Hrsg.) *IEEE Intelligent Systems, 20*(5), S. 67-75. doi:10.1109/MIS.2005.81

Abbasi, A., & Chen, H.-c. (März 2008). Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace. *ACM Transactions on Information Systems*(26), S. 1-29. doi:10.1145/1344411.1344413

Aihara, S., & Tanaka-Ishii, K. (Juli 2015). Computational Constancy Measures of Texts—Yule's K and R ´enyi's Entropy. *Computational Linguistics*(41), S. 1-22. doi:DOI: 10.1162/COLI_a_00228

Amstad, T. (1978). *Wie verständlich sind unsere Zeitungen?* Zürch: Studenten-Schreib-Service.

Argamon, S., Fine, J., & Shimoni, A. (Dezember 2003). Gender, Genre, and Writing Style in Formal Written Texts. *Interdisciplinary Journal for the Study of Discourse, 23*(3), S. 321–346. doi:10.1515/text.2003.014

Argamon, S., Saric, M., & Stein, S. S. (2003). Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (S. 475-480). Washington, DC. doi:10.1145/956750.956805

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., & Levitan, S. (26. Februar 2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology, 58*(6), S. 802-822. doi:10.1002/asi.20553

Azevedo, G. (2. August 2019). *Feature selection techniques for classification and Python tips for their application*. Abgerufen am 25. Dezember 2020 von towards data science: https://towardsdatascience.com/feature-selection-

techniques-for-classification-and-python-tips-for-their-application-10c0ddd7918b

Azevedo, G. (2. August 2019). *Feature selection techniques for classification and Python tips for their application*. Abgerufen am 25. Dezember 2020 von towards data science: https://towardsdatascience.com/feature-selection-techniques-for-classification-and-python-tips-for-their-application-10c0ddd7918b

Baeza-Yates, R., & Riberio-Neto, B. (1999). *Modern Information Retrieval.* Harlow: ACM press, A Division of the Association for Computing Machinary, Inc.

Bansal, S., & Aggarwal, C. (22. November 2020). *Textstat*. Von https://pypi.org/project/textstat/ abgerufen

Barrus, T. (23. Dezember 2020). *pyspellchecker Documentation*. Abgerufen am 27. Dezember 2020 von https://readthedocs.org/projects/pyspellchecker/downloads/pdf/latest/

Bauer, R. (6. September 2018). *Sie erforscht die Sprache des Bösen: So überführt Gudrun Müller Täter anhand ihrer Worte*. Abgerufen am 20. Dezember 2020 von Focus: https://www.focus.de/wissen/forensische-linguistik-gudrun-mueller-ueberfuehrt-taeter-anhand-ihrer-worte_id_9495954.html

Bernstein, B. (1. Oktober 1962). Social Class, Linguistic Codes and Grammatical Elements. *Language and Speech*(5), S. 221-240. doi:10.1177/002383096200500405

Bhandari, A. (3. April 2020). *Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization*. Abgerufen am 25. Dezember 2020 von Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Bishop, C. M. (2006). *Pattern Recognition and Maschine Learning.* Cambridge: Springer Science+Business Media, LLC.

Bonaccorso, G. (2017). *Machine Learning Algorithms.* Birmingham: Published by Packt Publishing Ltd.

Bonnin, R. (2017). *Machine Learning for Developers.* Birmingham: Packt Publishing Ltd.

Bönninghoff, B., Hessler, S., Kolossa, D., Kucharczik, K., Nickel, R., & Pittner, K. (11. November 2019). Autorschaftsanalyse Verstellungsstrategien und Möglichkeiten der automatisierten Erkennung. *Datenschutz und Datensicherheit - DuD, 43*, S. 691–699.

Boukkouri, H. E. (18. Juni 2018). *Text Classification: The First Step Toward NLP Mastery*. Abgerufen am 23. 12 2020 von Text Classification: The First Step Toward NLP Mastery: https://medium.com/data-from-the-trenches/text-classification-the-first-step-toward-nlp-mastery-f5f95d525d73

Bredthauer, S. (2013). *Verstellungen im inkriminierten Schreiben.* Springer Gabler, Wiesbaden. doi:10.1007/978-3-658-24324-1

Bromley, D. (1991). Aspects of written language production over adult life. *Psychology and Aging*(6), S. 296–308. doi:10.1037/0882-7974.6.2.296

Brownlee, J. (25. Oktober 2019). *Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning*. Abgerufen am 25. Dezember 2020 von Maschine Learning Mastery: https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/

Brownlee, J. (13. Juli 2020). *A Gentle Introduction to Information Entropy*. Abgerufen am 25. Dezember 2020 von Machine Learning Mastery: https://machinelearningmastery.com/what-is-information-entropy/

Brownlee, J. (3. August 2020). *A Gentle Introduction to k-fold Cross-Validation*. Abgerufen am 25. Dezember 2020 von Machine Learning Mastery: https://machinelearningmastery.com/k-fold-cross-

validation/#:~:text=Cross%2Dvalidation%20is%20a%20resampling,k%2Dfold%20cross%2Dvalidation.

Brownlee, J. (7. September 2020). *One-vs-Rest and One-vs-One for Multi-Class Classification*. Abgerufen am 25. Dezember 2020 von Machine Learning Mastery: https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/

Carstensen, K.-U., Ebert, C., Ebert, C., Jekat, S., Klabunde, R., & Langer, H. (2010). *Computerlinguistik und Sprachtechnologie.* Heidelberg: Spektrum Akademischer Verlag.

Cattell, R. (1978). *The Scientific Use of Factor Analysis.* New York: Plenum press.

Cheplyaka, R. (11. Dezember 2017). *Explained variance in PCA*. Abgerufen am 25. Dezember 2020 von https://ro-che.info/articles/2017-12-11-pca-explained-variance

Cheung, H., & Kemper, S. (Januar 1992). Competing Complexity Metrics and Adults' Production of Complex Sentences. *Applied Psycholinguistics*(13), S. 53-76. doi:10.1017/S0142716400005427

Coelho, L. P., & Richert, W. (2015). *Building Machine Learning Systems with Python* (2 Ausg.). Birmingham: Packt Publishing Ltd.

Collins English Dictionary. (2020). *Definition of 'phrase-structure rule'*. (H. Publishers, Herausgeber) Abgerufen am 25. Dezember 2020 von https://www.collinsdictionary.com/dictionary/english/phrase-structure-rule

Dern, C. (2003). Sprachwissenschaft und Kriminalistik: zur Praxis der Autorenerkennung. (D. Gruyter, Hrsg.) *Zeitschrift für germanistische Linguistik*(31), S. 44-77.

Dern, C. (2009). *Autorenerkennung Theorie und Praxis der linguistischen Tatschreibeanalyse.* Stuttgart: Richard Boorberg Verlag GmbH & Co KG.

Dinov, I. (2018). *Data Science and Predictive Analytics.* (C. Springer, Hrsg.) doi:10.1007%2F978-3-319-72347-1

Domingos , P., & Pazzani , M. (November 1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning, 29*, S. 103-130.

Dresp-Langley, B., Ekseth, O. K., Fesl, J., Gohshi, S., Kurz, M., & Sehring, H.-W. (Juli 2019). Occam's Razor for Big Data? On Detecting Quality in Large Unstructured Datasets. *Applied Sciences, 9*(15), S. 2. doi:10.3390/app9153065

Drommel, R. (2016). *Sprachprofiling - Grundlagen und Fallanalysen zur Forensischen Linguistik* (Bd. 126). Berlin: Frank & Timme.

DuBay, W. (25. August 2004). The Principles of Readability.

DuBay, W. (2006). *The Classic Readability Studies.*

Duque, T. (2. April 2020). *Text Normalization*. Abgerufen am 23. Dezember 2020 von https://towardsdatascience.com/text-normalization-7ecc8e084e31

DWDS. (2020a). *Morphologie, in: DWDS – Digitales Wörterbuch der deutschen Sprache*. Abgerufen am 15. Dezember 2020 von https://www.dwds.de/wb/Morphologie

DWDS. (2020b). *orthografisch, in: DWDS – Digitales Wörterbuch der deutschen Sprache*. Abgerufen am 15. Dezember 2020 von https://www.dwds.de/wb/orthografisch

DWDS. (2020c). *Inhaltswort, in: DWDS – Digitales Wörterbuch der deutschen Sprache*. (B.-B. A. Wissenschaften, Herausgeber) Abgerufen am 23. Dezember 2020 von https://www.dwds.de/wb/Inhaltswort

Edpresso Team. (2020). *What is matplotlib?* Abgerufen am 27. Dezember 2020 von edpresso: https://www.educative.io/edpresso/what-is-matplotlib

Efron, M., Marchionini, G., & Zhiang, J. (2003). Implications of the Recursive Representation Problem for Automatic Concept Identification in On-line Governmental Information. Von https://www.researchgate.net/publication/228584435_Implications_of_th

e_recursive_representation_problem_for_automatic_concept_identification_in_on-line_governmental_information abgerufen

Everitt, B. (März 1975). Multivariate analysis: the need for data, and other problems. *The British Journal of Psychiatry, 126*(3), S. 237 - 240. doi:10.1192/bjp.126.3.237

Explosion. (2020a). *spaCy - Annotation Specifications*. Abgerufen am 23. Dezember 2020 von https://spacy.io/api/annotation

Explosion. (2020b). *spaCy - Facts & Figures*. Abgerufen am 23. Dezember 2020 von https://spacy.io/usage/facts-figures

Explosion. (2020c). *spaCy 101: Everything you need to know*. Abgerufen am 23. 12 2020 von https://spacy.io/usage/spacy-101

Feiguina, O., & Hirst, G. (2007). Authorship Attribution for Small Texts: Literary and Forensic Experiments. *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007.* Amsterdam.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), S. 221–233. doi:10.1037/h0057532

Forman, G. (2002). Choose Your Words Carefully: An Empirical Study of Feature Selection Metrics for Text Classification. *Principles of Data Mining and Knowledge Discovery*, (S. 150-162). Helsinki. doi:10.1007/3-540-45681-3_13

Fraj, M. (5. Januar 2018). *In Depth: Parameter tuning for SVC*. Abgerufen am 26. Dezember 2020 von medium: https://medium.com/all-things-ai/in-depth-parameter-tuning-for-svc-758215394769

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, (S. 611–617). doi:10.3115/1220355.1220443

Goodfellow , I., Bengio , Y., & Courville, A. (2016). *Deep Learning.* MIT Press. Von http://www.deeplearningbook.org abgerufen

Gorsuch, R. (1983). *Factor analysis.* Hillsdale: Lawrence Erlbaum Associates.

Grieve, J. (26. Juli 2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing, 22*(3), S. 251–270.

Guenter, S., & Sanderson, C. (2006). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (S. 482-491). Sydney. doi:10.3115/1610075.1610142

Gupta, P. (5. Juni 2017). *Cross-Validation in Machine Learning*. Abgerufen am 25. Dezember 2020 von towardsdatascience: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

Halvani, O., & Steinebach, M. (2014). Autorschaftsanalyse — die Illusion der Anonymität. *Wirtschaftsinformatik & Management, 6*, S. 33–43.

Halvani, O., Steinebach, M., & Neitzel, S. (2014). Lässt sich der Schreibstil verfälschen um die eigene Anonymität in Textdokumenten zu schützen? (S. 229-241). Bonn: Gesellschaft für Informatik e.V.

Henschel, S. (12. April 2019). *Wie künstlich intelligent ist Computerlinguistik?* Abgerufen am 23. Dezember 2020 von https://blog.sbb.berlin/wie-kuenstlich-intelligent-ist-computerlinguistik/

Huilgol, P. (24. August 2019). *Accuracy vs. F1-Score*. Abgerufen am 25. Dezember 2020 von Analytics Vidhya: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

IBM. (2020). *IBM Power System AC922 Systems Hardware Datenbroschüre*. Abgerufen am 25. Dezember 2020 von IBM Hompage: https://www.ibm.com/downloads/cas/6PRDKRJ0

Igual, L., & Seguí, S. (2017). *Introduction to Data Science.* Schweiz: Springer International Publishing. doi:10.1007/978-3-319-50017-1

Imo, W. (2016). *Grammatik.* Stuttgart: J.B. Metzler Verlag GmbH.

intersoft consulting. (2020). *Art. 4 DSGVO Begriffsbestimmungen*. Abgerufen am 12. Dezember 2020 von DSGVO: https://dsgvo-gesetz.de/art-4-dsgvo/

Jaju, S. (22. Januar 2017). *Comprehensive Guide on t-SNE algorithm with implementation in R & Python*. Abgerufen am 25. Dezember 2020 von Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/

Kämper, H. (1996). Nachweis der Autorenschaft. *Kriminalistik, 50*(8-9), S. 561-566.

Kaster, A., Siersdorfer, S., & Weikum, G. (13-19. August 2005). Combining Text and Linguistic Document Representations for Authorship Attribution. *In SIGIR Workshop: Stylistic Analysis of Text for Information Access (STYLE)*, (S. 27-35). Salvador (Bahia). doi:10.1.1.84.4407

Kaushik, S. (1. Dezember 2016). *Introduction to Feature Selection methods with an example (or how to select the right variables?)*. Abgerufen am 25. Dezember 2020 von Analytics Vidhya: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

Kemper, S., & Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychology and Aging*(16), S. 312-322. doi:10.1037/0882-7974.16.2.312

Kemper, S., Kynette, D., Rash, S., O'Brien, K., & Sprott, R. (1989). Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics*(10), S. 49–66. doi:10.1017/S0142716400008419

Kipp, C. (11. September 2018). *Einstieg in Natural Language Processing – Artikelserie*. Abgerufen am 23. Dezember 2020 von https://data-science-

blog.com/blog/2018/09/11/einstieg-in-natural-language-processing-artikelserie/

Kitson, H. (1921). *The Mind of the Buyer.* (Macmillan, Hrsg.) New York.

Koppel, M., & Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, (S. 69-72).

Koppel, M., Akiva, N., & Dagan, I. (14. Juli 2006). Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology, 57*(11), S. 1519-1525. doi:10.1002/asi.20428

Koppel, M., Argamon, S., & Shimoni, A. (Dezember 2003). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing, 17*(4), S. 401-412.

Koppel, M., Schler, J., & Argamon, S. (März 2011). Authorship attribution in the wild. *Language Resources and Evaluation*(45), S. 83-94. doi:10.1007/s10579-009-9111-2

Kriesel, D. (29. Dezember 2016). *Spiegel Mining*. Abgerufen am 23. Dezember 2020 von http://www.dkriesel.com/spiegelmining

Kruczek, J., Kruczek, P., & Kuta, M. (2020). Are n-gram Categories Helpful in Text Classification? *Computational Science – ICCS 2020*, (S. 524-537). Amsterdam. doi:10.1007/978-3-030-50417-5_39

Lewis, D., Russel-Rose, T., Yang, Y., Li, F., & Dietterich, G. (April 2004). RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*(5), S. 361-397.

Liu, C. (10. März 2020). *SVM Hyperparameter Tuning using GridSearchCV*. Abgerufen am 26. Dezember 2020 von Velocity: https://www.vebuso.com/2020/03/svm-hyperparameter-tuning-using-gridsearchcv/

Liu, H., & Dash, M. (1997). Feature selection for classification. *Intelligent Data Analysis, 1*, S. 131-156. doi:10.1016/S1088-467X(97)00008-5

Liu, H., & Jiang, N. (2013). Understand System's Relative Effectiveness Using Adapted Confusion Matrix. *Design, User Experience, and Usability. Design Philosophy, Methods, and Tools*, (S. 294-302). Las Vegas.

Liu, Y. H. (2019). *Python Machine Learning By Example* (2 Ausg.). Birmingham: Packt Publishing Ltd.

Luber, S., & Litzel, N. (15. Februar 2018). *Was ist ein Neuronales Netz?* Abgerufen am 25. Dezember 2020 von Big Data Insider: https://www.bigdata-insider.de/was-ist-ein-neuronales-netz-a-686185/

MacCallum, R., Widaman, K., Zhang, S., & Hong, S. (März 1999). Sample Size in Factor Analysis. *Psychological Methods, 4*(1), S. 84-99. doi:10.1037/1082-989X.4.1.84

Madigan, D., Genkin , A., Lewis, D., Genkin , E., Argamon , S., Fradkin , D., & Ye , L. (2005). Scale, Author Identification on the Large. *In Proc. of the Meeting of the Classification Society of North America.*

Mendenhall, T. (11. März 1887). THE CHARACTERISTIC CURVES OF COMPOSITION. *Science, 9*(214), S. 237-246. doi:10.1126/science.ns-9.214S.237

Misra, R. (1. Mai 2019). *Support Vector Machines — Soft Margin Formulation and Kernel Trick*. Abgerufen am 25. Dezember 2020 von towards data science: https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series).* The MIT Press.

Mutjaba, H. (24. September 2020). *What is Cross Validation in Machine learning? Types of Cross Validation*. Abgerufen am 25. Dezember 2020 von greatlearningblog: https://www.mygreatlearning.com/blog/cross-validation/

Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E., & Song, D. (2012). On the Feasibility of Internet-Scale Author Identification. *2012 IEEE Symposium on Security and Privacy* (S. 300-314). San Francisco: IEEE. doi:10.1109/SP.2012.46

Nini, A. (26. Februar 2015). *Authorship profiling in a forensic context*.

NLTK.org. (2020a). *NLTK 3.5 documentation*. Abgerufen am 23. Dezember 2020 von https://www.nltk.org/

NLTK.org. (2020b). *NLTK Corpora*. Abgerufen am 23. Dezember 2020 von http://www.nltk.org/nltk_data/

Oren, H. (2012). Autorschaftsanalyse im Kontext der Attribution, Verifikation und intrinsischer Exploration (Master-Thesis). Technische Universität Darmstadt. Von http://tubiblio.ulb.tu-darmstadt.de/102054/ abgerufen

pandas development team. (2020). *Package overview*. Abgerufen am 26. Dezember 2020 von Pandas: https://pandas.pydata.org/docs/getting_started/overview.html

Peng, F., Schuurmans, D., & Wang, S. (September 2004). Augmenting Naive Bayes Classifiers with Statistical Language Models. *Information Retrieval Journal, 7*(3), S. 317–345. doi:10.1023/B:INRT.0000011209.19643.e2

Pennebaker, J., & Stone, L. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology, 85*(2), S. 291–301. doi:10.1037/0022-3514.85.2.291

Potha, N., & Stamatatos, E. (2014). A Profile-Based Method for Authorship Verification. *Hellenic Conference on Artificial Intelligence*, (S. 313-326). Karlovassi. doi:10.1007/978-3-319-07064-3_25

Prateek, J. (2017). *Artificial Intelligence with Python.* Birmingham: Packt Publishing Ltd.

Rabaglia, C., & Salthouse, T. (Dezember 2011). Natural and constrained language production as a function of age and cognitive abilities. *Language*

and Cognitive Processes , S. 1505-1531. doi:10.1080/01690965.2010.507489

Rao, T., Shaukat, S., & Khan, M. (28. Mai 2016). Impact of sample size on principal component analysis ordination of an environmental data set: effects on eigenstructure. *Ekológia (Bratislava), 35*(2), S. 173-190. doi:10.1515/eko-2016-0014

Raschka, S. (Oktober 2014). Naive Bayes and Text Classification I - Introduction and Theory. doi:10.13140/2.1.2018.3049

Raschka, S. (2016). *Python Machine Learning.* Birmingham: Packt Publishing Ltd.

Rexha, A., Kröll, M., Ziak, H., & Kern, R. (2. Februar 2018). Authorship identification of documents with high content similarity. *Scientometrics, 115*, S. 223–237. doi:10.1007/s11192-018-2661-6

Rish, I. (2001). An empirical study of the naive bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, (S. 41-46).

Rojas, R. (15. Februar 2015). The Curse of Dimensionality. Von https://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/dimensionality.pdf abgerufen

Rudman, J. (1998). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities, 31*(4), S. 351-365.

Saltz, J., & Hotz, N. J. (14. Dezember 2020). *What is CRISP DM?* Abgerufen am 23. Dezember 2020 von https://www.datascience-pm.com/crisp-dm-2/

Sanderson, C., & Guenter, S. (Januar 2006). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. *Conference: International Conference on Empirical Methods in Natural Language Processing (EMNLP*, S. 482-491. doi:10.3115/1610075.1610142

Sapkota, U., Bethard, S., Montes, M., & Solorio, T. (2015). Not All Character N-grams Are Created Equal: A Study in Authorship Attribution. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (S. 93–102). Denver, Colorado: Association for Computational Linguistics. doi:10.3115/v1/N15-1010

Schmid, H.-J. (2002). Do women and men really live in different cultures? Evidence from the BNC. In A. Wilson, P. Rayson, & P. J. Melia, *Corpus linguistics by the Lune: a Festschrift for Geoffrey Leech* (S. 185–221). Frankfurt: Peter Lang.

Schneider, A. (2013). *Eigenwerte und Eigenvektoren*. Abgerufen am 25. Dezember 2020 von Mathebibel: https://www.mathebibel.de/eigenwerte-eigenvektoren

scikit-learn developers. (2020a). *Feature selection*. Abgerufen am 25. Dezember 2020 von scikit learn: https://scikit-learn.org/stable/modules/feature_selection.html

scikit-learn developers. (2020b). *sklearn.feature_selection.SelectKBest*. Abgerufen am 26. Dezember 2020 von scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

scikit-learn developers. (2020c). *sklearn.model_selection.GridSearchCV*. Abgerufen am 25. Dezember 2020 von scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

scikit-learn developers. (2020d). *sklearn.model_selection.train_test_split*. Abgerufen am 26. Dezember 2020 von scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

scikit-learn developers. (2020e). *sklearn.neighbors.KNeighborsClassifier*. Abgerufen am 25. Dezember 2020 von scikit learn: https://scikit-

learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

scikit-learn developers. (2020f). *sklearn.preprocessing.MinMaxScale*. Abgerufen am 25. Dezember 2020 von scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

scikit-learn developers. (2020g). *sklearn.preprocessing.StandardScaler*. Abgerufen am 25. Dezember 2020 von scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

scikit-learn developers. (2020h). *Getting Started*. Abgerufen am 27. Dezember 2020 von scikit-learn: https://scikit-learn.org/stable/getting_started.html

scikit-learn developers. (2020i). *Support Vector Machines¶*. Abgerufen am 25. Dezember 2020 von scikit learn: https://scikit-learn.org/stable/modules/svm.html

Shearer, C. (Herbst 2000). The CRISP-DM Model: The New Blueprint for Data. *Journal of Data Warehousing*(5), S. 13-23.

Singh, S. (21. Mai 2018). *Understanding the Bias-Variance Tradeoff*. Abgerufen am 25. Dezember 2020 von towardsdatascience: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Stadler, M.-L. (29. Januar 2020). *Natural Language Processing*. Abgerufen am 26. Dezember 2020 von mindsquare: https://mindsquare.de/knowhow/natural-language-processing/

Stamatatos, E. (16. Dezember 2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), S. 538-556. doi:10.1002/asi.21001

Stamatatos, E. (Januar 2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*(21), S. 421-439.

Stolerman, A., Overdorf, R., Afroz, S., & Greenstadt, R. (2014). Breaking the Closed-World Assumption in Stylometric Authorship Attribution. *DigitalForensics 2014: Advances in Digital Forensics X*, (S. 185-205). doi:10.1007/978-3-662-44952-3_13

The SciPy community. (29. Juni 2020). *What is NumPy?* Abgerufen am 26. Dezember 2020 von numpy.org: https://numpy.org/doc/stable/user/whatisnumpy.html

van Halteren, H., Tweedie, F., & Bayeen, H. (1. September 1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, 11*(3), S. 121-132. doi:10.1093/llc/11.3.121

Vijayan, V., & Govilkar, S. (31. März 2019). A Survey on Author Profiling Techniques. (I. I. IJCSE, Hrsg.) *International Journal of Computer Sciences and Engineering*, S. 1065-1069. doi:10.26438/ijcse/v7i3.10651069

Vorhies, W. (26. Juli 2016). *CRISP-DM – a Standard Methodology to Ensure a Good Outcome*. Abgerufen am 23. Dezember 2020 von https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome

Wallace, D., & Mosteller, F. (1964). *Inference & Disputed Authorship: The Federalist.* (A. Wesley, Hrsg.)

Webis Group. (2020a). *PAN at CLEF*. Abgerufen am 12. Dezember 2020 von PAN: https://pan.webis.de/events.html

Webis group. (2020b). *Shared Tasks*. Abgerufen am 12. Dezember 2020 von PAN: https://pan.webis.de/shared-tasks.html#authorship-analysis

Webis group. (2020c). *Publications*. Abgerufen am 12. Dezember 2020 von PAN: https://pan.webis.de/publications.html

Yang, Y., & Pedersen, J. (1997). A Comparative Study on Feature Selection in Text Categorization. *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, (S. 412-420).

Yule , G. U. (1944). *The Statistical Study of Literary Vocabulary.* Cambrige University Press.

Yule, G. U. (1. Januar 1939). On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed Autorship. *Biometrika, 30*(3-4), S. 363–390. doi:10.1093/biomet/30.3-4.363

Zhao, Y., & Zobel, J. (2005). Effective and Scalable Authorship Attribution Using Function Words. *Asia Information Retrieval Symposium*, (S. 174-189). Jeju Island, Korea. doi:10.1007/11562382_14

Zipf, G. (1932). *Selected Studies of the Principle of Relative Frequency in Language.* Cambrige: Harvard University Press.

Bildverzeichnis

Bild 1: Lebenszyklus nach CRISP-DM 13
Bild 2: Durchschnittliche Anzahl der Wörter 42
Bild 3: Wortlängen der Artikel 44
Bild 4: Wortschatzreichtum der Autoren im Vergleich 48
Bild 5: Ausreißer Wortschatzreichtum 49
Bild 6: Durchschnittlicher FRE der Autoren 66
Bild 7: Hinzufügen Datenpunkt KNN 75
Bild 8: Klassifikation mittels SVC 80
Bild 9: Soft and Hard Margin SVC 82
Bild 10: Über- und Unteranpassung 85
Bild 11: Kreuzvalidierungsverfahren 91
Bild 12: Ablauf Wrapper 97
Bild 13: Hauptkomponenten im zweidimensionalen Raum 99
Bild 14 Hauptkomponentenanalyse erklärte Varianz 102
Bild 15: Ablaufdiagram iterativer Filter 114
Bild 16: Genauigkeitskurve Zeichen-Affix-3-Gramme 116
Bild 17: Genauigkeitskurve Funktionswörter 116
Bild 18: Bewertung Extraktionsmethode Wortlänge 123
Bild 19: Bewertung Extraktionsmethode Ziffern 125
Bild 20: Bewertung von Wort-2- und Wort-3-Grammen 127
Bild 21: Bewertung Zeichen-Affix-N-Gramme GNB 129
Bild 22: Bewertung Zeichen-Affix-N-Gramme SVC 129
Bild 23: Bewertung Zeichen-Affix-N-Gramme KNN 130
Bild 24: Bewertung Zeichen-Wort-N-Gramme 131
Bild 25: Bewertung Zeichen-Interpunktion-N-Gramme 132
Bild 26: Bewertung Wortarten-N-Gramme 133
Bild 27: Bewertung des Modells unter verschiedenen Skalierungen 139
Bild 28: Vergleich der Modelle mit und ohne themenabhängiger Merkmale 144

Tabellenverzeichnis

Tabelle 1: Bekannte Merkmale der Literatur 32
Tabelle 2: Top 5 der häufigsten Wörter nach Autoren 38
Tabelle 3: Top 5 Wort-N-Gramme 40
Tabelle 4: Hitzekarte Sonderzeichen 50
Tabelle 5: Zeichen-N-Gramme 55
Tabelle 6: Hitzekarte Ziffern 56
Tabelle 7: Top 5 Funktionswörter der Autoren 57
Tabelle 8: Hitzekarte Wortarten 60
Tabelle 9: Wortarten-N-Gramme 61
Tabelle 10: Wortarten Satzanfänge und –endungen 63
Tabelle 11: Übersetzung FRE 65
Tabelle 12: Konfusionsmatrix binäre Klassifikation 86
Tabelle 13: Konfusionsmatrix drei Klassen 87
Tabelle 14: Schnittmengen Klasse 1 87
Tabelle 15: Dimensionalität Filter geringe Häufigkeit 107
Tabelle 16: Dimensionalität Filter hohe Häufigkeit 109
Tabelle 17: Korrelationskoeffizienten relative und absolute Häufigkeit 111
Tabelle 18: Merkmale vor und nach dem iterativen Filter 119
Tabelle 19: Bewertung Wort-N-Gramme mit N von 4-6 126
Tabelle 20: Zusammenfassung der verwendeten Merkmale 137
Tabelle 21: Bewertung der individuellen Merkmale 141
Tabelle 22: Werte der Differenzfunktionen 143
Tabelle 23: Werte der Differenzfunktionen themenunabhängiger Merkmale 145
Tabelle 24: F1-Werte des Modells 147

Anlagen

A1 Mailverkehr: Re Masterarbeit Autorenschaftserkennung.msg

Die Datei „Re Masterarbeit Autorenschaftserkennung.msg“ beinhaltet den Mailverkehr mit dem Chefredakteur der Seite http://musikreviews.de Thoralf Koß. Unter anderem beinhaltet dieser Verlauf ein Einverständnis die Artikel der Seite im Rahmen der Forschung zu verwenden und zum anderen eine Bestätigung, dass die Artikel allein von einem Autor erstellt werden.

A2 Python Skript: Scraper.py

Bei dem Python Skript “Scraper.py“ handelt es sich um einen Webscraper für die Webseite http://musikreviews.de/. Er basiert auf der Bibliothek „BeautifulSoup“. Die Aufgabe ist das Scrollen durch das gesamte Archiv der Webseite und das Erfassen aller Artikel. Es werden der Titel, Link, Autor, Genre, Erscheinungsdatum und der Text des Artikels in eine CSV-Datei gespeichert. Anhand des Erscheinungsdatums wird erkannt welche Artikel bereits in der CSV-Datei vorhanden sind und es werden nur neue Artikel hinzugefügt. Die CSV-Datei bildet die Datengrundlage für weitere Untersuchungen.

A3 Python Skript: Extractor.py

Das Python Skript „Extractor.py“ beinhaltet alle Funktionen für das Extrahieren der Merkmale aus der Datengrundlage. Die verwendeten Werkzeuge sind hierbei Natural Language Processing, Reguläre Ausdrücke und Funktionen der Bibliotheken numpy und pandas. Die extrahierten Merkmale werden wiederum in CSV-Dateien gespeichert.

Neben den Extraktionsfunktionen beinhaltet das Skript Methoden zum Pseudonymisieren der Autoren und dem Selektieren von Artikeln auf Basis der Autoren- und Textanzahl.

Weiterhin basieren alle Auswertungen in Kapitel 5.6 auf diesem Skript. Die Auswertungen wurden auskommentiert, um die Möglichkeit zu bewahren, nur einzelne Auswertungen durchzuführen.

Alle Methoden und Auswertungen wurden entsprechend kommentiert und besitzen einen Verweis auf das relevante Kapitel, Tabelle oder Schaubild. Für die Suche des Codes zu einem spezifischen Kapitel kann deshalb nach der Kapitelnummer gesucht werden.

A4 Python Skript: Processor.py

Das Python Skript „Processor.py“ beinhaltet alle ausgeführten Schritte der Datenvorbereitung und der Modellierung. Es verwendet dabei die in Kapitel 6.1 vorgestellten Bibliotheken und greift für notwendigen Datenextraktionen auf Methoden des Skriptes Extraktor.py zurück.

Alle Methoden und Auswertungen wurden entsprechend kommentiert und besitzen einen Verweis auf das relevante Kapitel, Tabelle oder Schaubild.

Für ein Suchen des Codes zu einem spezifischen Kapitel kann deshalb nach der Kapitelnummer gesucht werden. Alle Schritte sind auskommentiert, um die Möglichkeit zu bewahren, nur einzelne Auswertungen durchzuführen. Ein Durchführen aller Schritte in führt zu einer langen Laufzeit.

A5 CSV-Datei: musikreviews.csv

Die CSV-Datei „musikreviews.csv“ beinhaltet alle mithilfe des Webscrapers extrahierten Daten der Seite http://musikreviews.de/. Es werden der Titel, Link, Autor, Genre, Erscheinungsdatum und der Text des Artikels in eine CSV-Datei gespeichert.

A6 CSV-Datei: musikreviews_balanced_authors.csv

Die CSV-Datei “musikreviews_balanced_authors.csv” wird anhand der Datei „musikreviews.csv“ erstellt. Sie enthält 100 Artikel eines jeden Autoren, welcher

mindestens 100 Artikel besitzt. Anstatt eines Autors wird eine pseudonymisierte Nummer für diesen Autor gespeichert. Zusätzlich enthält die Datei das Genre, Erscheinungsdatum und den Text eines jeden Artikels.

A7 Daten

Der Ordner „daten“ beinhaltet alle in dieser Arbeit erzeugten und verwendeten Daten der Skripte „Extractor.py“ und „Processor.py“. Für eine erleichterte Übersicht erfolgt eine Untergliederung der Daten in Unterordner. Die Unterordner umfassen dabei alle benötigten Ein- und Ausgabedaten der Arbeitsschritte. Da die Ausgabe eines Arbeitsschrittes die Eingabe eines anderen darstellen kann, können hierbei bewusste Redundanzen in den Daten auftreten. Folgend werden die Unterordner und die darin enthaltenen Daten beschrieben.

- „1_raw“: Enthält die Daten aller Merkmale nach der erstmaligen Extraktion des Skriptes „Extractor.py“
- „2_filter_low_occurrence“: Enthält die Merkmale nach dem Filtern der Ausprägungen mit einer geringen Häufigkeit (vgl. Kapitel 7.1.1).
- „3_fiter_high_frequency“: Enthält alle Merkmale nach dem Filtern einer hohen Dokumentenhäufigkeit (vgl. Kapitel 7.1.27.1.1).
- „4_relative_frequency“: Enthält alle Merkmale nach dem bestimmen der individuellen relativen Häufigkeit (vgl. Kapitel 7.1.4). Die Eingabedaten sind dabei abhängig vom Wert, zu welchem die relative Häufigkeit bestimmt wird, auf die Unterordner „char_count“, „word_count“ und „sentence_count“ aufgeteilt.

In allen weiteren Arbeitsschritten wird der iterative Filter durchlaufen (vgl. Kapitel 7.2.1). Der Filter setzt dabei eine bestimmte Ordnerstruktur für Ein- und Ausgabedaten voraus:

- „accuracy“: Enthält jeweils eine Datei für ein Merkmal, welche die Genauigkeitswerte der Klassifikatoren in den einzelnen Durchläufen des iterativen Filters beinhaltet.

- „best_perc“: Enthält jeweils eine Datei für die Kombination aus dem ermittelten prozentualen Anteil und Genauigkeitswert des Merkmals.
- „csv_after_filter“: Enthält die Merkmalsdateien nach dem Durchlaufen des iterativen Filters.
- „csv_before_filter“: Enthält die Merkmalsdateien vor dem Durchlaufen des iterativen Filters.
- „diagrams“: Für jedes Merkmal wird bei einem Durchlaufen des iterativen Filters ein Diagramm erstellt. Dieses stellt den Verlauf der Genauigkeit bei einer steigenden Anzahl der Merkmale dar. Das ermittelte Perzentil wird mit einem „X“ markiert.
- „results“: Enthält die Ergebnisse des Arbeitsschrittes.
- „selected_features“: Enthält eine Datei mit einer Auflistung der jeweils selektierten Ausprägungen für jedes Merkmal.

Wird in einem Arbeitsabschnitt eine variierende Anzahl von Autoren oder Texten bearbeitet, erhalten die Dateien einen Prefix nach der Konvention: „a<Autorenanzahl>_t<Textanzahl>_“.

- „5_iterative_filter“: Enthält die Daten für das Durchlaufen des iterativen Filters für ein jedes Merkmal und die Genauigkeitswerte des Merkmals vor und nach dem Durchlauf (vgl. Kapitel 7.2.2).
- „6_feature_analysis_char_n_grams“: Enthält die Ergebnisse und Daten für den Vergleich der Zeichen-N-Gramme.
- „6_feature_analysis_pos_tag_n_grams“: Enthält die Ergebnisse und Daten für den Vergleich der Wortarten-N-Gramme. Durch die gleiche Vorgehensweise erhalten dieser und der vorherige Schritt dieselbe führende Nummerierung.
- „7_modeling“: Enthält alle benötigten Daten und Ergebnisse der durchgeführten Auswertungen in der Modellierung (vgl. Kapitel 8).

Verzeichnis der Abkürzungen

BoW	Bag-of-Words
CRISP-DM	Cross Industry Standard Process for Data Mining
FRE	Flesch Reading Ease Score
GNB	Gaussian Naïve Bayes
KNN	K-Nearest-Neighbor
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OvO	One-vs-One
OvR	One-vs-Rest
PC	Hauptkomponenten (engl. principal components)
PCA	Hauptkomponentenanalyse (engl. principal component analysis)
PoS-Tagging	Part-of-speech-tagging
RBF	Radical Basis Function
SVC	Support Vector Classifier
SVM	Support Vector Machine

Thesen

- Eine Klassifikation eines Autors anhand von Merkmalen des Textes ist möglich.
- Zeichen-Affixe-N-Gramme erweisen sich als bestes Merkmal der Klassifikation.
- Eine Klassifikation mit Support Vector Machines zeigt bessere Ergebnisse als Gaussian Naïve Bayes oder K-Nearest-Neighbor.
- Die Vielzahl an Merkmalsausprägungen lassen sich effektiv vor einer Klassifikation minimieren.
- Wortarten-N-Gramme großer Länge können die syntaktische Struktur des Textes nicht erfassen.
- Themenabhängige Merkmale weisen eine starke Korrelation zu anderen Merkmalen auf und können deshalb ohne relevante Genauigkeitsverluste des Klassifikationsmodells außer Acht gelassen werden.
- Die Merkmale weißen eine starke Korrelation zueinander auf.
- Durch das Bestimmen der individuellen relativen Häufigkeit der Merkmale kann die Abhängigkeit zur Textlänge minimiert werden.
- Größere Zeichen-N-Gramme liefern in der deutschen Sprache keine besseren Ergebnisse.
- Die Tendenz eines Autors sehr Lange Worte zu schreiben sagt mehr über seinen Schreibstil aus, als die genaue Länge dieser langen Worte.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe.
Die Stellen der Arbeit, die anderen Quellen im Wortlaut oder dem Sinn nach entnommen wurden, sind durch Angaben der Herkunft kenntlich gemacht. Dies gilt auch für Zeichnungen, Skizzen, bildliche Darstellungen sowie für Quellen aus dem Internet.

Ich erkläre ferner, dass ich die vorliegende Arbeit in keinem anderen Prüfungsverfahren als Prüfungsarbeit eingereicht habe oder einreichen werde.

Die eingereichte schriftliche Fassung entspricht der auf dem Medium gespeicherten Fassung.

Östringen, den 07.01.2021

Ort, Datum (Unterschrift)