Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Masterarbeit

Andy Herzberg

Analyse der oberflächlichen Merkmale von Qualitätsjournalismus-Texten

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Andy Herzberg

Analyse der oberflächlichen Merkmale von Qualitätsjournalismus-Texten

Abschlussarbeit zum Erlangen des akademischen Grades Master of Arts
im Studiengang Next Media
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Prof. Dr. Tim Tiedemann
Fachbetreuer: Dr. Hannes Schettler

Eingereicht am 20. Dezember 2018

Andy Herzberg

Thema der Arbeit

Analyse der oberflächlichen Merkmale von Qualitätsjournalismus-Texten

Stichworte

Text Mining, Klassifikation, Textqualität, Qualitätsjournalismus

Kurzzusammenfassung

Seit mehreren Jahrzehnten werden umfassend Merkmale erforscht, die gut lesbaren Text ausmachen. Der Fokus dieser Studien liegt meist auf der Klassifikation von Lesematerial für geeignete Schulstufen oder Bildungsgrade beim Fremdsprachenerwerb. Anders als die bisherigen Studien untersucht die vorliegende Arbeit die oberflächlichen Textmerkmale, die einen Einfluss auf menschlich kuratierten Qualitätsjournalismus haben. Da sich kaum Studien mit diesem Schwerpunkt auf die deutsche Sprache konzentrieren, soll diese Arbeit dazu beitragen die Forschungslücke zu schließen. Zu diesem Zweck wird eine deutschsprachige Textsammlung mit preisgekrönten Reportagen aufgebaut und nach der Extraktion der Merkmale mittels Machine Learning Verfahren untersucht. Mit traditionellem Feature Engineering wurde ohne Parameter-Tuning eine Klassifikationsgenauigkeit von circa 70% erreicht. Googles Cloud-Produkt „AutoML“, ein selbstlernendes neuronales Netz, erreichte sogar eine Klassifikationsgenauigkeit von 82%. Die Ergebnisse zeigen, dass es oberflächliche Textmerkmale gibt, die preisgekrönte Reportagen ausmachen und die für diverse Anwendungsbereiche genutzt werden können.

Andy Herzberg

Title of the paper

Analysis based on the shallow text features of quality journalism

Keywords

Text mining, classification, text quality, quality journalism

Abstract

For several decades the characteristics that make up well-written texts have been researched extensively. The focus of these studies is mostly the classification of texts suitable for different school grades or second language acquisition. In contrast to previous studies, this paper examines the formal superficial features that influence humanly curated quality journalism. Since few studies focus on the German language, this paper contributes to close this gap in research. Therefore a corpus of award-winning German reportages will be built and, after feature extraction, examined by machine learning algorithms. With traditional feature engineering, a classification accuracy of about 70% can be achieved without parameter tuning. Google's cloud product "AutoML", a self-learning neural network, even achieves a classification accuracy of 82%. The results show that there are formal superficial text features that constitute quality journalism and might be used for miscellaneous applications.

Danksagung

Ich möchte mich an dieser Stelle bei allen bedanken, die mich im Laufe des Studium und während der Erstellung dieser Arbeit unterstützt haben. An erster Stelle danke ich meiner Freundin Rebecca, die mich stets ermuntert hat, das Thema trotz etwaiger Hindernisse weiter zu verfolgen.
Ein besonderer Dank geht an Prof. Dr. Kai von Luck und allen anderen, die mich während des Studiums begleitet und immer wieder aufs Neue motiviert haben. Ich habe viel gelernt und hatte die einmalige Gelegenheit, mich eingehend mit Themen zu beschäftigen, die mir ansonsten weitestgehend verborgen geblieben wären.
Außerdem bedanke ich mich vielmals bei Hannes Schettler, Kristin Göbel und Parsa Mizani für ihren fachlichen Rat und ihre Hilfe beim Lektorat.
Mein abschließender Dank gilt der „Little Kohorte“, Kathrin Baitinger und Quirine Philipsen: Es war mir eine Ehre noch einmal mit euch die Schulbank zu drücken.

Inhaltsverzeichnis

1 Einleitung

Seit etwa 5.000 Jahren werden Texte in Form von Zeichen überliefert (Schreibhaus 2015). Das geschriebene Wort dient dazu, Wissen zu konservieren und zu vermitteln oder einfach, um den Leser zu unterhalten. Verlage kommerzialisieren dieses Kulturgut und haben ein Interesse daran, passende Texte für bestimmte Leserkreise zu produzieren. In Folge dessen bot Robert Gunning Associates bereits 1944 Firmen und Verlagen Beratung im Hinblick auf gute und präzise Formulierung an (nach Pitler/Nenkova 2008: S. 1). Auch die Wissenschaft forscht seit mehreren Dekaden danach welche Eigenschaften einen Text besonders gut lesbar machen. Frühe Forschung auf diesem Gebiet beschäftigte sich weitestgehend mit englischer Sprache und zielte darauf ab, geeignete Lesestücke für bestimmte Schulstufen oder Bildungsniveaus beim Fremdsprachenerwerb zu empfehlen.
Heutzutage lernen Rechner semantische Zusammenhänge ohne menschliche Aufsicht, indem sie statistische Zusammenhänge großer Textmengen erfassen. Die gelernten Zusammenhänge werden verwendet, um Texte zu generieren[1], zusammenzufassen[2] oder zu übersetzen[3]. Maschinell generierte Texte werden unter anderem bei Produktbeschreibungen, Sport-, Finanz- und Wetterberichten eingesetzt. Die Menschen verlassen sich also nicht nur bei der Informationssuche auf Rechner, sondern in jüngster Zeit umso mehr, da sie automatisch generierte Texte konsumieren. Vor diesem Hintergrund ist es verwunderlich, dass es relativ wenige Studien zur computergestützten Erfassung von Textqualität oder Schreibstil gibt (Nenkova 2012: S.49). Deshalb soll sich die vorliegende Masterarbeit eben dieser Thematik widmen.
Inspiriert wurde diese Arbeit durch das Projekt „Future News“[4], welches Gruner + Jahr 2016/2017 im Rahmen der Digital News Initiative (DNI) in Kooperation mit Google durchgeführt hat. Das Ziel des Projekts war es, ein Recherchetool für Journalisten zu entwickeln, welches täglich die wichtigsten Themen des Tages identifiziert und die ursprünglichen und qualitativ besten Quellen[5] ermittelt. Ziel war es nicht, die Suchmaschinen-optimierten Klone der Agenturnachrichten sichtbar zu machen,

[1] z. B. https://www.retresco.de/textgenerierung/ und https://www.ax-semantics.com/
[2] z. B. https://smmry.com/
[3] z. B. https://deepl.com/
[4] https://newsinitiative.withgoogle.com/dnifund/dni-projects/future-news/
[5] z. B. Artikel, Blog-Einträge, Leitartikel und Kommentare

sondern einzigartige und gut geschriebene Inhalte. Diese Arbeit greift den Qualitätsaspekt des Future News-Projekts auf und beleuchtet, welche oberflächlichen Texteigenschaften die Qualität eines Textes ausmachen.

2 Forschungsfrage

Die vorliegende Arbeit beschäftigt sich mit der Frage, was Textqualität von journalistischen Nachrichtentexten ausmacht und ob sich diese Qualität auf Basis von oberflächlichen Textmerkmalen für Reportagen vorhersagen lässt. Eine ähnliche Idee verfolgten Jodie Archer und Matthew L. Jockers – im Fokus ihrer Studien standen allerdings nicht Nachrichtentexte, sondern Romane. Anhand von 3.000 verschiedenen Datenpunkten wurden 2016 die Eigenschaften der Buchtitel der „New York Times"-Bestsellerlisten analysiert. Mit diesen Informationen wurde ein Modell trainiert, welches vorhersagen sollte, ob ein Buch kommerziell erfolgreich wird. Das Ergebnis war erstaunlich: Ein Bestseller konnte mit einer Genauigkeit von 80% vorhergesagt werden (Serrao 2016). Diese Arbeit verfolgt ein ähnliches Ziel und untersucht, ob sich die Qualität journalistischer Texte ebenfalls vorhersagen lässt und welche Textmerkmale besonders aussagekräftig für die Vorhersagen sind.
Die Arbeit gliedert sich zur Beantwortung der Forschungsfrage in zwei logische Einheiten: Im ersten Teil wird das theoretische Fundament gelegt. So wird eingangs erläutert, wie der Lesevorgang funktioniert, da dies hilft, die Vielschichtigkeit von Textqualität zu verstehen. Darauf basierend werden die zahlreichen Facetten von Textqualität erörtert, die in der Fachliteratur nachzulesen sind. Es folgt ein umfassender Überblick zu zahlreichen klassischen und modernen Studien, die sich mit Textqualität und Merkmalen, die diese ausmachen, beschäftigt haben, sowie deren Ergebnissen. Der praktische, experimentelle Teil der Arbeit geht der Frage nach, ob es möglich ist, Textqualität anhand von oberflächlichen Textmerkmalen vorherzusagen. Da keine deutschsprachige Sammlung von qualitativ hochwertigen Nachrichtentexten verfügbar ist, wird diese aus diversen Online-Quellen zusammengestellt. Dieser

Prozess der Korpus[6]-Erstellung wird detailliert beschrieben. Darauf folgt eine Beschreibung des geplanten Versuchsablaufs und der Vorverarbeitungsschritte für die Reportage-Texte. Zur Extraktion der Textmerkmale wird klassisches Feature Engineering betrieben. Basierend darauf wird ein Modell trainiert, in mehreren Iterationen optimiert und zur Bestimmung von qualitativ hochwertigen Reportagen verwendet. Es kommen klassische Machine Learning-Verfahren zum Einsatz, die es erlauben, eine Rangfolge der einflussreichsten Merkmale für die Klassifikation zu erstellen. Auf Basis der Ergebnisse wird abschließend erörtert, ob eine Vorhersage von Textqualität auf Basis des Korpus und der verwendeten Features möglich ist.

3 Literaturrecherche

3.1 Leseverständnis

Seit 40 Jahren beschäftigt sich die Wissenschaft damit zu ergründen, wie das Leseverständnis und die damit verbundenen Prozesse funktionieren. Es wurden zahlreiche Theorien entwickelt, die sich mit der Verschlüsselung, Darstellung und Anwendung von erworbenen linguistischen Erkenntnissen befassen, die für das Verständnis eines Textes relevant sind. Nach Feng stimmen die theoretischen Rahmenwerke darin überein, dass das Ziel des Lesens darin besteht, ein zusammenhängendes Bild eines Textes zu generieren (Feng 2010: S.7ff). Je besser ein Leser die vom Autor beabsichtigten lokalen und globalen Zusammenhänge eines Textes nachvollziehen kann desto verständlicher wird ein Text.
Eine Voraussetzung für das Leseverständnis ist eine Koordination von kognitiven und sprachlichen Fähigkeiten sowie des Gedächtnisses. In der Frühphase des Lesens geht es darum, Wörter zu erkennen und einfache Sätze zu verarbeiten, um daraus eine

[6] Eine Sammlung von niedergeschriebenen Texten in einer bestimmten Sprache wird als Textkorpus bezeichnet. Anhand eines Korpus werden in der Wissenschaft unter anderem Eigenschaften von Texten untersucht (siehe auch: https://de.wikipedia.org/wiki/Textkorpus).

Bedeutung abzuleiten. Die aufeinanderfolgenden Sätze werden hierbei sequenziell verarbeitet und die Informationen aufeinander aufbauend genutzt. Die grundlegende Bedeutungseinheiten werden in bestimmten Mustern im Gedächtnis abgelegt und bilden kleine Bauteile, die mit anderen Informationen, beispielsweise zeitlich oder kausal, verknüpft sind. Aus den Mustern bildet sich nach und nach ein semantisches Netz[7], welches durch weitere gelesene Sätze aufgebaut und erweitert wird. Die Verweise zwischen Satzelementen sind vielerorts nicht explizit sondern implizit vorhanden und müssen aufgelöst werden, wobei zuvor erworbenes Wissen hilft. Für den Gedächtnisaufbau sind verschiedene Speichersysteme des Gehirns notwendig. Eine zentrale Rolle nimmt das Arbeitsgedächtnis[8] ein, welches für die vorübergehende Speicherung und Manipulation von Informationen verantwortlich ist. Während des Lesevorgangs ermöglicht das Arbeitsgedächtnis den Zugriff auf benachbarte Textinformationen und die Abfrage von relevanten Verweisen aus dem Langzeitgedächtnis. Das Arbeitsgedächtnis hat eine endliche Kapazität und ist definiert als die Anzahl der Sätze, die einzelne Personen verarbeiten und sich dabei noch an das letzte Wort des vorangegangenen Satzes erinnern können. Feng formuliert die Erkenntnis mehrerer wissenschaftlicher Studien zu Arbeitsgedächtnis und Sprachverstehen wie folgt: Je mehr Aufwand das Arbeitsgedächtnis erfordert, desto mehr leidet das Leseverständnis (Feng 2010: S.11ff). Daher lässt sich die Textverständlichkeit durch eine Analyse der Anforderung an das Arbeitsgedächtnis gut vorhersagen.

3.2 Textqualität und "Readability[9]"

Nach der US-amerikanischen Sprachwissenschaftlerin Nenkova umfasst das Lesen zwei unterschiedliche Ebenen an semantischer Verarbeitung (Nenkova 2012: S.49). Die eine beschäftigt sich mit dem Textinhalt, die andere mit der Form, dem Schreibstil. Für den Leser sind beide Ebenen miteinander verbunden und die unterschiedlichen Aspekte der Textqualität werden meist nur unbewusst wahrgenommen. Die ebenfalls US-amerikanische Forscherin Louis erweitert diese Textqualitätsdefinition und ergänzt Grammatik und Textfluss als wichtige Einflussfaktoren für Textqualität (Louis 2012: S. 54ff).

---

[7] Ein semantisches Netz ist ein Modell dessen, wie die abstrakte Speicherung von Begriffen und Beziehungen im Gehirn funktioniert. Alle Netze ergeben das semantische Gedächtnis, das gemeinhin Allgemeinwissen genannt wird (siehe auch: http://lexikon.stangl.eu/3128/semantisches-gedachtnis/).

[8] siehe auch https://de.wikipedia.org/wiki/Arbeitsgedächtnis

[9] Unter „Readability" (engl. für Lesbarkeit) versteht man den Grad der Einfachheit, mit dem ein Leser einen geschriebenen Text verstehen kann (siehe auch: https://en.wikipedia.org/wiki/Readability).

Die umfangreichste Zusammenstellung der Aspekte, die Textqualität ausmachen, lässt sich bei Steendam finden (Steendam u. a. 2012: S. 162):

- Inhalt
- Relevanz
- Vollständigkeit/Reichhaltigkeit an Ideen
- Originalität
- Qualität der Argumentation
- Rhetorischer Aufbau
- Schwerpunkt
- Textaufbau
- Thematische Entfaltung und Zusammenhang
- Leserführung und Stimmigkeit
- Wortschatz
- Wortfülle
- Sprachebene
- Präzision
- Verständlichkeit und Angemessenheit der geschriebenen Sprache
- Syntax und Grammatik
- Satzbau/Komplexität
- Syntaktische Richtigkeit
- Grammatische Richtigkeit
- Schreibweise
- Stil
- Klarheit
- Textfluss

In der Fachliteratur für die journalistische Ausbildung finden sich meist konkrete Anweisungen für einen guten Schreibstil. So legt beispielsweise Schneider, ehemaliger Direktor der Henri-Nannen-Schule, den Journalisten nahe (Schneider/Raue 2012: S. 49ff) kurze, verständliche Wörter und miteinander korrespondierende, prägnante Hauptsätze zu verwenden. Da eingeschobene Nebensätze den Lesefluss erschweren, sollten sie vermieden werden. Weiterhin empfiehlt er, den Ausdruck bei Verben, Adjektiven und Präpositionen zu wechseln, dieses aber bei Substantiven zu unterlassen. Salchert gibt in ihrer Schrift „Verständliches Schreiben“ ebenfalls konkrete Ratschläge für eine klare und einfache Sprache (Salchert 2012: S.43):

- Wichtiges nach vorn
- Belangloses und Überflüssiges weglassen
- Gendergerechte Sprache
- Konkret formulieren

- Auf Binsenweisheiten, Floskeln und Klischees verzichten
- Kurze Hauptsätze, wenig Nebensätze, gar keine Schachtelsätze
- Mit Adjektiven geizen
- Mit dynamischen Verben protzen
- Aktiv schlägt Passiv
- Füllwörter weglassen
- Positive Begriffe wählen
- Verneinungen vermeiden
- Fremdwörter und Abkürzungen sparsam einsetzen

Da für die Erstellung des Korpus im praktischen, zweiten Abschnitt dieser Arbeit prämierte Reportagen deutschsprachiger, journalistischer Literaturpreise verwendet werden, sollen an dieser Stelle ausgewählte Beurteilungskriterien aufgeführt werden. Der Fokus wird insbesondere auf folgende Aspekte gelegt: Journalisten sollen „[...] Geschichten [...] auf ungewöhnliche Art [...] erzählen." (Reporter Forum e.V. 2018) und „[...] beim Leser für ‚Kino im Kopf' sorgen." (Krug/Petzold 2016: S. 4). Beiträge sollen außerdem vorbildlich „[...] in Sprache, Stil und Form" (BDZV 2017) sein.
Bei der Recherche zu internationaler Fachliteratur zum Thema Textqualität stößt man unweigerlich auf den Begriff „Readability[10]". Zahlreiche Studien beschäftigen sich mit der Frage, wie die Texteignung für eine Schulstufe oder ein bestimmtes Ausbildungsniveau bei Fremdsprachenerwerb klassifiziert werden kann. Es gibt verschiedene Definitionen des Begriffs[11]: Es können einerseits die für Textqualität bereits genannten Eigenschaften wie Layout, Inhalt und Schreibstil gemeint sein. Andererseits bezeichnet Readability in manchen Kontexten lediglich linguistische Eigenschaften wie Rechtschreibung und Syntax eine Textes (vgl. Larsson 2006: S. 8). Tatsächlich wird beim Lesen der Studien nicht immer deutlich, was der Autor genau unter Readability versteht. Da sowohl bei Readability-Forschung, als auch bei Textqualitätsforschung, die gleichen Merkmalsgruppen Anwendung finden, soll den Unterschieden keine weitere Betrachtung geschenkt werden.
Dieser Abschnitt hat verdeutlicht, dass Textqualität ein breites Spektrum an Texteigenschaften umfasst. Die Mischung der Eigenschaften erzeugt beim Leser – nicht zuletzt durch seine Vorbildung – einen subjektiven Eindruck von Textqualität. Im Rahmen des experimentellen Teils werden lediglich Teilaspekte von Textqualität untersucht. Insbesondere sind das oberflächliche Merkmale[12], die sich mit Standard-Software der Computerlinguistik erfassen und quantifizieren lassen. Inhaltliche

[10] engl. für Lesbarkeit
[11] https://en.wikipedia.org/wiki/Readability#Definition
[12] Hiermit sind unter anderem grammatikalische, syntaktische, lexikalische und morphologische Merkmale gemeint.

Aspekte wie beispielsweise Aktualität des Themas, thematische Entfaltung oder Leserführung können und sollen nicht Gegenstand der Untersuchungen im praktischen Teil sein.

3.3 Traditionelle Forschung

Die Suche nach einer präzisen Definition von Textqualität, also den Faktoren, die guten Textfluss und einfache Lesbarkeit ausmachen, hat eine lange Tradition (Pitler/Nenkova 2008: S. 3). Nachfolgend werden die markantesten Entwicklungen der frühen Forschung auf diesem Gebiet aufgeführt (vgl. DuBay 2006).

Im Jahre 1880 veröffentlichte der englische Professor Sherman die These, dass kurze Sätze und konkrete Ausdrücke beim Leseverständnis helfen. Durch seine Studien konnte er außerdem belegen, dass Umgangssprache einfacher zu verstehen ist, als geschriebener Text und empfahl infolge dessen, dass Schrift- sich an Umgangssprache orientieren solle. Der Autor Rubakin veröffentlichte 1889 eine Wortliste mit 1.500 Einträgen der gebräuchlichsten russischen Wörter, die er aus 10.000 bekannten Texten extrahiert hatte. In den einhergehenden Studien kam er zu dem Ergebnis, dass die meisten Probleme beim Leseverständnis in nicht geläufigen Wörtern und zu langen Sätzen begründet sind. 1921 publizierte Kitson eines der ersten Bücher zum Thema Marketingpsychologie und belegte, dass jeder Leser spezifische Vorlieben für Texte hat und dementsprechend Zeitungen oder Magazine kaufte und las. Durch seine Forschung wurde ebenfalls deutlich, dass kurze Wörter und Sätze zum Leseverständnis beitragen. Der Bildungspsychologe Thorndike griff die Idee der Wortlisten von Rubakin auf und veröffentlichte 1921 das „Teachers Word Book“, welches die passende Literaturauswahl für bestimmte Bildungsgrade auf Basis von Worthäufigkeiten erlaubte. Lively und Pressey veröffentlichten 1923 die erste Formel mit deren Hilfe die Textkomplexität ermittelt werden konnte. Die Formel betrachtet jeweils für 1.000 Wörter den Median der Wortindizes, die in der Thorndike-Liste zu finden waren, und die Wortanzahl der Wörter, die nicht in der Liste waren. Aufgrund der Komplexität der Berechnung wurde in den Folgejahren versucht einfachere und präzisere Formeln zu finden. Bis 1980 wurden mehr als 200 unterschiedliche Formeln veröffentlicht (DuBay 2004: S. 42). Die bekanntesten dieser Lesbarkeitsindizes sind laut Wikipedia[13]: Flesch (1948), Dale und Chall (1948), Gunning Fog (1952), SMOG (1969), Flesch-Kinceid (1975), Automated Readability Index (1975) und Coleman-Liau (1975).

Die Satzlänge wurde in der frühen Forschung bereits als Indikator für Leseverständnis und Textkomplexität identifiziert. Da man anhand der Wortlisten ebenfalls beobachten konnte, dass geläufige Wörter in der Regel kurz sind, wurden in den moderneren

[13] vgl. https://en.wikipedia.org/wiki/Readability

Lesbarkeitsformeln häufig Satz- und Wort- oder Silbenanzahl als Parameter verwendet. So nähert die Flesch-Kinceid-Formel beispielsweise die syntaktische und lexikalische Komplexität anhand der durchschnittlichen Satzlänge und Buchstabenanzahl pro Wort an.
Da deutsche Wörter in der Regel länger als englische sind, passte Amstad 1978 die Wortfaktoren der Flesch-Kinkeid-Formel an (nach ryte.com 2016). Die modifizierte Formel lautet:

FLE: Flesch-Reading-Ease Score
ASL: Average Sentence Length
ASW: Average Number of Syllables per Word

Tabelle 1: Lesbarkeitseinstufung für Alter/Ausbildung (nach ryte.com 2016)

Von … bis unter …	**Lesbarkeit**	**Verständlich für**
0–30	Sehr schwer	Akademiker
30–50	Schwer	
50–60	Mittelschwer	
60–70	Mittel	13–15-jährige Schüler
70–80	Mittelleicht	
80–90	Leicht	
90–100	Sehr leicht	11-jährige Schüler

Anhand dieses konkreten Beispiels wird deutlich, dass Lesbarkeitsformeln einfach zu berechnende Faktoren verwenden, die lediglich eine grobe Annäherung an die Komplexität bieten, die Lesbarkeit ausmacht. Ein Nachteil ist weiterhin, dass die Formeln für spezifische Anwendungsfälle optimiert sind und von Annahmen ausgehen, die nicht gegeben sein müssen. So geht die Flesh-Kincaid-Formel beispielsweise davon aus, dass ein Text aus mindestens 100 Wörtern besteht (Callan/Collins-Thompson 2004: S. 1). Zudem wird der Einfluss bestimmter Faktoren wie zum Beispiel Wort- und Satzlänge überbewertet, während andere Einflussfaktoren, wie u. a. Syntax, Textstruktur, Textzusammenhang, Vorkenntnisse und Lesemotivation (Feng 2010: S. 68) ignoriert werden.
Bezeichnender Weise hat Amazon.com bis 2012 Komplexitätsstufen für Bücher auf Basis von Lesbarkeitsformeln ausgewiesen. Anscheinend wurde diese Funktion von der Seite entfernt, da es andauernde Kritik an der Aussagefähigkeit dieser Formeln gab (vgl. Barbaresi 2011; 2012).

3.4 Aktuelle Forschungsansätze

Die Forschung zur Textkomplexität wird maßgeblich davon getrieben, Texte für ein bestimmtes Bildungsniveau zu empfehlen. Das kann sowohl passende Literatur für die muttersprachliche Schulausbildung als auch Literatur für den Fremdsprachenerwerb sein (Heilman u. a. 2007: S. 1). Während sich die klassische Forschung mangels Rechenleistung auf einzelne wenige Aspekte konzentrieren musste, werden aktuellere Forschungsansätze durch die Entwicklung von Natural Language Processing[14] und Maschinellem Lernen[15] begünstigt. Mit Hilfe dieser neuen Techniken und leistungsfähiger Rechner ist eine effizientere Auswertung großer Textmengen möglich. Allerdings gibt es bis heute kein einheitliches und anerkanntes Modell, das das Zusammenspiel der zahlreichen Lesbarkeitsfaktoren erfasst (Pitler/Nenkova 2008: S. 3).

Typischerweise findet sich in den modernen Studien folgende Unterteilung in Merkmalsgruppen, wenngleich nicht alle Studien immer alle Eigenschaften betrachten (Hancke/Vajjala/Meurers 2012; Nenkova 2012; Pitler/Nenkova 2008; vor der Brück/Helbig/Leveling 2008):

Tabelle 2: Gebräuchlichsten Merkmalsgruppen für die Lesbarkeitsklassifikation

Merkmalsgruppe	**Ziel**
Sprachmodelle	Wortschatz vergleichen
Grammatikalische Merkmale	Analyse der verwendeten Wortarten und grammatikalischer Strukturen
Lexikalische Merkmale	Analyse der lexikalischen Reichhaltigkeit des verwendeten Vokabulars
Syntaktische Merkmale	Analyse von Satzbau und -struktur, strukturelle Beziehungen zwischen Wörtern eines Satzes finden
Morphologische Merkmale	Analyse der internen Struktur von Wörtern und wie diese modifiziert werden können
Textzusammenhang, Textfluss und Textbezüge	Analyse semantischer Zusammenhänge anhand von oberflächlichen Merkmalen

[14] Das Fachgebiet Natural Language Processing untersucht die algorithmische Verarbeitung von natürlicher Sprache (siehe auch: https://en.wikipedia.org/wiki/Natural_language_processing).

[15] Maschinelles Lernen bezeichnet die Fähigkeit von IT-Systemen auf Basis von Mustern eigenständig Lösungen für Probleme zu finden (siehe auch: https://de.wikipedia.org/wiki/Maschinelles_Lernen).

Die verschiedenen Methoden werden vielfach in Kombination verwendet, um ein bestmögliches Klassifikationsergebnis für Textqualität bzw. Textkomplexität zu erzielen. Nachfolgend werden die wichtigsten Studien bezogen auf die jeweilige Merkmalsgruppe aufgeführt.

Sprachmodelle

Bei den Sprachmodellen[16] werden, ähnlich wie beim „Bag-of-Words-Modell“[17], verschiedene Textquellen anhand des vorkommenden Wortschatzes verglichen. Bei den Sprachmodellen wird allerdings die Wahrscheinlichkeit einbezogen, dass ein Wort oder eine Wortsequenz in einem bestimmten Text auftaucht. Basierend auf den Wahrscheinlichkeiten lässt sich ein spezifischer Wortschatz und somit die Eignung für ein bestimmtes Bildungsniveau ermitteln.
Die Forscher Si, Callan und Collins-Thompson haben die Lesbarkeitsklassifikation mit Sprachmodellen zum Gegenstand ihrer Studien gemacht (Callan/Collins-Thompson 2004; Si/Callan 2001). Sie verwendeten ein Unigram-Sprachmodell[18], mit dem die Lesbarkeit einer Webseite eingestuft wird. Sie kamen zu dem Ergebnis, dass sie, gegenüber den klassischen Lesbarkeitsformeln, mit Sprachmodellen eine bessere Korrelation zu den menschlichen Lesbarkeitsbewertungen herstellen konnten. Sie konnten außerdem belegen, dass der Schwierigkeitsgrad des Lesens linear mit der Textlänge ansteigt. Das heißt mit anderen Worten: Je länger der Text, desto schwieriger ist er zu lesen. Weiterhin bewiesen sie, dass die vielfach in Lesbarkeitsformeln verwendete Wortsilbenanzahl keinen Zusammenhang mit der empfundenen Komplexität der untersuchten Webseiten aufwies.

Grammatikalische Merkmale

Schwarm und Ostendorf verwendeten für ihre Studien erstmals eine Kombination aus Sprachmodell und weiteren Features, wie die grammatikalischen POS-Tags[19]. Ihr Ziel

[16] Ein Sprachmodell sammelt statistische Informationen bezogen auf das spezifische Vokabular eines Textes. Dass Modell basiert auf Wahrscheinlichkeiten und dient der Vorhersage von Wortkombinationen (siehe auch: https://de.wikipedia.org/wiki/Spracherkennung#Sprachmodell).
[17] Das Bag-of-Word-Modell ist die vereinfachte Darstellung eines Textes. Berücksichtigt wird lediglich die Wortanzahl, nicht aber Grammatik und Wortstellung
(siehe auch: https://en.wikipedia.org/wiki/Bag-of-words_model).
[18] Beim Unigram-Sprachmodell hängt die Wahrscheinlichkeit jedes Wortes von der Wahrscheinlichkeit dieses Wortes im Dokument ab
(siehe auch https://en.wikipedia.org/wiki/Language_model#Unigram_models).
[19] POS-Tagging benennt einen Prozessschritt beim Natural Language Processing bei dem die Wörter eines Textes mit entsprechenden Wortarten (engl. part of speech) ausgezeichnet werden (siehe auch: https://de.wikipedia.org/wiki/Part-of-speech-Tagging).

war es, mit Hilfe der Wortartinformationen, syntaktische Komplexität abzubilden. Im Vergleich mit den Lesbarkeitsmetriken von Flesch-Kincaid und Lexile konnten sie eine deutlich höhere Übereinstimmung mit menschlicher Bewertung von Textkomplexität erreichen (Schwarm/Ostendorf 2005, S. 15). Heilmann u. a. untersuchten, inwieweit sich die Vorhersage von Textkomplexität, im Gegensatz zum reinen Sprachmodell, durch weitere Merkmale verbessern lässt. Die zusätzlichen Merkmale waren grammatikalische POS-Tags, Tempus von Verben und einfach zu berechnende Textmetriken wie Satz- und Wortanzahl. Das Team um Heilman fand heraus, dass sich der Grad der Korrelation mit menschlicher Bewertung von Textkomplexität durch die zusätzlichen Merkmale leicht verbessern lies, wenngleich das Sprachmodell für sich alleine genommen bessere Ergebnisse lieferte (Heilman u. a. 2007, S. 466).

Lexikalische Merkmale

Lexikalische Merkmale sind solche, die sich auf das Wort als solches beziehen und bilden die Komplexität des verwendeten Wortschatzes ab. Es gilt: Je größer die Vielfalt, desto schwieriger ist das Sprachverständnis (Johansson 2008: S. 71ff). Aus diesem Grunde werden Metriken wie beispielsweise die Type-Token-Relation[20] berechnet, die den Umfang des Wortschatzes erfassen. Es gibt, wie bei den grammatikalischen Merkmalen keine Studien, die die lexikalischen Features isoliert betrachten. Sie werden stets zusammen mit anderen Eigenschaften betrachtet. So kombinierten Vajjala und Meurers traditionelle Metriken, wie beispielweise Satz- und Wortanzahl, mit lexikalischen und syntaktischen Eigenschaften, um passende Webseiten für Kinder oder Erwachsene zu empfehlen. Mit ihrem Feature-Set konnten sie die Klassifikationsgenauigkeit gegenüber vorangegangenen Experimenten ohne lexikalische Merkmale auf Basis des selben Korpus um etwa 17% steigern (Vajjala/Meurers 2012: S. 170).

Syntaktische Merkmale

Neben dem Wortschatz spielt die Komplexität des Satzbaus ebenfalls eine entscheidende Rolle für die Lesbarkeit. Bereits 1998 belegte Gibson, dass syntaktische Komplexität beim Lesen mit verzögerter Verarbeitungszeit einhergeht und somit negativen Einfluss auf die Lesbarkeit hat (nach Pitler/Nenkova 2008: S. 3). Schwarm und Ostendorf verwendeten – wie bereits erwähnt – Sprachmodelle in Kombination mit syntaktischen Features, die sie mit Hilfe von POS-Tags berechneten (Schwarm/Ostendorf 2005). Petersen und Ostendorf knüpften an diese

[20] Die Type-Token-Relation (TTR) ist eine Kennzahl für die Wortvielfalt in einem Textes (siehe auch: https://de.wikipedia.org/wiki/Type-Token-Relation).

Untersuchungen an und konnten belegen, dass Online-Nachrichten für Kinder und Erwachsene erfolgreich mit ihrer Kombination aus Sprachmodell, traditionellen und syntaktischen Features klassifiziert werden konnten (Petersen/Ostendorf 2006: S. 18). Heilman u. a. untersuchten, inwiefern eine Kombination aus Sprachmodell und syntaktischen Features eine Verbesserung gegenüber einem reinen Sprachmodell bewirken würde. Ihr Ziel war es passende, fremdsprachige Lektüre für ein bestimmtes Bildungsniveau zu empfehlen. Hierzu wurden syntaktische Muster aus den Texten extrahiert und für die Klassifikation verwendet. Das Ergebnis belegte, dass eine Kombination beider Modelle die besten Klassifikationsergebnisse lieferte, wenngleich das Sprachmodell einzeln betrachtet für den Anwendungsfall besser funktionierte (Heilman/Collins-Thompson/Eskenazi 2008).

Morphologische Merkmale

Morphologie untersucht Zusammensetzung und Gemeinsamkeiten beim Aufbau von Wörtern. Die deutsche Wissenschaftlerin Hancke untersuchte erstmals den Einfluss morphologischer Merkmale auf die Lesbarkeit während die englische Lesbarkeitsforschung diese weitestgehend ignoriert hatte. Sie konnte belegen, dass sich die Vorhersage der Eignung eines Textes für Kinder oder Erwachsene durch die zusätzliche Verwendung von morphologischen Merkmalen verbessern lies (Hancke/Vajjala/Meurers 2012: S. 1064). Für viele andere Sprachen, die ebenfalls über eine umfangreiche Morphologie verfügen, gibt es mittlerweile entsprechende Untersuchungen (Dell'Orletta/ Montemagni/Venturi 2011; François/Fairon 2012; Aluisio u. a. 2010).

Textzusammenhang, Textfluss, Textbezüge

Linguisten beschäftigen sich seit Dekaden mit Textelemente, die für einen guten Lesefluss sorgen und somit das Textverständnis erleichtern. Als eines der einflussreichsten Modelle nennt die US-amerikanische Sprachwissenschaftlerin Pitler u. a. die Centering-Theorie von Grosz (vgl. Pitler/Nenkova 2008: S. 3). Demnach muss es für das Textverständnis benachbarter Sätze starke Verlinkungen zwischen den relevanten Informationen geben. Diese ermöglichen wiederum ein gutes Leseverständnis. Die meisten rechnergestützten Ansätze, um diese Verlinkungen messbar zu machen, sind durch die Centering-Theorie inspiriert. Karamanis u. a. untersuchte die Centering-Theorie und konnte belegen, dass Satzfolgen, die keine gemeinsamen Entitäten haben, mit zusammenhanglosen und somit schlecht geschriebenem Text assoziiert werden (Karamanis u. a. 2008: S. 42ff). Barzilay & Lapata generalisierten die Centering-Theorie und entwickelten das „Entity Grid Model", welches die Entitäten-Übergänge zwischen Sätzen erfasst und analysiert. Ihr Ziel war es, lokale Zusammenhänge über einen gesamten Text zu ermitteln und Texte basierend auf diesem Index einzustufen. Das Ergebnis war erstaunlich: Gegenüber

dem Feature-Set von Schwarm und Ostendorf mit Sprachmodell und syntaktischen Features konnte die Klassifikationsgenauigkeit mit dem Entity-Grid-Model um etwa 10% gesteigert werden (Barzilay/Lapata 2008: S. 28).
Während sich der größte Teil der bisher genannten Studien mit der Klassifikation von Lesbarkeit, also dem Erfassen von Textkomplexität beschäftigt hat, waren Pitler & Nenkova die ersten, die tatsächlich mit dem Fokus auf Textqualität forschten. Sie nutzen ein umfängliches Set an Merkmalen und ihr überraschendes Ergebnis war, dass sie menschliche Bewertungen von Textqualität zu knapp 89% vorhersagen konnten (Pitler/Nenkova 2008: S. 10). Es wurden, neben Sprachmodell, syntaktischen Merkmalen und dem Entity-Grid-Model, neue Eigenschaften berechnet, die den lokalen Zusammenhang zwischen Sätzen erfassen sollten. So wurde mit Hilfe der annotierten Penn Discourse Treebank (PTDB)[21] die Wahrscheinlichkeit berechnet, dass es vergleichende, kausale oder temporale Beziehungen zwischen den Sätzen gibt. Für sich alleine betrachtet sorgten das Entity-Grid-Model und die Discourse Relations[22] für die besten Klassifikationsergebnisse. Wenngleich die Studie von Pitler & Nenkova zu den einflussreichsten dieser Zeit gehörte, gab es doch Kritik an ihren Untersuchungsmethoden: Die PDTB enthielt lediglich 30 annotierte Texte, die für das Training des Modells für die lokalen Textzusammenhänge verwendet werden konnten (Feng 2010: S. 29ff). Als Resultat der Kritik versuchten Pitler, Louis und Nenkova in einer späteren Studie ihr Modell für implizite Textbezüge durch weitere Features zu verbessern (Pitler/Louis/Nenkova 2009) und untersuchten den Einfluss der Textbezüge auf die wahrgenommene Textqualität (Nenkova u. a. 2010). Eine spätere Studie beschäftigte sich damit, die Qualität von maschinell generierten Texten auf Basis von Grammatik, Textwiederholung, Referenzen, Fokus und Struktur zu untersuchen (Pitler/Louis/Nenkova 2010). Sie machten sich außerdem Textbezüge zunutze, um den Informationsgehalt verschiedener auf sich aufbauender Textpassagen zu untersuchen. Ihr Ergebnis war, dass sich die Verschachtelung detailreicher und allgemein gehaltener Textpassagen gut auf die empfundene Qualität von maschinell generierten Texten auswirkte (Louis/Nenkova 2011b; 2011a; Nenkova 2012). Die Erkenntnisse können wie folgt zusammengefasst werden: Besteht ein Text aus zu vielen allgemeinen Sätzen, wird er als wenig informativ wahrgenommen. Zu viele Detailinformationen können den Leser wiederum verwirren.
Basierend auf den Erkenntnissen der vorangegangenen Studien fordert Louis, dass es je nach Textsorte unterschiedliche Kriterien für Textqualität geben müsse (Louis 2012: S. 55). Der Fokus sollte nicht mehr auf den Leser gelegt werden, sondern vielmehr auf die Eigenschaften der jeweiligen Texte selbst. Louis stellt außerdem die These auf, dass gute wissenschaftliche Arbeiten in bestimmten Textabschnitten Formulierungen

[21] https://www.seas.upenn.edu/~pdtb/
[22] engl. für Diskursbeziehungen bzw. Argumentationsstruktur

mit einer bestimmten Syntax[23] aufweisen. Das Vorhandensein bestimmter syntaktischer Muster könnte ebenfalls als Merkmal für Textqualität verwendet werden, so ihre Annahme.
Louis und Nenkova begannen 2013 damit, einen frei verfügbaren Korpus wissenschaftsjournalistischer Artikel für Textqualitätsforschung zusammenzustellen. Als Grundlage dienten „New York Times"-Artikel, die in dem Sammelband „Best American Science Writing" veröffentlicht wurden und mit der Kategorie „GREAT" ausgezeichnet wurden. Weitere NYT-Artikel des Autors bekamen die Kategorie „VERY GOOD" zugeordnet. Als „TYPICAL" wurden weitere NYT-Artikel aus derselben fachlichen Domäne basierend auf einem Sprachmodell ausgezeichnet. Anhand dieses Science-News-Korpus[24] untersuchten Louis und Nenkova neue Merkmale, die sich auf unterschiedliche Aspekte von gutem Schreibstil konzentrieren, wie Überraschungsmomente, visuelle Sprache und Emotionen. Wie in den vorangegangenen Studien untersuchten sie ebenfalls Textbezüge und Textaufbau. Es zeigte sich, dass sich die preisgekrönten NYT-Artikel mit einer Genauigkeit von 58% klassifizieren ließen (Louis/Nenkova 2013: S. 113).
In den folgenden Jahren wirkte Nenkova vornehmlich an weiteren Ansätzen zu Bestimmung der Qualität von maschinell generierten Texten mit (Hong u. a. 2014). Außerdem forschte sie weiterhin zu Aspekten von Textqualität, wie Informationsdichte und Spezifität von Sätzen (Li/Nenkova 2015; Yang/Nenkova 2014).

Ansätze ohne Feature Engineering

Während bei den älteren Studien massiver Aufwand in das Herausarbeiten von Textmerkmalen investiert wurde, versucht die Forschung in jüngster Zeit Textqualität mit wenig oder ohne Feature Engineering vorherzusagen (Östling/Grigonyte 2017: S. 242). Hierzu werden neuronale Netze verwendet, die mittels überwachtem Lernen trainiert werden. Alikaniotis u. a. verwendeten für ihre Untersuchungen einen Kaggle[25]-Datensatz[26] mit etwa 20.000 benoteten Aufsätzen von englischsprachigen Mittelschülern (Alikaniotis/Yannakoudakis/Rei 2016). Unter Verwendung von Deep

[23] Das können beispielsweise Fragen oder Definitionen sein.
[24] http://www.cis.upenn.edu/~nlp/corpora/scinewscorpus.html
[25] Kaggle ist eine Plattform für Data Scientists auf der regelmäßig öffentliche Wettbewerbe mit Machine-Learning-Fokus ausgeschrieben werden.
[26] https://www.kaggle.com/c/asap-aes

Learning[27] trainierten sie Noten-abhängige Word Embeddings[28] und konnten Klassifikationsergebnisse erzeugen, die besser als alle vorherigen Untersuchungsmethoden mit den menschlichen Bewertungen übereinstimmten (Alikaniotis/ Yannakoudakis/Rei 2016: S. 721).
Östling und Grigonyte forschten ebenfalls auf dem Gebiet der Neuronalen Netze. Sie stellten einen Korpus aus schwedischen Texten unterschiedlicher Qualitätsniveaus zusammen. Verwendet wurden Texte von Nachrichtenseiten, Aufsätzen und Blogs, um ein Convolutional Neural Network[29] zu trainieren (Östling/Grigonyte 2017). Dabei setzten sie den Qualitätsindex der Nachrichtenseiten und universitären Aufsätze gegenüber den Blogtexten höher und trainierten das Modell mit Satzpaaren der unterschiedlichen Quellen. Die Forscher zeigten, dass das Modell erfolgreich lokale Probleme in Sätzen aufzeigen und einen Qualitätsindex pro Text vorhersagen konnte. Weiterhin wurde belegt, dass die Qualitätsvorhersagen gut mit den Benotungen der Arbeiten übereinstimmten und dass der Qualitätsscore auch den Wissens-Zugewinn während des Fremdsprachenerwerbs gut abbildete.

3.5 Deutsche Lesbarkeitsforschung

Zur Vorhersage von Lesbarkeit bzw. Textkomplexität für die deutsche Sprache in deutscher Sprache gibt es Studien von zwei Forscherteams. Einerseits sind es Arbeiten, die sich mit dem „DeLite Readability Checker“ (vor der Brück/Hartrumpf 2007b; 2007a; vor der Brück/Hartrumpf/Helbig 2008; vor der Brück/Helbig/Leveling 2008) der Fernuniversität Hagen beschäftigen. Zum anderen gibt es eine Studie, die am Fachbereich Linguistik der Universität Tübingen entstanden ist (Hancke/Vajjala/

[27] Deep Learning (engl. für tiefes Lernen) ist ein Fachgebiet des Maschinellen Lernens. Diese Technik basiert auf neuronalen Netzen, die die Funktionsweise des menschlichen Gehirns nachahmen. Unter Deep Learning versteht man Architekturen, die über umfangreiche innere Strukturen, sogenannte Hidden Layer, verfügen. Der Nutzen der inneren Struktur besteht darin, dass aus Informationen erzeugt werden können, die eine transformierte Repräsentation der ursprünglichen Informationen darstellen (siehe auch: https://de.wikipedia.org/wiki/Deep_Learning).

[28] Word Embedding (engl. für Worteinbettung) ist der Sammelbegriff für Algorithmen aus dem Natural Language Processing. Bei dieser Technik werden Wörter oder Phrasen einer gegebenen Sprache auf Vektoren abgebildet. Wörter oder Phrasen, die häufig im gleichen Kontexten auftauchen, haben ähnliche Vektoren und repräsentieren daher die semantische Bedeutung eines Wortes im Kontext (siehe auch: https://en.wikipedia.org/wiki/Word_embedding).

[29] Ein Convolutional Neural Network gehört zu Klasse der Deep Learning Architekturen. Diese wurde speziell für die Verarbeitung von Bildern entwickelt. Im Gegensatz zu normalen neuronalen Netzen sind sie in der Lage Matrizen ohne Transformation als Eingabe zu verwenden. Es hat sich herausgestellt, dass Convolutional Neural Networks auch in anderen Bereichen, wie beispielsweise der Textverarbeitung, gut funktionieren.
(siehe auch: https://de.wikipedia.org/wiki/Convolutional_Neural_Network)

Meurers 2012). Die Veröffentlichungen zum DeLite Readability Checker stammen aus den Jahren 2007 und 2008. Danach lassen sich keine Verweise auf DeLite finden. Teilweise scheinen die Ideen und Konzepte in der semantischen Suchmaschine SEMPRIA Search[30] aufgegangen zu sein, die der Co-Autor Hartrumpf als Softwarelösung betreibt. Hancke schreibt über DeLite, dass die Datenbasis aus 500 von Menschen annotierten Texten aus dem kommunalen und juristischen Umfeld[31] besteht (Hancke/Vajjala/Meurers 2012, S. 1065). Hancke kritisiert, dass der Korpus größtenteils juristische Texte enthält, die ein höheres Leseverständnis erfordern als gewöhnliche Texte, was die Aussagekraft der Ergebnisse mindert. DeLite verwendet ein umfangreiches Set an Merkmalen die darauf abzielen, die Lesbarkeit auf Basis von lexikalischen, syntaktischen, semantischen, morphologischen und Diskurs-beziehungen zu bewerten.

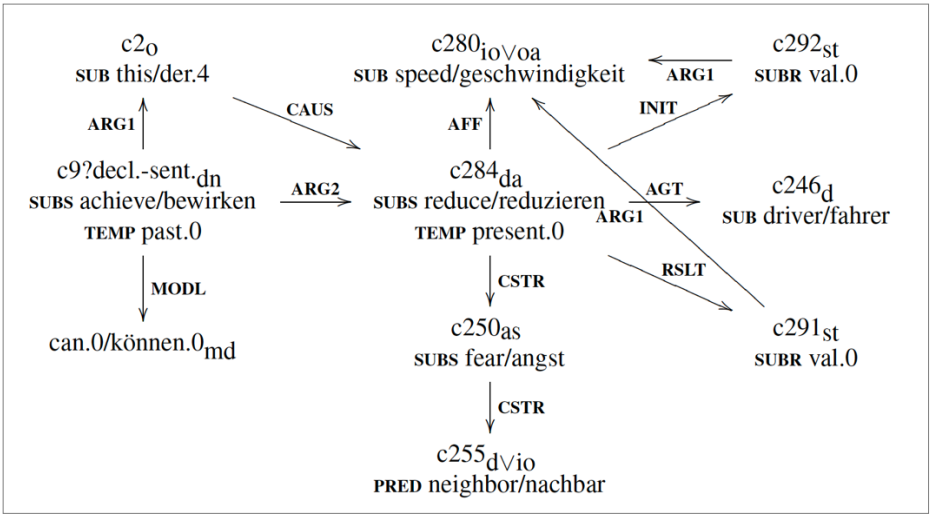


Abbildung 1: Vereinfachte Version eines semantischen Netzwerks (vor der Brück/Hartrumpf/Helbig 2008b: S. 432)

Von der Brück und Hartrumpf bilden neben zahlreichen oberflächlichen Features sogenannte semantische Netzwerke (SN), die die Beziehungen zwischen den Wörtern auflösen und Textkomplexität messbar machen sollen (vor der Brück/Hartrumpf 2007a). Basierend auf ihren Studien haben sich die folgenden Merkmale als am stärksten mit den Lesbarkeitsbewertungen korreliert erwiesen.

[30] https://www.sempria.de/
[31] z. B. Stadtverordnungen

Tabelle 3: Am stärksten mit den Lesbarkeitsbewertungen korrelierte Features (nach vor der Brück/Hartrumpf 2007a: S. 11; vor der Brück/Helbig/Leveling 2008)

Merkmal	**Beschreibung**	**Korrelation mit Lesbarkeits-bewertungen**	**Indicator Type**[32]
Number of words per sentence	Wortanzahl pro Satz	0.430	Sur
SN[33] quality	Qualität des semantischen Netzes	0.399	Syn/Sem
Inverse concept frequency	Anzahl negierter Konzept-Knoten	0.330	Sem
Word form frequency	Häufigkeitsklasse einer Wortform	0.262	Sur
Number of reference candidates for a pronoun	Anzahl Referenzkandidaten für ein Pronomen	0.209	Sem
Number of propositions per sentence	Anzahl Aussagen pro Satz	0.180	Sem
Clause center embedding depth	Verschachtelungstiefe von Teilsätzen	0.157	Syn
Passive without semantic agent	Passive Formulierung ohne sematischen Agenten	0.155	Syn/Sem
Number of SN nodes	Anzahl Knoten eines SN	0.148	Sem
Pronoun without antecedent	Pronomen ohne Nominalbezug	0.140	Sem
Number of causal relations in a chain	Anzahl verketteter kausaler Beziehungen	0.139	Sem
Distance between pronoun and antecedent	Abstand zwischen Pronomen und nominalem Vorläufer	0.138	Sem
Maximum path length in the SN	Maximale Pfadlänge in SN	0.132	Sem
Number of connections between discourse entities	Anzahl an Verbindungen zwischen Diskurseinheiten	0.132	Sem

Das zweite Forscherteam verfolgte mit ihrer Studie das Ziel, Ansätze für Textvereinfachungen zu finden. Im Rahmen ihrer Forschung wird ein Zwei-Klassen-Korpus (einfache/schwierige Texte) auf Basis von GEO- und GEOlino-Onlineartikeln aufgebaut. Anhand des Korpus werden eine Reihe traditioneller (TRAD),

[32] Syn = syntactic / Sem = semantic / Sur = surface

[33] Semantische Netzwerke (SNs) ermöglichen es, die Semantik einzelner Wörter, Phrasen, Sätze, Texte oder Textkorpora einheitlich darzustellen.

syntaktischer (SYN), lexikalischer (LEX) und Sprachmodellfeatures (SM) angelehnt an englische Studien erforscht. Aufgrund der reichhaltigen Morphologie der deutschen Sprache wurden morphologische Features (MORPH) ebenfalls in die Untersuchung mit einbezogen.
Ihre Untersuchungen zu Sprachmodellen ergaben, dass sich durch die Kombinationen von reinen Wörtern und Wörtern durchsetzt mit Wortart-Informationen (POS-Tags) in Form von Uni-, Bi- und Trigrammen[34] eine Klassifikationsgenauigkeit der Lesbarkeit von 77,6% erzielen lies.
Die morphologischen Features erwiesen sich ebenfalls als gute Indikatoren für Lesbarkeit. Durch eine Kombination aller morphologischen Features konnte eine Vorhersagegenauigkeit von 85,4% erreicht werden. Als Top-10-Features erwiesen sich gemessen am Informationszugewinn:

Tabelle 4: Die 10 wichtigsten Features gemessen am Informationszugewinn (nach Hancke u. a., 2012, S. 1074)

Merkmal	**Beschreibung**	**Gruppe**
Avg. Word Length	Ø Wortlänge	LEX/TRAD
Num. 2nd person Vs / Num. finit Vs	Anzahl Verben in der 2. Person pro Anzahl gebeugter Verben	MORPH
Num. Syllables Per Word	Ø Silbenanzahl pro Wort	LEX/TRAD
Num. 3rd person Vs / Num. finite Vs	Anzahl Verben in der 3. Person pro Anzahl gebeugte Verben	MORPH
Avg. length of a T-unit[35]	Ø Länge der kleinsten grammatikalischen Satzeinheit	SYN
Avg. length of a Sentence	Ø Satzlänge	SYN/TRAD
Complex Nominals per Clause	Nomen / Nominalphrasen pro Teilsatz	SYN
Complex Nominals per T-unit	Nomen / Nominalphrasen pro kleinster grammatikalischer Satzeinheit	SYN
Num. PPs per sentence	Anzahl Präpositionaler Phrasen pro Satz	SYN
Avg. length of a clause	Ø Teilsatzlänge	SYN

Die meisten der Top-10-Merkmale gehören zur syntaktischen Feature-GruppeDie hohe Relevanz der syntaktischen Merkmalen bestätigt die Ergebnisse von Vajjala & Meurers für die englische Sprache (Vajjala/Meurers 2012, S. 169). Durch die

[34] N-Gramme nennt man das Ergebnis der Zerlegung eines Textes in kleinere Teilstücke. Das Resultat bilden Wort- oder Buchstabenfolgen (siehe auch: https://de.wikipedia.org/wiki/N-Gramm).
[35] Als T-unit bezeichnet man den kürzesten grammatikalisch zulässigen Satz einschließlich der Nebensätze. Die Analyse von T-units wurde vielfach verwendet, um die syntaktische Komplexität von Texten messbar zu machen (siehe auch: https://en.wikipedia.org/wiki/T-unit)

Kombination aller Merkmale erreicht das Forscherteam um Hancke eine Vorhersagegenauigkeit von 89,7% auf ihren Textkorpus.

3.6 Fazit

Textqualität ist ein Begriff mit vielen Facetten. Neben oberflächlichen Merkmalen sind insbesondere Aufbau, Ideenreichtum, Originalität, Argumentationslinie, Thema, Zusammenhang und Textfluss wichtige Zutaten. Die letztgenannten Punkte lassen sich nicht durch oberflächliche Merkmale erfassen. Einige konkrete Ratschläge aus der journalistischen Schulliteratur lassen sich quantifizierbar machen: So zum Beispiel die Verwendung kurzer, prägnanter Wörter oder den wechselnder Ausdruck bei Verben, Adjektiven und Präpositionen, um nur einige zu nennen.
Computerlinguisten haben eine große Bandbreite an Techniken entwickelt, um Textkomplexität und Textqualität zu erfassen. Die Merkmalsgruppen können in Sprachmodell, grammatikalische, syntaktische, lexikalische und morphologische Features unterschieden werden.
Es erscheint sinnvoll die komplette Bandbreite der oberflächlichen Merkmale für den experimentellen Teil zu extrahieren. Das Sprachmodell könnte verwendet werden, um festzustellen, ob bei qualitativ hochwertigen Texten eine bestimmte, einheitliche Ausdrucksweise vorliegt. Die grammatikalische Wortartinformation können verwendet werden, um die prozentuale Anzahl bestimmter Wortarten zu vergleichen, oder auch um den Wortumfang für die lexikalischen Features zu bestimmen. Es wird in der journalistischen Ausbildung empfohlen, syntaktische Komplexität zu vermeiden. Deswegen erscheint es sinnvoll, auch die syntaktische Merkmale der Texte zu extrahieren und zu vergleichen. Morphologische Merkmale können verwendet werden, um den Aufbau von Wörtern zu vergleichen. Dier deutsche Sprache verfügt über eine reichhaltige Morphologie und auch hier könnten sich interessante Erkenntnisse ergeben. Im ersten Schritt sollen eher einfache Merkmale extrahiert werden. Sofern es der zeitliche Rahmen der Arbeit erlaubt, soll versucht werden, komplexere Merkmale aus den Texten zu extrahieren, wie es das Forscherteam um die US-amerikanische Sprachwissenschaftlerin Nenkova mit hohem Reifegrad praktiziert hat. Die Verwendung von neuronalen Netzen scheidet auf den ersten Blick aus, weil sie große Textmengen zum Training voraussetzen. Abhängig vom verwendeten neuronalen Netz ist es außerdem schwierig, eine Rangfolge einzelner Merkmale mit ihrem Anteil am Klassifikationsergebnis zu erzeugen.
Es ist im theoretischen Teil ebenfalls deutlich geworden, dass für die deutsche Sprache wenig Studien zur maschinellen Untersuchung von oberflächlichen Textqualitätsmerkmalen gibt. Und genau diesem Thema soll sich der experimentelle Teil widmen.

4 Experimente

Angelehnt an die Ergebnisse bisheriger Studien wird im experimentellen Teil untersucht, ob mit Hilfe von Machine Learning eine Vorhersage von Textqualität möglich ist. Untersucht werden hierzu oberflächliche Textmerkmale, wie sie mit Standard-Softwarebibliotheken für Natural Language Processing extrahiert werden können. Der Fokus dieser Arbeit soll, im Gegensatz zu den bisherigen deutschsprachigen Studien zu Lesbarkeitseinstufung auf den journalistischen Merkmalen von Textqualität liegen. Da es nach bisheriger Recherche keine derartigen Studien für die deutsche Sprache gibt, soll die vorliegende Arbeit dazu beitragen, diese Forschungslücke zu schließen. Im ersten Schritt wird ein passender Korpus journalistischer Qualitätstexte erstellt und dann hinsichtlich seiner Merkmale untersucht. Die Merkmale werden mittels Feature Engineering aus den Texten extrahiert, um anschließend bewerten zu können, welche Eigenschaften für die Vorhersage am signifikantesten sind. Wie im vorherigen Kapitel erläutert, liegt der Fokus auf oberflächlichen Merkmalen, da Aspekte wie beispielsweise Aktualität eines Themas oder thematische Entfaltung kaum messbar sind.
Als Blaupause für einen strukturierten Prozess beim Data Mining wird oftmals der KDD-Prozess[36] herangezogen. Der Ablauf umfasst Datenselektion, Vorverarbeitung, Transformation, Data Mining und Interpretation. Die Schritte erfolgen iterativ, bis die gewünschte Güte des Vorhersagemodells erreicht ist und Wissen abgeleitet werden kann. Da dieser Ablauf sich gut für eine strukturierte Darstellung des experimentellen Teils eignet, wird die Herangehensweise anhand dieses Prozess beschrieben.

[36] **K**nowledge **D**iscovery in **D**atabases

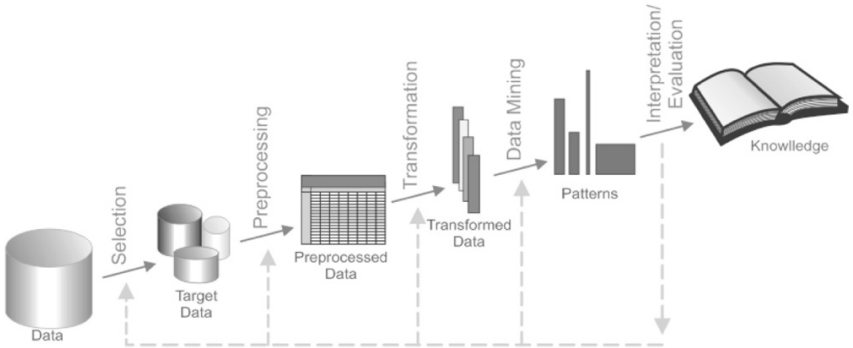


Abbildung 2: Der KDD-Prozess (Góes u. a. 2015, S. 336)

Hinweis: In den folgenden Grafiken, Tabellen und Texten wird der Begriff „Perle“ als Synonym für preisgekrönte bzw. prämierte Nachrichtentexte verwendet, da der Begriff griffiger und kürzer ist. „Nicht-Perlen“ sind Nachrichtentexte, die keine literarischen Preis gewonnen haben und daher eher dem Standard entsprechen.

4.1 Datenselektion

Die Zusammenstellung eines Korpus ist eine klassische Aufgabe beim Natural Language Processing (Hancke/Vajjala/Meurers 2012: 1056). Da für die deutsche Sprache kein klassifizierter Korpus für qualitativ hochwertige Textstücke zur Verfügung steht, muss dieser für das Experiment aufgebaut werden. Als Grundlage dienen Texte, die für Reporterpreis, Theodor-Wolff-Preis, Henri-Nannen-Preis eingereicht und/oder prämiert wurden sowie Stern-Artikeln aus der Gruner + Jahr Pressedatenbank. Der *Reporterpreis* ist einer der angesehensten Journalistenpreise und wurde 2017 zum neunten Mal verliehen. Es werden die besten Stücke u. a. in den folgenden Kategorien gekürt (O A 2017b):

- Reportage
- Wissenschaftsreportage
- Kulturkritik
- Essay
- Lokalreportage
- Interview
- Freie Reporter
- Datenjournalismus
- Investigation

Der *Theodor-Wolff-Preis* wird seit 1962 vom Bundesverband Deutscher Zeitungsverleger verliehen (BDZV 2018). Die Auswahl trifft eine unabhängige, aus neun namhaften Journalisten bestehende Jury für die Kategorien:

- Lokales
- Reportage
- Meinung
- Thema des Jahres

Der *Henri-Nannen-Preis* wird seit 1977 verliehen und zeichnet die hervorragendsten journalistischen Texte des vorherigen Jahres aus. Der Preis wird u. a. in den folgenden Kategorien vergeben (O A 2017a):

- Reportage / Egon Erwin Kisch-Preis
- Investigative Leistung
- Dokumentation
- Web-Projekt

Sämtliche Texte aus genannten Kategorien werden in den jeweils vorliegenden Formaten, entweder PDF, HTML oder CSV, heruntergeladen. Hierfür wird die Software KNIME[37] verwendet, die es ermöglicht, alle Verlinkungen einer Seite zu erfassen und anschließend sämtliche Linkziele herunterzuladen.
Als weitere Textressource wird die G+J Pressedatenbank genutzt, die mehrere Millionen Artikel aus über 150 Quellen beinhaltet (O A 2016c). Von dieser Quelle wird eine zufällige Auswahl von 7.500 Artikeln des Magazins „stern“ exportiert.

4.2 Vorverarbeitung

Die Vorverarbeitung erfolgt in mehreren Schritten: Extraktion der Texte und Wiederherstellen selbiger im Rohformat. Danach Hinzufügen der jeweiligen Textkategorie, für die der Text prämiert wurde. Weiterhin wird jedem Texte ein Klassenlabel (qualitativ hochwertiger Text: ja/nein) hinzugefügt. Auf Basis dieses Labels wird im Nachhinein die Klassifikation durchgeführt. Da sich die Aufbereitung je nach Quelle unterscheidet, werden die notwendigen Schritte gesondert beschrieben.

[37] https://www.knime.com/

Reporterpreis

Die nominierten und preisgekrönten Texte können auf der Webseite des Reporter-Forums[38] entweder in Form von PDF-Readern oder einzelnen PDFs heruntergeladen werden. Die Texte der PDFs wurden mit der Software KNIME unter Verwendung des Tika Parsers extrahiert. Da KNIME sich für die Weiterverarbeitung mittels regulärer Ausdrücke nicht eignete, wurden die Reader-Texte mit Hilfe von Python und regulären Ausdrücke in einzelne Texte zerteilt. Anschließend wurde mit Hilfe von Python und regulären Ausdrücken[39] versucht, die Originalform der Texte zu rekonstruieren. Die folgende Grafik illustriert, welche Herausforderung sich bei der Textbereinigung und -wiederherstellung ergab.

\nwww.reporter-forum.de\n\nDas konnte ja keiner ahnen\n\nKlaus Wowereit hat das Image, der nette Bürgermeister von Berlin zu sein. Er hat aber \nauch eine andere Seite. Damit verblüffte er einst schon seine Lehrer - und damit über-\nrumpelt er jetzt wieder mal seine Gegner.\n\nConstanze Bullion, Süddeutsche Zeitung, 10.10.2011\n\nDer Klaus, haben sie hinten im Rudel gesagt, am Ende wird der Klaus entschei-\n\nden. Also hat der Klaus entschieden, der alte Leitwolf, allein wie immer, damit er vorn \n\nbleibt wie immer. Er kann das, wegbeißen, was ihm quer kommt. Manchmal reißt er \n\nauch ein Schaf, das letzte war grün. Nach der Jagd legt er sich hin und schläft.\n\nJetzt geht es also in die nächste Runde im Roten Rathaus, Klaus Wowereit sieht \n\naus, als hätte er tausend Jahre nicht geschlafen. Er ist es nicht gewohnt, so hart zu \n\narbeiten. Alle Linien seines Gesichts weisen in diesen Tagen abwärts, die Augen \n\nwollen weg, sich zurückziehen in die Höhlen. Nein, verjüngt hat er ihn nicht, dieser \n\nkräftezehrende Lauf in seine letzte Amtszeit im Land Berlin.\n\n"Keine Basis für eine Koalition", hat der Berliner Bürgermeister gesagt, als er \n\nvor ein paar Tagen dem grünen Schaf die Kehle durchgebissen hat. Da hatte er schon \n\nWitterung nach rechts aufgenommen, zu den schwarzen Schafen, die er diese Woche \n\nbeschnuppern wird. Wowereit will eine Koalition mit der CDU, hat er erklärt, bevor er \n\nsich zurückzog in seinen Bau. Eilig, mit langem Schritt, den Rücken rund vom vielen \n\nRunterbeugen, hinter ihm sein Rudel. Es sind Aktenträger, gescheitelte Köpfe, auch \n\ndie vom linken Flügel nicken folgsam. Der Alte hat es ihnen mal wieder gezeigt.\n\nKlaus Wowereit ist jetzt Deutschlands dienstältester Ministerpräsident nach Kurt \n\nBeck. Berlin hat ihn eben zum dritten Mal gewählt, auch weil viele glauben, er sei ein \n\nnetter, leutseliger Kuschelteddy. Wowereit pflegt dieses Bild, denn es schützt den \n\nhttp://www.reporter-forum.de/\n\nwww.reporter-forum.de\n\nanderen in ihm, den Wolf, der ab und zu ein bisschen Blut braucht, der Führung nicht \n\nteilt, immer auf der Hut ist, nur der eigenen Nase vertraut. Sie hat ihn weit geführt.\n\nAus dem grauen Haus in

Abbildung 3: Ausschnitt eines rohen Textextrakts eines Reporterpreis-PDF

Konkret ging es darum, die durch das Speicherformat hervorgerufenen Satz- und Worttrennung zu eliminieren und intakte Satzstrukturen zu bilden, die anschließend

[38] http://www.reporter-forum.de/

[39] Reguläre Ausdrücke werden in der Programmierung zur Mustererkennung eingesetzt. Sie bieten eine flexible und präzise Methode, um Textfolgen zu entdecken und mit Hilfe anderer Programmiersprachen zu modifizieren (https://de.wikipedia.org/wiki/Regulärer_Ausdruck).

weiterverarbeitet werden können. Weiter Infos zu den Texten wurden aus mehreren Quellen aggregiert. Die Kategorien der prämierten Texte konnten von Wikipedia[40] extrahiert werden, während die der nominierten Texte manuell von der Reporter-Forum-Webseite zusammengestellt wurden. Da selbst eine Jury-Nominierung für den Reporterpreis als Auszeichnung angesehen werden kann, wurden alle Texte mit einem positiven Label hinsichtlich der Textqualität versehen.

Theodor-Wolff-Preis

Die nominierten und preisgekrönten Texte sind seit 2009 auf der Webseite des BDZW[41] im Volltext verfügbar. Für die Korpus-Erstellung wurden die Texte ebenfalls mit KNIME heruntergeladen und dann mit XPath aus dem HTML extrahiert. Python wurde verwendet, um die Texte mit denselben regulären Ausdrücken, die bei den PDFs verwendet wurden, in ihrer ursprünglichen Textform wiederherzustellen. Die Textkategorien der Preisträger wurden von Wikipedia[42] übernommen und mit der jeweiligen Texten verknüpft. Für die nominierten, aber nicht preisgekrönten Texte konnte keine Textgattung ermittelt werden. Daher werden diese nicht für den Korpus verwendet. Die verwendeten, preisgekrönten Texte erhalten demzufolge alle ein positives Label hinsichtlich der Textqualität.

Henri-Nannen-Preis

Für den Korpus wurden die Nannen-Preis-Shortlist des Jahres 2011 und sämtliche Einreichungen des Jahres 2017 in der Kategorie Reportagen verwendet. Die Textstücke des Henri-Nannen-Preises lagen für das Jahr 2011 in Form eines PDF-Readers[43] vor, der, wie für die anderen Quellen bereits beschrieben, in die ursprüngliche Form umgewandelt wurde. Die Texte der Nannen-Preis-Shortlist bekamen alle ein positives Label hinsichtlich ihrer Textqualität.
Die Einreichungen des Jahres 2017 standen als CSV-Export zur Verfügung, aus denen die Texte einfach ausgelesen und ebenfalls mit denselben regulären Ausdrücken behandelt wurden, wie auch die PDFs und HTML-Texte. Somit soll verhindert werden, dass sich Textcharakteristika, die sich durch Extraktionsfehler ergeben, im Klassifikationsergebnis wiederspiegeln. Die Artikel des Jahres 2017 wurden anhand der im Internet veröffentlichten Shortlist[44] von 2017 abgeglichen. Alle Artikel der

[40] https://de.wikipedia.org/wiki/Deutscher_Reporterpreis
[41] http://www.bdzv.de/twp/
[42] https://de.wikipedia.org/wiki/Liste_der_Preisträger_des_Theodor-Wolff-Preises
[43] http://www.reporter-forum.de/fileadmin/pdf/Egon-Erwin-Kisch-Preis/Kisch_Preis_2011/HNP2011Reader.pdf
[44] https://www.nannen-preis.de/download/juroren_shortlists_und_gewinner_2017.pdf

Shortlist bekommen ein positives Label, die übrigen ein neutrales Label hinsichtlich ihrer Textqualität.

Gruner + Jahr Pressedatenbank

Die vierte Quelle für den Korpus stellen Artikel aus der Gruner + Jahr Pressedatenbank[45] dar. Für den Korpus wurden zufällig etwa 7.500 stern-Magazintexte aus der Pressedatenbank exportiert. Die Hoffnung war, dass sich unter den Exporten genügend Artikel passend zu den journalistischen Preisträgertexten finden würden.
Die stern-Artikel lagen im XML-Format vor und die Texte wurden mittels XPath aus ihrer XML-Struktur extrahiert und wie auch alle anderen Texte zuvor mit denselben regulären Ausdrücken behandelt. Die Kategorie-Codes der Artikel lagen nur chiffriert vor, konnten allerdings aus den URL-Parametern rekonstruiert werden. Sämtliche Texte bekamen ein neutrales Label hinsichtlich ihrer journalistischen Qualität.

Es kann nicht ausgeschlossen werden, dass sich, trotz sorgfältiger Analyse, einzelne Fehler in den Texten verbergen. Bei einer manuellen Inspektion der Artikel wurden keine offensichtlichen Fehler mehr festgestellt. Auf diesen Umstand soll im weiteren Verlauf der Experimente ein Augenmerk gelegt werden falls Auffälligkeiten auftreten.

4.3 Datenselektion für den Korpus

Erst nach der Vorverarbeitung der unterschiedlichen Texte und dem Vereinheitlichen der Textkategorien wird deutlich, wie viele Texte tatsächlich für eine Analyse zur Verfügung stehen. Es ergibt sich ein Bild, wie in nachfolgender Tabelle gezeigt.

[45] https://pressedatenbank.guj.de/PDB/Home.htm

Tabelle 5: Metriken des vorläufigen Korpus

	Perlen[46]			**Nicht-Perlen**		
Textform	**n_docs**[47]	**n_sents**[48]	**n_words**[49]	**n_docs**	**n_sents**	**n_words**
Bericht[50]	-	-	-	2.599	509.656	5.873.981
Freie Reporter	58	22.973	259.857	-	-	-
Lokaljournalismus	39	6.617	77.543	-	-	-
Reportage	254	117.684	1.427.609	312	73.643	870.721
Reportage Kultur	28	10.954	139.624	-	-	-
Reportage Lokal	71	21.240	232.193	-	-	-
Reportage Wissenschaft	13	4.987	59.250	-	-	-
TOTAL	**463**	**184.455**	**2.196.076**	**2.911**	**583.299**	**6.744.702**

Da lediglich in den Kategorien Reportage und Bericht eine ausreichende Anzahl von Texten zu finden sind, erfolgt eine Recherche zu den Gemeinsamkeiten und Unterschieden beider Textformen. Die wichtigsten Unterschiede werden im Folgenden kurz skizziert (nach Redaktion alpha Lernen 2017a; 2017b): Während der Bericht streng objektiv sein soll, darf die Reportage auch subjektive Beobachtungen und Sichtweisen einbringen. Das Ziel der Reportage ist es, ein Geschehen so emotional darzustellen, wie es der Autor selbst erlebt hat. Typisch ist für die Reportage ein fortwährender Perspektivwechsel vom Subjektiven ins Objektive. Anders hingegen ist der Bericht, der neutral informiert und keine Meinung äußert. Lebendig wird der Bericht dadurch, dass konkrete Einzelheiten und Interviews, sowie direkte und indirekte Rede im Wechsel eingestreut werden. Ein weiterer Unterschied ist der Tempus in dem ein Ereignis geschildert wird: Eine Reportage beschreibt das Geschehen meist in der Gegenwartsform, während der Bericht eher in der Vergangenheitsform berichtet (O A 2016a).
Angelehnt an die Empfehlung aus der Fachliteratur (Louis 2012, S. 55) sollen die Textfeatures genrespezifisch betrachtet werden. Nach reiflicher Überlegung werden

[46] In den folgenden Grafiken, Tabellen und Texten wird der Begriff „Perle“ als Synonym für preisgekrönte bzw. prämierte Texte verwendet, da der Begriff griffiger und kürzer ist.
[47] Dokumentanzahl
[48] Satzanzahl
[49] Wortanzahl
[50] Artikel des Magazins „stern“ aus der Pressedatenbank

daher lediglich Reportagetexte verwendet, da von dieser Textsorte zumindest eine kleine Anzahl an Perlen und Nicht-Perlen vorliegen.

Oberflächliche Analyse des Korpus

Da lediglich Reportagen für die weiteren Untersuchungen verwendet werden, ergibt sich folgendes Bild über den Korpus:

Tabelle 6: Korpus-Metriken für Perlen und Nicht-Perlen des Texttyp Reportage

	n_docs	n_sents	n_words	n_sents / n_docs	n_words / n_docs	n_words / n_sents
Reportage (Perle)	254	117.684	1.427.609	463	5.621	12
Reportage (Nicht-Perle)	312	73.643	870.721	236	2.791	12

Zu beobachten ist, dass die Wortanzahl pro Satz bei allen Textsorten etwa gleich ist, während die Satzanzahl bei den Reportagen mit neutralem Label, gefolgt von den Berichten, deutlich geringer ist. Die hochwertigen Reportagen sind im Schnitt länger wie auch aus den folgenden Histogrammen ersichtlich wird.

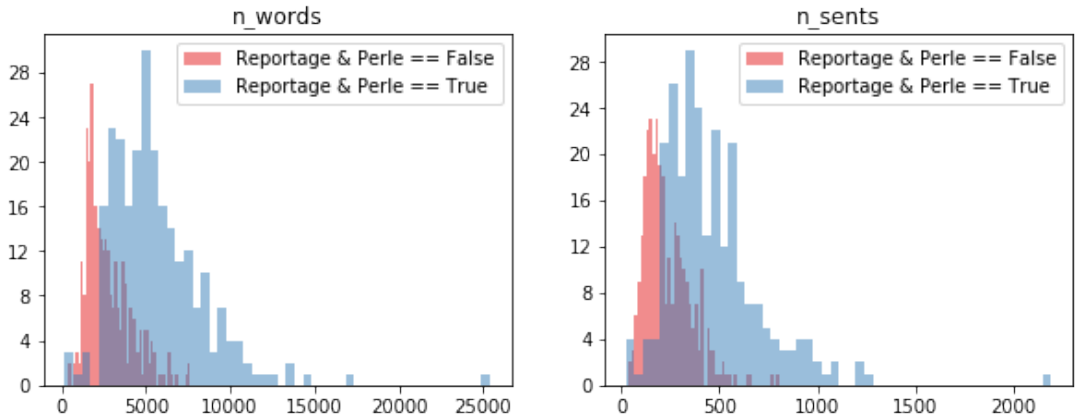


Abbildung 4: Wortanzahl und Satzlänge im Vergleich

Augenscheinlich können daher lediglich normalisierte, relative Features verwendet werden, da ansonsten lediglich die Textlänge als ausschlaggebender Klassifikations-parameter gewichtet wird.

4.4 Transformation

Nachdem der Korpus erstellt ist, beginnt die Transformation der Daten mit Hilfe von verschiedenen Software-Bibliotheken. Der geplante Prozess wird im folgenden Schaubild dargestellt. Nicht alle Schritte wurden letztendlich implementiert. Es folgt eine detaillierte Beschreibung der tatsächlich durchgeführten Verarbeitungsschritte.

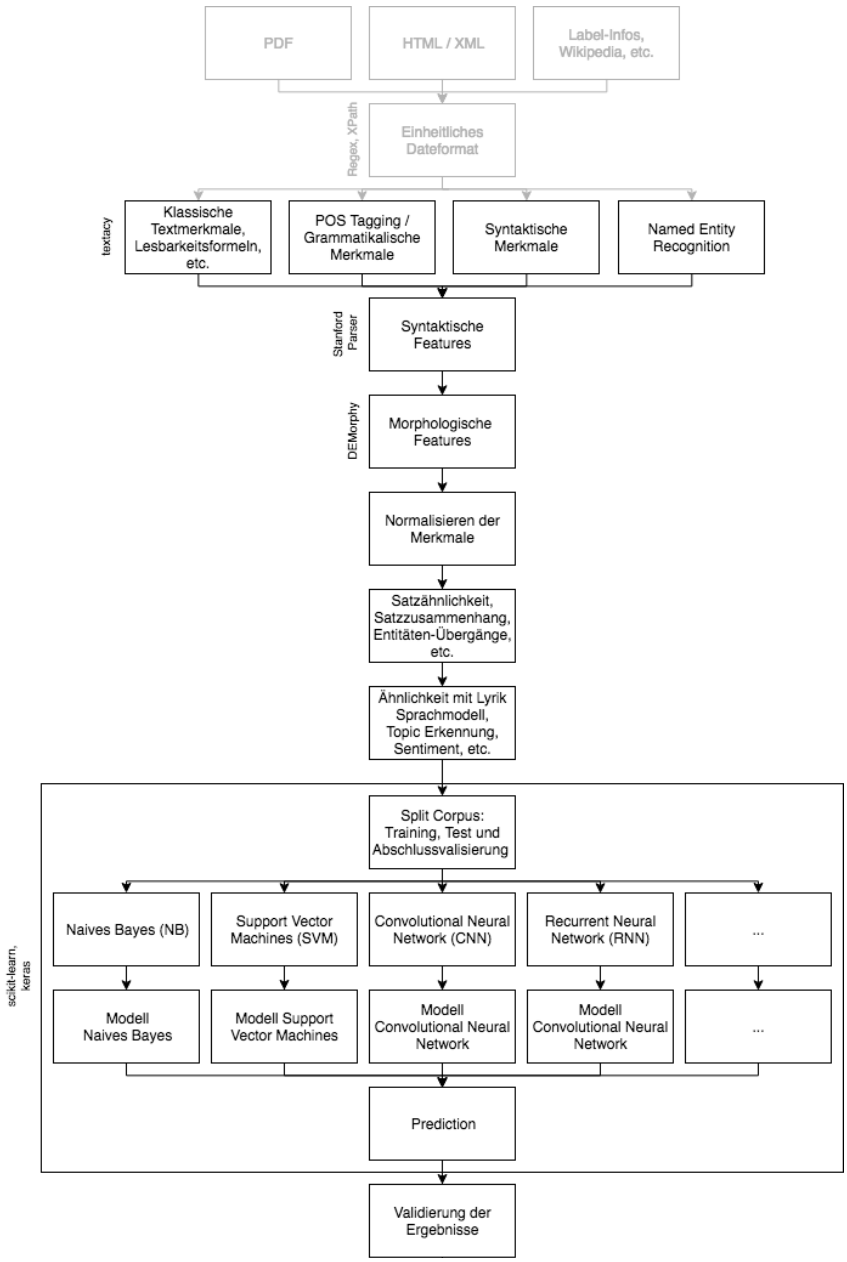


Abbildung 5: Geplanter Versuchsaufbau

Klassische Features

Als erstes werden die Texte mit dem Python-Library textacy[51] prozessiert. Textacy basiert auf der verbreiteten Python-Bibliothek spaCy[52]. Während spaCy unter anderem für Tokenisation, Part-of-Speech Tagging, Dependency Parsing und Named Entity Recognition sorgt, setzt textacy auf dieser Basis auf und stellt zusätzliche Funktionen bereit. Das sind beispielsweise einfache Textmetriken, Lesbarkeitsmetriken oder Distanzfunktionen, mit denen Ähnlichkeiten zwischen Sätzen oder Dokumenten ermittelt werden können. Nachdem die Texte von textacy verarbeitet wurden, stehen automatisch die folgenden Metriken pro Dokument zur Verfügung:

Abkürzung	Beschreibung
n_chars	Anzahl Buchstaben
n_long_words	Anzahl Wörter > 7 Buchstaben
n_monosyllable_words	Anzahl einsilbiger Wörter
n_polysyllable_words	Anzahl mehrsilbiger Wörter
n_sents	Satzanzahl
n_syllables	Anzahl Silben
n_unique_words	Anzahl einzigartiger Wörter
n_words	Anzahl Wörter
wiener_sachtextformel[53]	Kennzahl für die Eignung eines Sachtextes für eine Schulstufe

Abbildung 6: Automatische Metriken über das textacy-Modul TextStats

Anhand dieser Textfeatures werden die folgenden Metriken berechnet:

Abkürzung	Beschreibung
avg_words_length	Ø Wortlänge
avg_sent_length	Ø Satzlänge
avg_syllables_p_word	Ø Silbenanzahl pro Wort
avg_long_words	Ø Anzahl Wörter > 7 Buchstaben
avg_monosyllable_words	Ø Anzahl einsilbiger Wörter
avg_polysyllable_words	Ø Anzahl mehrsilbiger Wörter
flesch_readability_ease_de[54]	Deutsche Flesch Reading Ease Kennzahl (vgl. Kapitel 3.3)

Abbildung 7: Berechnete Metriken auf Basis der textacy-TextStats

[51] https://textacy.readthedocs.io/en/stable/#
[52] https://spacy.io/
[53] siehe auch: https://de.wikipedia.org/wiki/Lesbarkeitsindex#Wiener_Sachtextformel
[54] siehe auch: https://de.wikipedia.org/wiki/Lesbarkeitsindex#Flesch-Reading-Ease

Grammatikalische Features

Beim Preprocessing mit textacy werden die Dokumente mit grammatikalischen STTS-POS-Tags, Universal-POS-Tags und syntaktischen Dependency-Tags ausgezeichnet. Das sogenannte POS-Tagging bezeichnet die Auszeichnung der Tokens eines Dokuments mit den entsprechenden Wortarten. Eine vollständige Liste mit Beschreibungen der Tags befindet sich im Anhang ab Seite 78.
Das Stuttgart-Tübingen-Tagset (STTS) ist mit 54 verschiedenen Tags deutlich umfangreicher als das Universal-POS-Tagset mit 17 Merkmalen. Das liegt daran, dass das STTS-Tagset weitere grammatikalischer Merkmale wie beispielsweise Kasus, Tempus oder Person mit einbezieht und somit einen höheren Detailgrad liefert.
Für die STTS- und Universal-POS-Tags werden die absoluten Tag-Häufigkeiten mit der Python-Bibliothek collections gezählt und dann auf die Wortanzahl des jeweiligen Dokuments normalisiert. Eine komplette Auflistung der verfügbaren Universal- und STTS-POS-Tags inklusive Beispiele findet sich im Anhang ab Seite 78.

Lexikalische Features

Lexikalische Features[55] beschreiben den Umfang des Wortschatzes. Im einfachsten Fall ist das die Anzahl unterschiedlicher Wörter bezogen auf die gesamte Wortanzahl. Die Berechnung der lexikalischen Eigenschaften pro Wort erfolgt mit Hilfe der Universal-POS-Tags. In der Lesbarkeitsforschung gilt: Je größer die Wortvielfalt, desto schwieriger ist das Sprachverständnis (Johansson 2008, S. 62ff). Daher hat sich die Gruppe der lexikalischen Features vielfach als guter Indikator für die Lesbarkeitseinstufung erwiesen (Dell'Orletta/Montemagni/Venturi 2011; François/ Fairon 2012; Hancke/Vajjala/Meurers 2012; vor der Brück/Helbig/Leveling 2008). Es kann davon ausgegangen werden, dass die lexikalischen Merkmale bei der die Vorhersage von Textqualität ebenfalls eine Rolle spielen. Wie im Kapitel 3.2 zur Textqualität zitiert, wird den Journalisten in der Ausbildung nahe gelegt, Füllwörter zu vermeiden. Außerdem soll der Ausdruck bei Substantiven wenig gewechselt werden. Beide Empfehlungen führen tendenziell zu einer niedrigeren Type-Token-Relation. Bei Verben, Adjektiven und Präpositionen soll die Ausdrucksweise gewechselt werden, was in der Tendenz eher zu einem höheren TTR-Wert führt. Es gibt unterschiedliche Ausprägungen der TTR-Kennzahl, die die Wortfülle eines ganzen Textes oder unterschiedlicher Wortarten messbar macht.

[55] Lexikalische Wörter geben einem Text Inhalt und Bedeutung. Zu ihnen zählen Nomen, Adjektive, Verben und Adverbien.

Type Token Relation
Das Type Token Relation beschreibt die lexikalische Vielfalt eines Textes und wird mit folgender Formel berechnet:

$$\text{ttr} = \frac{\text{n\_unique\_words}}{\text{n\_words}}$$

Lexical density
Lexikalische Wörter geben einem Text Inhalt und Bedeutung (O A 2016d). Zu ihnen zählen Nomen, Adjektive, Verben und Adverbien.

$$lexical\ density = \frac{n\_words_{lex}}{n\_words}$$

Die nachfolgenden Kennzahlen beschreiben das Verhältnis von Nomen, Verben, Adverbien und Adjektiven zur Gesamtheit der lexikalischen Wörter, während die letzte Messgröße das Verhältnis von Nomen zu Pronomen darstellt.

Noun variation$_{lex}$

$$noun\ variation_{lex} = \frac{n\_nouns}{n\_words_{lex}}$$

Adjective variation$_{lex}$

$$adjective\ variation_{lex} = \frac{n\_adjectives}{n\_words_{lex}}$$

Verb variation$_{lex}$

$$verb\ variation_{lex} = \frac{n\_verbs}{n\_words_{lex}}$$

Adverb variation$_{lex}$

$$adverb\ variation_{lex} = \frac{n\_adverbs}{n\_words_{lex}}$$

Noun/Pron Ratio

$$noun\ pronoun\ ratio = \frac{n\_noun}{n\_pronoun}$$

Syntaktische Features

Beim syntaktischen Parsen[56] wird der Text in grammatikalische Strukturen umgewandelt und entsprechend ihrer syntaktischen Funktion getagged[57]. Beim Preprocessing mit textacy werden die Dokumente mit den spaCy-Tags hinsichtlich ihrer syntaktischen Abhängigkeiten ausgezeichnet. Diese „Syntactic Dependency Parsing"-Tags sind mit 38 Tag-Labeln recht umfangreich. Im zweiten Schritt werden die Texte mit dem Stanford Parser[58] verarbeitet. Das deutschsprachige Modell des Parsers basiert auf dem Negra-Korpus und verwendet das Negra-Tagset. Da sich beide Tagsets voneinander unterscheiden, sollen sie gleichermaßen für die Klassifikation verwendet werden. Bei sämtliche syntaktischen Features werden die Tags auf die Anzahl der Sätze normalisiert, da sich die Tags jeweils auf einen Satz beziehen. Eine vollständige Liste sämtlicher syntaktischer Tags findet sich im Anhang im Kapitel 9.1.

Morphologische Features

Die Morphologie erforscht die Struktur von Wörtern und Gesetzmäßigkeiten bei deren Aufbau. Für das Tagging der morphologischen Features wurden drei verschiedene Tools getestet. Es handelt sich um DEMorphy[59], RFTagger[60] und RDRPOSTagger[61]. Die Python-Bibliothek DEMorphy gibt für jedes morphologisch veränderte Wort sämtliche Möglichkeiten hinsichtlich der Deklination unsortiert aus, was sich für die weitere Verarbeitung als impraktikabel herausstellte. Das Java-Programm RFTagger gibt morphologisch erweiterte POS-Tags aus, erwartet aber als Input ein satzweise formatiertes Textfile, was ebenfalls ungünstig erschien. Verwendet wurde letztendlich die Java-Version des RDRPOSTaggers, die mittels eines Python-Wrappers angesprochen wurde. Im Gegensatz zu den anderen Tools erlaubt es der Tagger ein komplettes Textfile als Ganzes einzulesen und zu prozessieren. Für die Verarbeitung wurde mittels KNIME je Dokument eine separate Textdatei erzeugt. Diese wurden dann mittels Python eingelesen und morphologisch prozessiert. Nachfolgende Prozessschritte waren Extraktion der Tags, Umwandeln in Universal POS-Tags und Zählen der Muster mittels regulärer Ausdrücke.
Angelehnt an die von Hancke u. a. untersuchten Merkmale (Hancke/Vajjala/Meurers 2012) wurden die folgenden Features berechnet:

---

[56] Als Parser wird ein Computerprogramm bezeichnet, welches für die Zerlegung und Umwandlung einer Eingabe in ein best. Ausgabeformat sorgt (siehe auch: https://de.wikipedia.org/wiki/Parser).
[57] Im Natural Language Processing versteht man unter Tagging die Annotation von Wörtern mit Zusatzinformationen der jeweiligen Domäne (siehe auch: https://de.wikipedia.org/wiki/Tagging).
[58] https://nlp.stanford.edu/software/lex-parser.shtml
[59] https://github.com/DuyguA/DEMorphy
[60] http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/
[61] https://github.com/jacopofar/RDRPOSTagger-python-3

Relativer Anteil infiniter Verben

$$avg\_infinitive\_verbs = \frac{n\_infinite\_verbs}{n\_verbs}$$

Relativer Anteil finiter Verben

$$avg\_finite\_verbs = \frac{n\_finite\_verbs}{n\_verbs}$$

Relativer Anteil partiziper Verben

$$avg\_participle\_verbs = \frac{n\_participle\_verbs}{n\_verbs}$$

Relativer Anteil imperativer Verben

$$avg\_imperative\_verbs = \frac{n\_imperative\_verbs}{n\_verbs}$$

Relativer Anteil modaler Verben

$$avg\_modal\_verbs = \frac{n\_modal\_verbs}{n\_verbs}$$

Relativer Anteil von Hilfsverben

$$avg\_auxiliary\_verbs = \frac{n\_auxiliary\_verbs}{n\_verbs}$$

Relativer Anteil von Verben im Präsens

$$avg\_present\_tense\_verbs = \frac{n\_present\_tense\_verbs}{n\_finite\_verbs}$$

Relativer Anteil von Verben der Vergangenheitsform

$$avg\_past\_tense\_verbs = \frac{n\_past\_tense\_verbs}{n\_finite\_verbs}$$

Relativer Anteil von Verben in der 1. Person Singular

$$avg\_1st\_person\_sg\_verbs = \frac{n\_1st\_person\_sg\_verbs}{n\_finite\_verbs}$$

Relativer Anteil von Verben in der 2. Person Singular

$$avg\_2nd\_person\_sg\_verbs = \frac{n\_2nd\_person\_sg\_verbs}{n\_finite\_verbs}$$

Relativer Anteil von Verben in der 3. Person Singular

$$avg\_3rd\_person\_sg\_verbs = \frac{n\_3rd\_person\_sg\_verbs}{n\_finite\_verbs}$$

Relativer Anteil von Verben in der 1. Person Plural

$$avg\_1st\_person\_pl\_verbs = \frac{n\_1st\_person\_pl\_verbs}{n\_finite\_verbs}$$

Relativer Anteil von Verben in der 2. Person Singular

$$avg\_2nd\_person\_pl\_verbs = \frac{n\_2nd\_person\_pl\_verbs}{n\_finite\_verbs}$$

Relativer Anteil von Verben in der 3. Person Singular

$$avg\_3rd\_person\_pl\_verbs = \frac{n\_3rd\_person\_pl\_verbs}{n\_finite\_verbs}$$

Relativer Anteil von Verben pro Satz

$$avg\_verbs\_per\_sent = \frac{n\_verbs}{n\_sents}$$

Relativer Anteil von Nomen im Akkusativ

$$avg\_accusative\_nouns = \frac{n\_accusative\_nouns}{n\_nouns}$$

Relativer Anteil von Nomen im Dativ

$$avg\_dative\_nouns = \frac{n\_dative\_nouns}{n\_nouns}$$

Relativer Anteil von Nomen im Genitiv

$$avg\_genitive\_nouns = \frac{n\_genitive\_nouns}{n\_nouns}$$

Relativer Anteil von Nomen im Nominativ

$$avg\_nominative\_nouns = \frac{n\_nominative\_nouns}{n\_nouns}$$

4.5 Erste Beobachtungen

Nachfolgend werden die Mittelwerte und Histogramme der berechneten Features verglichen, bevor die Texte mit ihren berechneten und normalisierten Merkmalen klassifiziert werden. Diese Untersuchung dient dazu, den Datenbestand nach der Transformation zu sichten und die Ergebnisse zu verifizieren.

Auf Basis des Eigenschaftsvergleichs zwischen prämierten Perlen und Nicht-Perlen lässt sich erstmals erahnen, dass es Unterschiede gibt, die ausschlaggebend für die Klassifikation der Texte sein könnten. Es erfolgt pro Merkmalsgruppe eine auszugsweise Darstellung der markantesten Beobachtungen. Die Histogramme sämtlicher Features befinden sich im Anhang ab Seite 86.

Klassische Features

Beim Vergleich der Mittelwerte der klassischen Features zwischen Perlen und Nicht-Perlen fällt auf, dass der durchschnittliche Satz bei den preisgekrönten Reportagen länger, während das durchschnittliche Wort dort kürzer ist.

Tabelle 7: Mittelwertvergleich und Signifikanztest der klassischen Textfeatures

Merkmal	Beschreibung	Mean (±SD) Nicht-Perle	Mean (±SD) Perle	t-Wert	p-Wert
n_chars / n_words	Ø Wortlänge	5,38 (±0,25)	5,28 (±0,24)	-5,01	0,00000
n_long_words / n_words	Ø Anzahl langer Wörter	0,26 (±0,03)	0,25 (±0,03)	-4,98	0,00000
n_syllables / n_words	Ø Silben pro Wort	1,74 (±0,09)	1,71 (±0,08)	-4,55	0,00001
n_polysyllable_words / n_words	Ø Mehrsilben-Wörter	0,18 (±0,03)	0,17 (±0,03)	-4,52	0,00001
n_monosyllable_words / n_words	Ø Einsilben-Wörter	0,54 (±0,03)	0,55 (±0,03)	4,49	0,00001
wiener_sachtextformel[62]	Kennzahl	6,30 (±1,30)	5,90 (±1,19)	-3,81	0,00016
flesch_readability_ease[63]	Kennzahl	65,89 (±6,29)	67,54 (±5,64)	3,24	0,00126
n_words / n_sents	Ø Satzlänge	12,21 (±2,42)	12,47 (±2,36)	1,30	0,19410

Lange Wörter gelten in der Lesbarkeitsforschung als schwieriger lesbar und könnten somit ein erster Anhaltspunkt für die unterschiedliche Textqualität sein. Die übrigen Werte in der Tabelle ergeben sich weitestgehend aus Satz- und Wortlänge. Eine Untersuchung der statistischen Signifikanz der Unterschiede zwischen Perlen und Nicht-Perlen zeigt allerdings, dass die Abweichungen bei der Satzlänge statistisch

[62] Wiener Sachtextformel: 5 ≙ Lese-Bildungsgrad der 5. Klasse, 6 ≙ Lese-Bildungsgrad der 6. Klasse
[63] Flesh-Readability-Score Schwierigkeitsgrad: 60-70 ≙ „Mittel“, 70-80 ≙ „Mittelleicht“

nicht signifikant sind. Die folgenden Histogramme illustrieren den Zusammenhang zwischen Textlänge, Silbenanzahl pro Wort und zwei deutschen Lesbarkeitsformeln in der unteren Reihe.

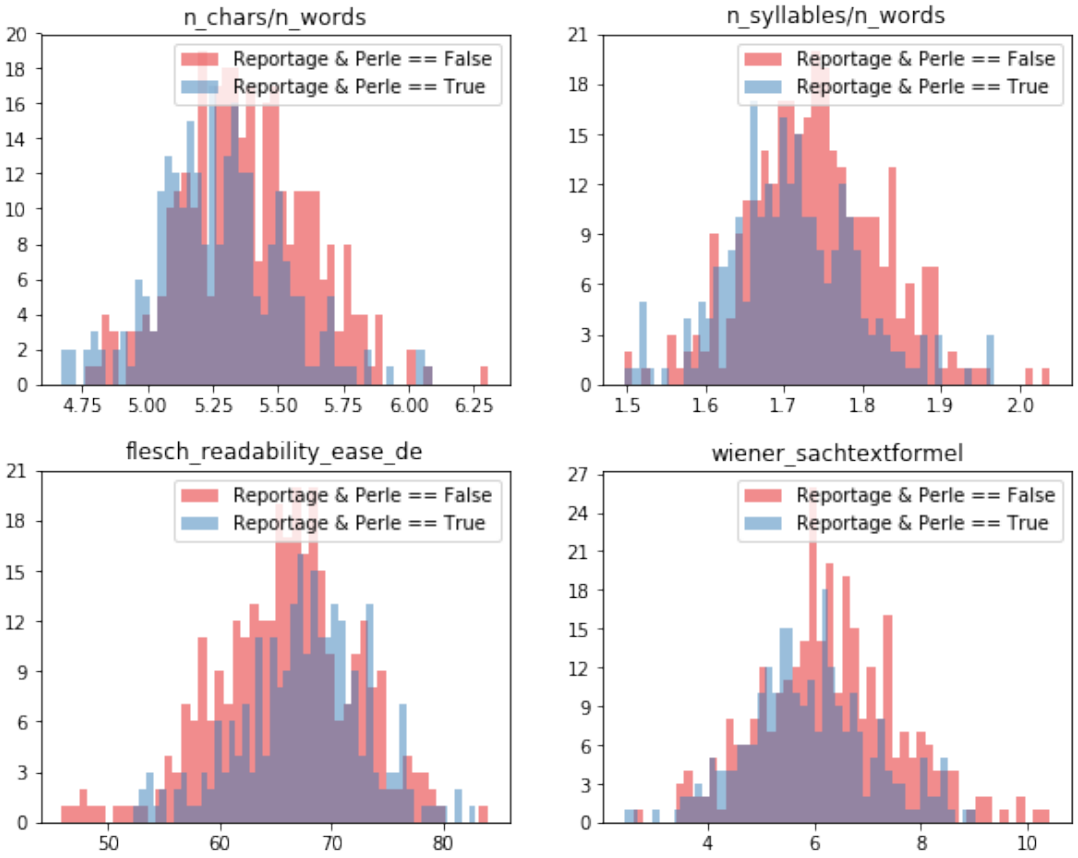


Abbildung 8: Histogramme ausgewählter klassischer Features

Laut Flesch-Readability-Score ist die Schwierigkeitsstufe des Reportage-Korpus Mittel bis Mittelleicht. Obwohl die Übergänge fließend sind, sind die Nicht-Perlen im Schnitt etwas schwieriger lesbar. Bei der Wiener Sachtextformel ist die Achse gespiegelt: „Ein Wert von 4 steht [...] für [einen] sehr leichten Text, dagegen bezeichnet 15 einen sehr schwierigen Text.“ (O A 2018b).

Grammatikalische Features

In den Histogrammen der POS-Tags (vgl. Anhang, S.88ff) und beim Vergleich der Mittelwerte lassen sich mehrere Auffälligkeiten feststellen. Die folgende Tabelle listet die POS-Tags auf, bei denen der Unterschied in den Mittelwerten besonders signifikant ist. Am auffälligsten ist das Komma: Da die Satzlänge bei den prämierten

Reportagen länger ist, verwundert es nicht, dass unter den Perlen im Mittel auch mehr Kommas verwendet werden.

Tabelle 8: Mittelwertvergleich und Signifikanztest der STTS POS-Tags

Merkmal	Beschreibung	Beispiel	Mean (±SD) Nicht-Perle	Mean (±SD) Perle	t-Wert	p-Wert
n_COMMA / n_words	Komma	,	0,083 (±0,017)	0,095 (±0,020)	7,85	0,00000
PROAV / n_words	Pronominales Adverb	deswegen [sprechen wir] darüber	0,006 (±0,002)	0,005 (±0,002)	-4,68	0,00000
TRUNC / n_words	Abkürzung	UKW, km/h	0,001 (±0,001)	0,000 (±0,001)	-4,41	0,00001
n_PPER / n_words	irreflexives Personal-pronomen	ich, er, ihm, mich, dir	0,058 (±0,021)	0,065 (±0,019)	4,40	0,00001
n_KON / n_words	Neben-ordnende Konjunktion	und, oder, aber	0,030 (±0,007)	0,028 (±0,006)	-4,10	0,00005
n_ADJA / n_words	attributives Adjektiv	[das] große [Haus]	0,042 (±0,011)	0,039 (±0,009)	-4,09	0,00005
n_VVFIN / n_words	finites Verb, voll	[du] gehst, [wir] kommen [an]	0,073 (±0,011)	0,077 (±0,011)	3,85	0,00013
n_PPOSAT / n_words	Attribuieren-des Possessiv-pronomen	mein [Buch], deine [Mutter]	0,015 (±0,006)	0,017 (±0,005)	3,74	0,00021

Bei den preisgekrönten Reportagen werden außerdem mehr Pronomen verwendet. Pronomen deuten auf eine gute Vernetzung der einzelnen Textpassagen hin (Pitler/Louis/ Nenkova 2010), die für das Erfassen der Textzusammenhänge benötigt werden. Adjektive und Adverbien werden bei den preisgekrönten Reportagen weniger verwendet, was gut zu den Empfehlungen „Mit Adjektiven geizen“ und „Füllwörter weglassen“ von Salchert passt (vgl. Kapitel 3.2).
Die genannten Beobachtungen sind folglich auch bei den Universal POS-Tags zu beobachten. Basierend auf den Universal Tags kann man weiterhin feststellen, dass bei den Perlen mehr Verben verwendet werden. Das wiederum könnte gut zu Salcherts Empfehlung „Mit dynamischen Verben protzen“ passen, sofern es sich denn um dynamische Verben handelt.

Tabelle 9: Mittelwertvergleich und Signifikanztest der Universal POS-Tags

Feature	Beschreibung	Beispiel	Mean (±SD) Nicht-Perle	Mean (±SD) Perle	t-Wert	p-Wert
n_SPACE / n_words	Leerzeichen		0,017 (±0,008)	0,021 (±0,008)	6,38	0,00000
n_ADJ / n_words	Adjektiv	[das] große [Haus]	0,071 (±0,012)	0,066 (±0,011)	-5,23	0,00000
n_PUNCT / n_words	Punkt	.	0,173 (±0,020)	0,182 (±0,020)	5,16	0,00000
n_VERB / n_words	Verb	laufen; springen; tanzen	0,126 (±0,013)	0,130 (±0,012)	4,27	0,00002
n_CONJ / n_words	Bindewort	und so; weder – noch; aber; denn; als; wie	0,035 (±0,007)	0,032 (±0,006)	-4,12	0,00004
n_PRON / n_words	Pronomen	ich; er; ihm; mich; dir	0,100 (±0,024)	0,107 (±0,023)	3,86	0,00013
n_ADV / n_words	Adverb	schon, bald, doch	0,079 (±0,016)	0,074 (±0,015)	-3,63	0,00031
n_ADP / n_words	Adposition	aufgrund [des Unwetters]; um [Himmels] willen	0,099 (±0,012)	0,096 (±0,011)	-3,12	0,00188

Lexikalische Features

Bei der visuellen Inspektion der Histogramme ist bei einigen lexikalischen Features der preisgekrönten Texte eine bimodale Verteilung zu erkennen. Nachfolgend werden hier beispielhaft zwei dieser Histogramme dargestellt. Die Histogramme sämtlicher Features befinden sich im Anhang auf S. 89ff.

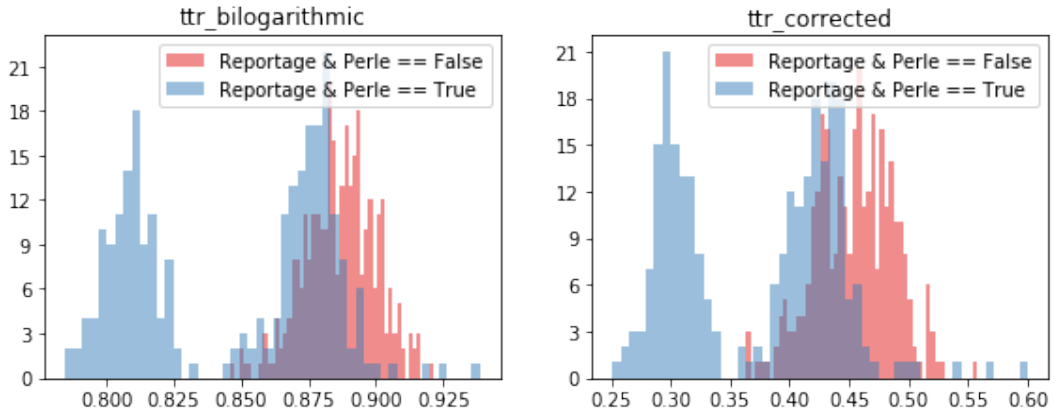


Abbildung 9: Bimodale Histogramme zweier TTR-Kennzahlen

Wie aus den Histogrammen ersichtlich wird, scheint sich zumindest ein Teil der Reportagen anhand der Type-Token-Relation perfekt klassifizieren zu lassen. In der Fachliteratur findet sich folgende Erklärung für dieses Phänomen (Perkuhn/Keibel/Kupietz 2012), dass sich auf die höhere Wortanzahl bei den preisgekrönten Reportagen zurückführen lässt:

> „Ein größerer TTR-Wert deutet auf einen differenzierteren und reichhaltigeren Wortschatz hin, ein niedrigerer TTR-Wert hingegen auf ein größeres Maß an Wiederholung und auf eine formelhaftere Sprachverwendung.
> Leider aber muss gleich eine Einschränkung nachgeschoben werden: Wissenschaftler haben früh bemerkt, dass das Type-Token-Verhältnis von der Korpusgröße abhängt: Wenn sonst alles gleichbleibt, nimmt der TTR-Wert mit steigender Korpusgröße ab [...]. Als Maß, um die lexikalische Vielfalt zweier Korpora zu vergleichen, eignen sich TTR-Werte damit nur bei gleich großen Korpora.“

Eine Untersuchung des Korpus bestätigt den Zusammenhang wie nachfolgende Tabelle verdeutlicht: Obwohl die Reportagen mit einem TTR-Wert unter 0,25 deutlich mehr Wörter beinhalten, gibt es unwesentlich mehr einzigartige Tokens.

Tabelle 10: Token-Vergleich auf Basis des TTR-Werts

	Merkmal	**Mean**
ttr > 0,25	n_unique_words	1.148
	n_words	2.924
ttr < 0,25	n_unique_words	1.329
	n_words	7.138

Das bimodale Phänomen bei den Histogrammen ist auch bei den berechneten Werten Corrected Type Token Ratio, Bilogarithmic Type Token Ratio und den TTR-Kennzahlen von Nomen, Verben, Adverbien und Adjektiven zu beobachten. Da sich der Zusammenhang dieser Features mit der Textlänge negativ auf die Klassifikation auswirken könnte, werden diese Merkmale bei der Klassifikation gesondert oder gar nicht betrachtet.
Die von McCarthy 2010 eingeführte Kennzahl MTLD[64] (McCarthy/Jarvis 2010) behebt diesen Makel beim TTR-Wert und stellt die lexikalische Vielfalt unabhängig von der Wortanzahl im Korpus dar. Implementiert wurde das MTLD-Feature mit Hilfe der Python-Bibliothek von John Frens[65]. Da die Python-Bibliothek auch Wortlisten mit mehr als 50 Tokens als Eingabe akzeptiert, wurde die lexikalische Vielfalt ebenfalls für Adjektive, Adverbien, Nomen und Verben sowie deren Lemmata berechnet.

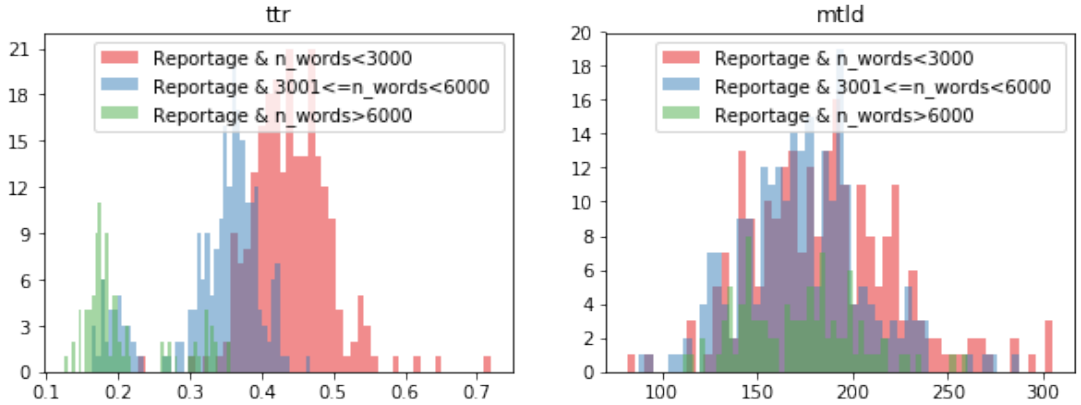


Abbildung 10: Histogramme von TTR im Vergleich zu MTLD

Anhand von Mittelwert-Vergleich und Signifikanztest wird deutlich, dass die preisgekrönten Stücke insgesamt über einen etwas weniger reichhaltigen Wortschatz verfügen. Weiterhin zeichnet die prämierten Reportagen aus, dass sie eine größere Bandbreite unterschiedlicher Verben verwenden, während die Varianz von Adjektiven und Adverbien bei den Nicht-Perlen höher ist.

[64] MTLD („Measure of textual diversity“) ist eine Kennzahl für die Wortfülle eines Textes (siehe auch: https://en.wikipedia.org/wiki/Lexical_diversity).
[65] https://github.com/jfrens/lexical_diversity

Tabelle 11: Mittelwertvergleich und Signifikanztest der lexikalischen Features[66]

Merkmal	Beschreibung	Mean (±SD) Nicht-Perle	Mean (±SD) Perle	t-Wert	p-Wert
verb_variation _lex_words	Kennzahl für Variation der Verben	0,26 (±0,03)	0,27 (±0,03)	5,74	0,00000
mtld	Kennzahl für Wortumfang	185,69 (±35,65)	169,34 (±32,00)	-5,68	0,00000
lexical_density	Verhältnis lexikalischer Wörter zur Wortanzahl	0,48 (±0,02)	0,48 (±0,02)	-5,55	0,00000
adjective_variation _lex_words	Kennzahl für Variation der Adjektive	0,15 (±0,02)	0,14 (±0,02)	-4,40	0,00001
noun_pronoun_ratio	Verhältnis von Nomen zu Pronomen	2,27 (±0,78)	2,04 (±0,62)	-3,84	0,00014
mtld_adverb_variation	Kennzahl für Variation der Adverbien	34,23 (±9,24)	31,48 (±7,66)	-3,79	0,00017
adverb_variation _lex_words	Kennzahl für Variation der Adverbien	0,16 (±0,03)	0,16 (±0,03)	-2,58	0,01022
mtld_verb_variation	Kennzahl für Variation der Verben	319,85 (±119,6)	342,43 (±124,4)	2,19	0,02862

Syntaktische Features

Die syntaktischen Tags wurden durch das „Syntatic Dependency Parsing“ von textacy (spaCy) hinzugefügt. Anschließend wurden die Reportagen mit dem Stanford Parser prozessiert. Bei beiden handelt es sich um syntaktische Parser, die aber weitestgehend unterschiedliche Tags auswerfen. Daher werden die Ergebnisse beider Parser verwendet. Ausschlaggebend für diese Entscheidung war unter anderem der Umstand, dass Pitler und Nenkova in ihren Studien belegen konnten, dass, zumindest in der englischen Sprache, die Anzahl an Verbalphrasen signifikant mit der empfundenen Textqualität zusammenhängt (Pitler/Nenkova 2008: S. 9). Der Standford Parser zeichnet die Verbalphrasen[67], im Gegensatz zum spaCy-Parser, direkt aus und markiert eher längere syntaktische Einheiten. Das zeigt sich auch an der Anzahl der

[66] Für die Betrachtung wurden die auffälligen bimodalen und längenabhängigen Features ignoriert.
[67] Mit Verbalphrase wird eine syntaktische Einheit bezeichnet, in dessen Zentrum ein Verb steht (siehe auch: https://de.wikipedia.org/wiki/Verbalphrase).

verfügbaren Tags: 43 bei spaCy und 25 beim Stanford Parser. Alle syntaktischen Tags werden anhand der Anzahl der Sätze normalisiert, um eine Satzlängen-unabhängige Kenngröße zu erlangen.
In den Histogrammen des spaCy-Taggers lassen sich visuell keine markanten Unterschiede zwischen den positiven und neutralen Labeln feststellen (vgl. Anhang, S. 91ff). Die markantesten Abweichungen des Mittelwert-Vergleichs und Signifikanztests sind in der folgenden Tabelle zusammengefasst.

Tabelle 12: Mittelwertvergleich und Signifikanztest der spaCy Syntax-Features

Merkmal	Beschreibung	Beispiel	Mean (±SD) Nicht-Perle	Mean (±SD) Perle	t-Wert	p-Wert
n_APP / n_sents	Apposition / Beisatz	[der Mann,] ein Mittvierziger,	0,056 (±0,029)	0,069 (±0,033)	4,966	0,00000
n_OA / n_sents	Akkusativ-objekt	[ich sehe] ihn	0,674 (±0,137)	0,735 (±0,157)	4,921	0,00000
n_SB / n_sents	Subjekt	[Der] Mann [kennt uns]	1,419 (±0,230)	1,512 (±0,278)	4,338	0,00002
n_JU / n_sents	Satzverknüpfer	Und [Peter ging weg.]	0,064 (±0,032)	0,053 (±0,028)	-3,963	0,00008
n_PNC / n_sents	Komponente vom Eigennamen	1. FC St. Pauli	0,122 (±0,090)	0,150 (±0,103)	3,488	0,00052
n_PH / n_sents	Platzhalter: pronominales Adverb oder „es“	[Ich glaube] daran [, dass er kommt]; [er liebt] es [, ein Auto zu fahren]	0,003 (±0,004)	0,005 (±0,005)	3,228	0,00132
n_OC / n_sents	Satz oder Verbalphrase: einen Verb, Adjektiv oder Nomen untergeordnet	[er verspricht,] uns zu besuchen; [der Beschluss,] ein Haus zu bauen	0,584 (±0,174)	0,630 (±0,181)	3,075	0,00220
n_VO / n_sents	Vokative	Hans [, wo gehst du jetzt hin?]	0,000 (±0,001)	0,001 (±0,002)	3,005	0,00278

Es ist unter anderem auffällig, dass es bei den prämierten Reportagen mehr Beisätze und Akkusativobjekte gibt. Zudem gibt es mehrere Subjekte pro Satz, was vermutlich dadurch begründet ist, dass bei den Perlen häufiger Hauptsätze in einem Satz verkettet werden.

Auch anhand der Tags des Standford-Parsers lässt sich feststellen, dass die Perlen häufiger verkettete Hauptsätze pro Satz verwenden.

Tabelle 13: Mittelwertvergleich und Signifikanztest der Standford Parser Tags

Merkmal	Beschreibung	Beispiel	Mean (±SD) Nicht-Perle	Mean (±SD) Perle	t-Wert	p-Wert
n_CS / n_sents	Verkettete Hauptsätze	Peter kommt und Paul geht	0,216 (±0,071)	0,245 (±0,080)	4,475	0,00000
n_S / n_sents	Satz	Hans liebt Maria.	1,665 (±0,260)	1,769 (±0,315)	4,318	0,00000
n_MPN / n_sents	Eigenname aus mehreren Worten	1. FC St. Pauli	0,098 (±0,067)	0,112 (±0,068)	2,515	0,01200
n_AP / n_sents	Adjektivphrase [68]	[Peter ist] an sich ganz nett; rund [100 Euro]	0,223 (±0,079)	0,208 (±0,073)	-2,310	0,02100
n_CAC / n_sents	Beigeordnete Adposition[69]	[die Züge] von und nach [Hamburg]	0,001 (±0,002)	0,000 (±0,001)	-2,094	0,03700
n_VP / n_sents	Verbale Phrase	[Peter will] uns besuchen	1,709 (±0,350)	1,768 (±0,343)	2,016	0,04400
n_QL / n_sents	Nichtwort, Kürzel	DXP13	0,000 (±0,000)	0,000 (±0,001)	1,596	0,11100
n_CNP / n_sents	Beigeordnete Nominalphrase [70]	ein Mann und eine Frau; die Jungen und Mädchen	0,147 (±0,061)	0,140 (±0,055)	-1,502	0,13400

Für die englische Sprache erwiesen sich die Verbalphrasen als hoch korreliert mit den menschlichen Bewertungen von Textqualität. Da sie bei den prämierten Reportagen im Mittel häufiger verwendet werden, sind Verbalphrasen anscheinend auch für

[68] Als Adjektivphrasen werden syntaktische Einheiten bezeichnet, in dessen Zentrum ein Adjektiv steht (siehe auch: https://de.wikipedia.org/wiki/Adjektivphrase).
[69] Unter dem Begriff Adposition werden eine Reihe morphologisch unveränderbare Wörter wie bspw. Präpositionen zusammengefasst (siehe auch: https://de.wikipedia.org/wiki/Adposition).
[70] Mit Nominalphrase wird eine syntaktische Einheit bezeichnet, in dessen Zentrum ein Nomen steht (siehe auch: https://de.wikipedia.org/wiki/Nominalphrase).

deutsche Reportagen ein Qualitätsindikator. Weiterhin kann beobachtet werden, dass die Nicht-Perlen mehr Adpositionen und Adjektivphrasen aufweisen, also mehr Füllwörter, die es für eine gute Schreibe zu vermeiden gilt.

Morphologische Features

Bei der Untersuchung mittels Mittelwert-Vergleich und Signifikanztest zeigen sich folgende Unterschiede:

Tabelle 14: Median-Vergleich der markantesten morphologischen Features

Merkmal	Beispiel	Median Nicht-Perle	Median Perle	t-Wert	p-Wert
n_3rd_person_pl_verbs / n_finite_verbs	[sie] putzen [die Fenster]	0,228 (±0,066)	0,203 (±0,055)	-4,683	0,00000
n_present_tense_verbs / n_finite_verbs	[Du] fährst [Fahrrad]	0,684 (±0,119)	0,644 (±0,125)	-3,812	0,00015
n_1st_person_pl_verbs / n_finite_verbs	[wir] freuen [uns sehr]	0,016 (±0,020)	0,012 (±0,014)	-2,750	0,00615
n_1st_person_sg_verbs / n_finite_verbs	[ich] spiele [Tischtennis]	0,031 (±0,037)	0,024 (±0,025)	-2,707	0,00699
n_infinitive_verbs / n_verbs	[Sie gehen, wir] kommen	0,197 (±0,043)	0,189 (±0,036)	-2,497	0,01279
n_nominative_nouns / n_nouns	[es ist ein schöner] Tag	0,429 (±0,056)	0,440 (±0,057)	2,367	0,01829
n_3rd_person_sg_verbs / n_finite_verbs	[Cindy] kommt [in den Garten]	0,556 (±0,081)	0,570 (±0,071)	2,179	0,02974
n_past_tense_verbs / n_finite_verbs	[ich] ging [nach Hause]	0,149 (±0,103)	0,167 (±0,111)	2,003	0,04567

Die Unterschiede ergeben sich weitestgehend in der Verwendung von finiten Verben[71]. Die Perlen verwenden mehr Verben in der Vergangenheitsform, während die Nicht-Perlen häufiger die Gegenwartsform verwenden. Außerdem verwenden die Perlen selten die erste und dritte Person Plural, dafür häufiger die dritte Person Singular.

[71] Bei einem finiten Verb handelt es sich um die konjugierte Form dieser Wortart. Das Beugen eines Verbs basiert auf Person, Numerus, Genus, Modus und Tempus (siehe auch: https://www.deutschplus.net/pages/Finite_infinite_Verbformen).

4.6 Data Mining

Nachfolgend werden die für die Klassifikation notwendigen Vorbereitungs- und Durchführungsschritte beschrieben.

Erstellen der Trainings- und Testdaten

Der Korpus umfasste 566 Textdokumente mit den Merkmalen, die anhand der Tagger- bzw. Parser-Ausgaben berechnet wurden.
Vor der Erstellung der Training- und Testdaten werden die absoluten Anzahlen der Features entfernt und lediglich auf den jeweiligen Text bezogene, relative Merkmale verwendet. Es bleiben 186 Features, von denen zehn lexikalische Features mit der Länge korreliert sind und eine bimodale Verteilung im Histogramm aufweisen.
Im ersten Schritt soll eine Klassifikation mit allen Merkmalen durchgeführt werden.
Das Datenset wird hierzu zufällig im Verhältnis 50/50 in Trainings- und Testdaten aufgeteilt. Die Trainingsdaten dienen dazu, das Klassifikationsmodell zu berechnen, und mit den Testdaten wird anschließend die Güte des Modells überprüft.

Trainieren der Modelle

Für die Klassifikation sollen explorativ mehrere Klassifikatoren mit unterschiedlichen Algorithmen trainiert werden. Die Implementierung der Verfahren Gaussian Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree, Random Forest und Gradient Boosting Trees erfolgt mit der Python-Bibliothek scikit-learn[72]. Aus Zeitgründen und da der zusammengestellte Korpus relativ klein ist, soll auf eine Klassifikation mittels neuronaler Netze vorerst verzichtet werden.
Jeder Algorithmus hat spezifischen Vor- und Nachteile, weswegen die verschiedenen Verfahren parallel mit ihren Standardeinstellungen ausprobiert werden sollen. Eine Auflistung der Konfigurationen, das Klassifikationsergebnis sowie die Standardabweichung befindet sich im Anhand auf S. 98ff. Die verwendeten Verfahren werden nachfolgend in Kürze mit ihren Vor- und Nachteilen beschrieben (vgl. Tufts 2018). Für weiterführende Informationen zu den Machine Learning Verfahren sei an dieser Stelle auf das Buch von Cleve & Lämmel verwiesen (Cleve/Lämmel 2014).

Gaussian Naïve Bayes

Naive Bayes ist der Oberbegriff für eine Reihe von Klassifikationsalgorithmen, die alle auf dem Bayes-Theorem basieren. Die Verfahren fußen auf der Annahme, dass jedes Feature unabhängig vom Wert anderer Merkmale ist. Der Gaussian Naive Bayes wird eingesetzt, wenn man von einer Normalverteilung kontinuierlicher Features

[72] http://scikit-learn.org/stable/

ausgehen kann. Folgende Vor- und Nachteile gelten für den Naive Bayes Algorithmus (Waldron 2015):

Vorteile

- Einfach zu verstehen und zu implementieren
- Für kleinere Datensätze geeignet
- Schnelle Berechnung
- Unempfindlich gegen irrelevante Merkmale und Ausreißer
- Geeignet für große Datenmengen

Nachteile

- Schlechte Ergebnisse, wenn die Features nicht unabhängig sind

Support Vector Machine

Das Support Vector Machine (SVM) Verfahren wird sowohl für Klassifizierungs- als auch Regressionszwecke eingesetzt (Bambrick 2016). SVM-Klassifikatoren basieren auf der Annahme, dass es eine Hyperebene gibt, mit der sich ein Datensatz bestmöglich in zwei Klassen unterteilen lässt. Die Vektoren, die die Hyperebene definieren, werden als Support Vectors bezeichnet. Das Verfahren lässt sich auch auf nicht linear klassifizierbare Daten anwenden. Mit dem sogenannten Kernel-Trick können Merkmale in einen höherdimensionalen Raum transformiert werden, um einfacher eine trennende Hyperebene zu finden.

Vorteile

- Genaue Berechnung
- Geeignet für kleinere, sauberere Datensätze
- Kann effizienter sein, weil es eine Teilmenge von Trainingspunkten verwendet
- Funktioniert gut mit nicht linearen Grenzen je nach verwendetem Kernel
- Gute Handhabung hochdimensionaler Daten

Nachteile

- Bedingt für große Datenmengen geeignet, da die Trainingszeit für SVM-Modelle hoch sein kann
- Weniger effektiv bei verrauschten Datensätzen mit überlappenden Klassen
- Abhängig vom Kernel anfällig für Overfitting

Logistische Regression

Die Logistische Regression ist, wie auch die vorherigen Verfahren, ein überwachter Klassifikationsalgorithmus. Im Gegensatz zur normalen Linearen Regression versucht die logistische Regression nicht, den Wert einer numerischen Variable für eine Reihe von Features vorherzusagen. Stattdessen wird die Wahrscheinlichkeit ausgegeben, mit

der ein Datensatz zu einer bestimmten Klasse gehört. Bei der Logistischen Regression gilt die Annahme, dass sich die Datenpunkte durch eine lineare Begrenzung in zwei Bereiche, einen für jede Klasse, aufteilen lassen.

Vorteile

- Einfach
- Zuverlässig
- Keine Parameter zum Einstellen
- Geringe Varianz
- Liefert Wahrscheinlichkeiten für Ergebnisse

Nachteile

- Starke Verzerrung bei nicht linearen Daten

Decision Tree Classifier

Das Decision Tree-Verfahren ist eine einfache und weit verbreitete Klassifikationstechnik. Entscheidungsbäume gliedern einen Datensatz in eine baumartige hierarchische Struktur auf. An jeder Verzweigung wird der Wert eines Merkmals verwendet, um die Fälle aufzuteilen. Dafür werden die optimale Variable und das optimale Teilungskriterium gesucht. Ziel ist es, ein Modell zu erstellen, welches den Wert einer Zielvariablen aus mehreren Eingabevariablen vorhersagt.

Vorteile

- Einfach zu interpretieren und zu erklären
- Unempfindlich gegen Ausreißer
- Auch anwendbar, wenn die Daten nicht linear trennbar sind
- Kann auch mit nominalen und ordinalen Merkmalen umgehen

Nachteil

- Neigt zum Overfitting[73]

Random Forest Classifier

„Random Forest“ ist einer der leistungsfähigsten und am häufigsten verwendeten Algorithmen. Den Namen prägte der Wissenschaftler Leo Breiman, der verschiedene Ansätze der zufälligen Generierung von Entscheidungsbäumen erforschte.

[73] Mit Überanpassung (engl. Overfitting) wird die Anpassung eines Modells an einen Datensatz bezeichnet. Das Modell ist zwar gut an den gegebenen Datensatz angepasst, lässt sich aber nicht auf andere, neue Datensätze anwenden (siehe auch: https://de.wikipedia.org/wiki/Überanpassung).

Das Verfahren wird sowohl für Regression als auch für Klassifikation verwendet. Ein Random Forest besteht aus vielen Entscheidungsbäumen, die auf zufällig ausgewählten Variablen basieren. Typischerweise wird jeder Baum an einem zufälligen Subset an Features trainiert. Bei einer Klassifikation wird für eine konkrete Vorhersage die Mehrheitsentscheidung aller Bäume verwendet.

Vorteile

- Schnell und skalierbar
- Wenig Konfigurationsparameter
- Entfernt Feature-Korrelationen (im Gegensatz zu Decision Trees)
- Reduzierte Varianz (im Gegensatz zu Decision Trees)

Nachteile

- Nicht einfach visuell zu interpretieren

Gradient Boosting Classifier

Eine Weiterentwicklung der Baummodelle[74] stellt das Gradient Boosting-Verfahren dar. Der Algorithmus erzeugt eine Abfolge einfacher Bäume und versucht dabei fortwährend, die Fehler des vorherigen Baumes zu korrigieren.

Vorteile

- Besser interpretierbar als Random Forest, da die maximale Größe der Bäume vorgegeben werden kann
- Kann mitnominalen und ordinalen Merkmalen umgehen

Nachteile

- Viele Konfigurationsparameter können die Klassifikationsgenauigkeit beeinflussen
- Bei einer großen Anzahl von Bäumen kann es zu Overfitting kommen

Klassifikation

Nach der Modellberechnung erfolgt die Klassifikation auf Basis der Testdaten. Zur Validierung der Ergebnisse wird das Modell mittels 10-facher Cross Validation getestet und anhand der Ergebnisse das Mittel der Genauigkeit und die Standardabweichung bestimmt. Die Machine-Learning-Verfahren die mittels der Python-Bibliothek scikit-learn implementiert wurden, werden mit ihren Standard-Einstellungen verwendet und es erfolgt vorerst kein Parameter Tuning.

[74] Decision Tree, Random Forest, etc.

Erster Durchlauf
Im ersten Durchlauf wurden alle 186 Features (inklusive der auffälligen lexikalischen Merkmale) für die Klassifikation verwendet. Es ergeben sich folgende Klassifikationsergebnisse:

Tabelle 15: Ergebnisse der Klassifikation mit allen 188 Merkmalen

Merkmale	Algorithmus	Accuracy	Std
186 features	Random Forest Classifier	0,84	0,08
186 features	Support Vector Machines Classifier (Linear Kernel)	0,79	0,07
186 features	Gaussian Naives Bayes	0,79	0,06
186 features	Logistic Regression	0,79	0,08
186 features	Gradient Boosting Classifier	0,78	0,08
186 features	Decision Tree Classifier	0,76	0,07

Der Random Forest Classifier liefert mit 84%-iger Genauigkeit und einer Standardabweichung von 8% die besten Ergebnisse.
Alle Arten von Baummodellen berechnen ihre Äste, indem sie mathematisch bestimmen, welche Aufteilung am effektivsten hilft, die Klassen zu unterscheiden. Daher lässt sich nach dem Training des Modells über das Attribut *.feature_importances_* die Wichtigkeit eines Features ausgeben.
Betrachtet man die Relevanz der unterschiedlichen Merkmale wird deutlich, dass genau die lexikalischen Features darunter sind, die eine hohe Korrelation mit der Textlänge ausweisen (vgl. S. 45ff). Es ist offensichtlich, dass die Ergebnisse mit diesem Featureset nicht repräsentabel sind. Vor dem zweiten Durchlauf erfolgt daher eine detaillierte Untersuchung der Merkmale wie im nächsten Abschnitt beschrieben.

Tabelle 16: Feature Importance des Random Forest Classifiers

Rang	Merkmal	Beschreibung	Wichtigkeit	Merkmals-gruppe
1	verb_variation_lemma	Variation bei Verben in Stammform	0,053	lexical
2	ttr	TTR; Kennzahl für Wortfülle	0,046	lexical
3	verb_variation	Variation bei Verben	0,043	lexical
4	ttr_corrected	Modifikation des TTR	0,042	lexical
5	adjective_variation_lemma	Variation bei Adjektiven in Stammform	0,042	lexical

6	ttr_bilogarithmic	Modifikation des TTR	0,039	lexical
7	adjective_variation	Variation bei Adjektiven	0,037	lexical
8	noun_variation_lemma	Variation bei Nomen in Stammform	0,034	lexical
9	adverb_variation	Variation bei Adverbien	0,033	lexical
10	noun_variation	Variation bei Nomen	0,030	lexical

Zweiter Durchlauf

Vor dem zweiten Durchlauf soll der Grad der Korrelation sämtlicher Merkmale mit der Wortanzahl ermittelt werden. Das Ziel ist es, Features auszuschließen, die hochgradig von der Textlänge anhängig sind. So soll vermieden werden, dass die Textlänge implizit das Ergebnis bestimmt.

Das Ergebnis der Korrelationsuntersuchung nach Pearson[75] (vgl. S. 95ff) zeigte, dass genau die lexikalischen Merkmale Textlängen-abhängig sind, die durch eine bimodale Verteilung in den Histogrammen auffallen. Anhand dieses Wahrscheinlichkeit werden anschließend die Merkmale für den zweiten Durchlauf ausgewählt ausgewählt, die keine oder nur wenig Korrelation mit der Wortanzahl aufweisen (p-Wert < 0,05). Die stark korrelierten Werte werden für den zweiten Klassifikationslauf nicht berücksichtigt.

Tabelle 17: Ergebnisse der Klassifikation mit 176 Merkmalen

Merkmale	**Algorithmus**	**Accuracy**	**Std**
176 features	Support Vector Machines Classifier (Linear Kernel)	0,71	0,06
176 features	Logistic Regression	0,70	0,07
176 features	Random Forest Classifier	0,70	0,09
176 features	Gradient Boosting Classifier	0,69	0,10
176 features	Gaussian Naives Bayes	0,65	0,07
176 features	Decision Tree Classifier	0,58	0,09

[75] Der Pearson-Korrelationskoeffizient misst die lineare Beziehung zwischen zwei Variablen. Es wird außerdem die Wahrscheinlichkeit (p-Wert) ausgegeben, dass die Korrelation durch puren Zufall entstanden ist.

Die Feature Importance wird, aus Gründen der Vergleichbarkeit mit den anderen Klassifikationsdurchläufen, in folgender Tabelle ebenfalls vom Random Forest Algorithmus angegeben, obwohl es in diesem Durchlauf nicht das beste Klassifikationsverfahren war.
Im zweiten Klassifikationsdurchlauf ergibt sich ein anderes Bild hinsichtlich der relevanten Features. Unter den zehn wichtigsten Features befinden sich auffällig viele syntaktische Merkmale, aber auch grammatikalische, lexikalische und morphologische sind vertreten. Das Ergebnis macht insgesamt einen valideren Eindruck.

Tabelle 18: Feature Importance des Random Forest Classifiers

Rang	Merkmal	Beschreibung	Wichtigkeit	Merkmals-gruppe
1	n_JU / n_sents	Ø Satzverknüpfer	0,014	syntactical
2	n_APP / n_sents	Ø Apposition / Beisatz	0,013	syntactical
3	n_OA / n_sents	Ø Akkusativobjekt	0,012	syntactical
4	n_SBP / n_sents	Ø passives Subjekt	0,012	syntactical
5	n_VVIZU / n_words	Ø Vollverb / Partikelverb im "zu"-Infinitiv	0,012	grammatical
6	n_ADJ / n_words	Ø Adjektiv	0,011	grammatical
7	n_3rd_person_pl_verbs / n_finite_verbs	Verb in der 3. Person Plural	0,011	morphological
8	lexical_density	Verhältnis lexikalischer Wörter zur Wortanzahl	0,011	lexical
9	n_PNC / n_sents	Ø Komponente vom Eigennamen	0,011	syntactical
10	n_NP / n_sents	Ø Nominalphrase	0,010	syntactical

Dritter Durchlauf

Da nicht ausgeschlossen werden kann, dass auch andere Merkmale von Satz- oder auch Wortanzahl abhängig sind, wird, wie im vorherigen Durchlauf, die Korrelation jedes Merkmals mit Wort- und Satzanzahl bestimmt.
Eine tabellarische Darstellung der Korrelationsergebnisse findet sich im Anhang auf S.98ff. Nachdem die Features entfernt wurden, die sowohl von Satz- als auch Wortanzahl abhängig sind, bleiben für die Klassifikation 46 Features übrig. Mit diesen Features wird die Klassifikation erneut durchgeführt.

Tabelle 19: Ergebnisse ohne korrelierte Merkmale

Merkmale	Algorithmus	Accuracy	Std
46 features	Gradient Boosting Classifier	0,70	0,06
46 features	Random Forest Classifier	0,69	0,10
46 features	Gaussian Naive Bayes	0,65	0,08
46 features	Support Vector Machines Classifier (Linear Kernel)	0,65	0,10
46 features	Logistic Regression	0,65	0,10
46 features	Decision Tree Classifier	0,59	0,09

In der folgenden Tabelle werden die Ergebnisse der Feature Importance-Berechnung des Random Forest Klassifikatoren dargestellt. Es zeigt sich, dass nach dem Entfernen der korrelierten Merkmale lediglich grammatikalische und syntaktische Features unter den zehn wichtigsten zu finden sind.

Tabelle 20: Feature Importance des Random Forest Classifiers

Rang	Merkmal	Beschreibung	Wichtigkeit	Merkmals-gruppe
1	n_CJ / n_sents	Ø Substantive, die durch Bindewörter verknüpft sind	0,036	syntactical
2	n_3rd_person_pl_verbs / n_finite_verbs	Verhältnis von Verben in der 3. Person Plural zu finiten Verben	0,032	morphological
3	n_RC / n_sents	Ø Relativsatz	0,029	syntactical
4	lexical_density	Verhältnis lexikalischer Wörter zur Wortanzahl	0,029	lexical
5	mtld_verb_variation	Kennzahl für Variation bei Verben	0,028	lexical
6	n_PROAV / n_words	Ø Pronominaladverb	0,027	grammatical
7	n_ADJ / n_words	Ø Adjektiv	0,027	grammatical
8	n_CONJ / n_words	Ø Bindewort	0,026	grammatical
9	n_KON / n_words	Ø nebenordnende Konjunktion	0,025	grammatical
10	n_ADV / n_words	Ø Adverb	0,025	grammatical

Vierter Durchlauf

Das Ziel des vierten Durchlaufs war es, die Performance eines neuronalen Netzes mit den beschriebenen traditionellen Maschine Learning-Verfahren zu vergleichen. Zu diesem Zweck wurde der auf neuronalen Netzen basierende Google Cloud-Dienst AutoML verwendet. Laut Produktseite[76] nutzt der, aktuell in der Beta-Version verfügbare Service, eine Kombination aus „Transfer Learning“ und „Neural Architecture Search“. Der Ansatz, den sich Transfer Learning zu Nutze macht, ist, dass neuronale Netzarchitekturen für ähnliche Klassifikationsaufgaben einen ähnlichen Aufbau nutzen. AutoML verwendet daher als Basis vortrainierte Modelle, die für ähnliche Fragestellungen entwickelt wurden, z. B. für Text- oder Bildklassifikation. Für das Verfahren spricht insbesondere, dass auch bei kleineren Datensätzen und wenig verfügbarer Rechenleistung gute Resultate erzielt werden können. Daher erscheint die Technologie perfekt geeignet für den zahlenmäßig kleinen Reportage-Korpus.

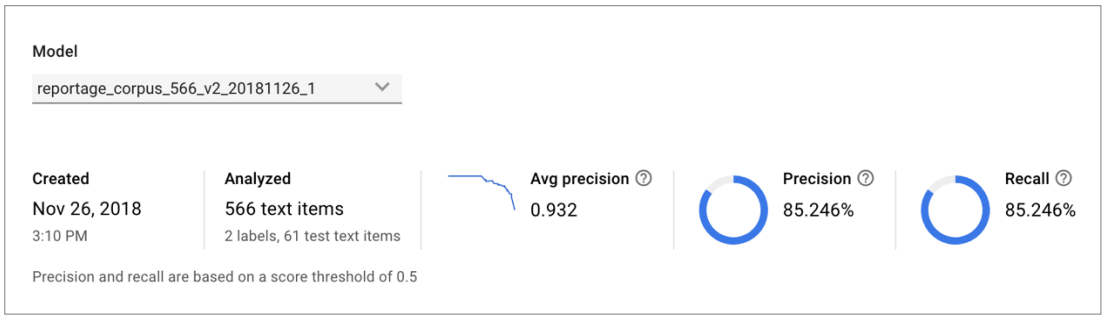


Abbildung 11: Evaluatation des AutoML-Modells vom Reportage-Korpus

Bei „Neural Architecture Search“ werden Maschine Learning-Modelle zur Steuerung verwendet, um andere Deep Learning-Modelle zu trainieren. Die Steuerungseinheit übernimmt die Auswahl einer geeigneten Netzwerkarchitektur, nimmt Parameter-anpassungen vor und versucht, das Modell iterativ zu verbessern. Die Steuerungseinheit kann außerdem dafür sorgen, dass Ebenen eines neuronalen Netzes ausgelassen werden, um die geringstmögliche Rechenleistung für ein bestmögliches Ergebnis zu verwenden. Die Wissenschaftler von Google konnten im Rahmen ihrer Forschung belegen, dass das AutoML-Modell für eine Foto-Klassifikationsaufgabe präzisere Ergebnisse lieferte als alle bisherigen manuell erzeugten Modelle (TensorFlow 2018: ab 26:15min).

[76] https://cloud.google.com/automl/

Es existiert u. a. eine Implementierung von AutoML für Textklassifikationsaufgaben. Diese wurde verwendet, um auf Basis des Reportage-Korpus ein AutoML-Modell zu berechnen.

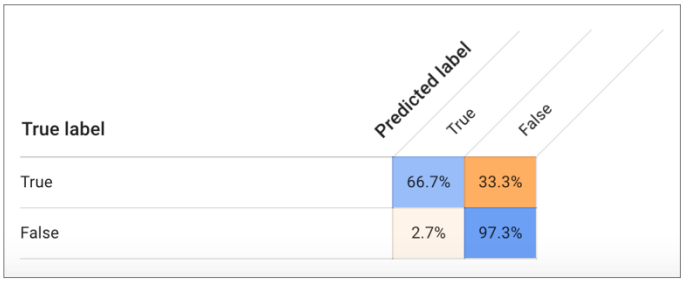


Abbildung 12: Confusion Matrix für das AutoML-Modell

Auffällig ist, dass das Modell eine hohe Präzision bei der Vorhersage der Nicht-Perlen, allerdings eine höhere Fehlerrate bei der Vorhersage von Perlen aufweist. Um die Güte des Modells mit traditionellen Machine Learning-Ansätzen mit Feature-Extraktion vergleichen zu können, werden die Werte der Confusion Matrix in Accuracy[77] umgerechnet[78]. Für das AutoML-Modell ergibt sich eine Vorhersagegenauigkeit von 82%. Die Genauigkeit der Modelle mit klassischem Feature Engineering wird also deutlich übertroffen. Das liegt natürlich auch daran, dass nicht versucht wurde, die klassischen Machine Learning-Modelle mittels Parameter Tuning zu verbessern.

4.7 Interpretation und Evaluation

Schon bei der ersten Inspektion des erstellten Korpus wurde deutlich, dass die preisgekrönten Reportagen eine höhere Wort- und Satzanzahl aufweisen als die Nicht-Perlen. Im Rahmen der Datentransformation wurden zahlreiche Merkmale berechnet, die durch Normierung auf Wort- und Satzanzahl in eine relative, vergleichbare Kenngröße umgewandelt wurden. Die Klassifikation mittels klassischer Maschine Learning-Verfahren wurde in drei Iterationen durchgeführt. Als Klassifikations-verfahren kamen jeweils Naives Bayes, Support Vector Machines, Logistic Regression, Decision Tree, Random Forest und Gradient Tree Boosting zum Einsatz. Die erste Iteration umfasste alle Features, von denen ausgegangen wurde, dass diese von der Wort- und Satzanzahl unabhängig sind. Bei den lexikalischen Features bestand aufgrund der auffälligen Histogramme der Verdacht, dass diese eine hohe Korrelation mit der Textlänge aufweisen würden, was sich bei einer tiefergehenden Untersuchung

[77] engl. für Vorhersagegenauigkeit
[78] vgl.: https://www.researchgate.net/post/How_can_I_calculate_the_accuracy

bestätigte. Um bei der Klassifikation dieser Features nicht implizit die Wort- oder Satzanzahl zu klassifizieren, wurden die betroffenen Merkmale aus der Feature-Matrix entfernt. Mit den verbleibenden Merkmalen erfolgte erneut eine Klassifikation. Da auf Basis der neuen Ergebnisse ebenfalls der Verdacht bestand, dass auch die syntaktischen Features mit der Satzanzahl korreliert sein könnten, wurde erneut eine Korrelationsanalyse mit der Wort- und Satzanzahl durchgeführt. Alle Merkmale, deren Signifikanzniveau kleiner als 0,05 war, wurden aus dem Feature-Set entfernt, wodurch sich die Anzahl der Merkmale auf 46 reduzierte. Mit diesen verbleibenden Merkmalen erfolgte erneut eine Klassifikation. Ohne Parameter Tuning der Modelle konnte eine Klassifikationsgenauigkeit von 70% erreicht werden. Die folgenden zehn Features haben sich im dritten Klassifikationsdurchlauf als die aussagekräftigsten Merkmale für Perlen herausgestellt:

Tabelle 21: Feature Importance des Random Forest Classifiers (Wort- und Satzanzahl unabhängige Merkmale)

Feature	Beschreibung	Beispiel	Mean (±SD) Nicht-Perle	Mean (±SD) Perle
n_CJ / n_sents	Ø Substantive, die durch Bindewörter verknüpft sind	Peter [und] Anna	0,540 (±0,198)	0,595 (±0,237)
n_3rd_person_pl_verbs / n_finite_verbs	Verhältnis von Verben in der 3. Person Plural zu finiten Verben	[die Kinder] lernen	0,228 (±0,066)	0,203 (±0,055)
n_RC / n_sents	Ø Relativsatz	[ein Mann,] den wir nicht kennen; [er schläft,] was mir nicht gefällt	0,138 (±0,067)	0,150 (±0,065)
lexical_density	Verhältnis lexikalischer Wörter zur Wortanzahl	-	0,485 (±0,020)	0,476 (±0,018)
mtld_verb_variation	Kennzahl für Variation bei Verben	-	319,9 (±119,6)	342,4 (±124,4)
n_PROAV / n_words	Ø Pronominaladverb	deswegen [sprechen wir] darüber	0,006 (±0,002)	0,005 (±0,002)
n_ADJ / n_words	Ø Adjektiv	[der] schlaue [Mitarbeiter]	0,071 (±0,012)	0,066 (±0,011)
n_CONJ / n_words	Ø Bindewort	und so; weder – noch; aber;	0,035 (±0,007)	0,032 (±0,006)
n_KON / n_words	Ø nebenordnende Konjunktion	[sie] und/ oder [Emma kommen] und streichen	0,030 (±0,007)	0,028 (±0,006)
n_ADV / n_words	Ø Adverb	bald schon [kommt sie] wohl	0,079 (±0,016)	0,074 (±0,015)

Wenngleich die ersten Klassifikationsdurchläufe deutlich bessere Ergebnisse geliefert haben, erscheint doch das Ergebnis des letzten Durchlaufs das valideste zu sein. Allerdings erscheinen einige Texteigenschaften der Perlen relativ unbedeutend. So beispielsweise die relative Anzahl der Substantive, die durch ein Bindewort verknüpft sind oder die Anzahl der Verben in der dritten Person Plural. Andere Merkmale ergeben, insbesondere im Hinblick auf die Empfehlungen aus der journalistischen Ausbildungsliteratur, wiederum Sinn, wie beispielsweise die lexikalische Vielfalt, die Varianz bei den Verben und die relative Anzahl an Adverbien.
Auf Basis der Korrelationsanalysen lässt sich formulieren, dass bei langen Texten die Wahrscheinlichkeit sinkt, dass neue grammatikalische oder syntaktische Muster auftauchen. Dieser Zusammenhang wurde bereits in den 1930er Jahren als Zipfschen Gesetz[79] formuliert – allerdings wird dieser Zusammenhang in der Literatur stets auf Worte bezogen, nicht aber auf grammatikalische oder syntaktische Muster.
Im vierten Durchlauf wurde Googles Cloud-Dienst AutoML für Natural Language Processing verwendet. Verglichen mit den Feature Engineering Ansätzen zeigte sich, dass die auf Deep Learning basierenden Modelle, mit 82% Vorhersagegenauigkeit, bessere Ergebnisse liefern als die traditionellen Maschine Learning-Modelle. Allerdings kann auf Basis des AutoML-Modells nicht festgestellt werden, welches der zahlreichen Texteigenschaften verantwortlich für das jeweilige Klassifikationsergebnis ist. Erste Versuche, das AutoML-Modell mit neuen, unbekannten Perlen zu testen, lassen vermuten, dass das Modell sehr an die wenigen Texte angepasst ist und wenig generalisiert. Dass außerdem wenig über die Genese des Modells bekannt ist und sich bei der aktuellen Betaversion wenig Anpassungen vornehmen lassen, ist ein entscheidender Nachteil dieses Verfahrens.
Basierend auf den Erkenntnissen der verschiedenen Durchläufe lassen sich zahlreiche Optimierungen ableiten, die im folgenden Kapitel dargestellt werden.

4.8 Nächste Schritte

Auf Basis der Merkmale konnte eine Klassifikationsgenauigkeit von knapp 70% erreicht werden. Das heißt, es gibt oberflächliche Muster bei den Reportagen, die die prämierten von den nicht-prämierten unterscheiden. Aus den Untersuchungen des experimentellen Teils ergeben sich weitere Untersuchungen und Optimierungspotentiale, die die Güte der Klassifikation steigern könnten. Einer der kritischsten Faktoren ist die Anzahl an Reportagen, die für die Klassifikation zur Verfügung standen. Verwendet wurden lediglich 566 Reportagen, von denen 254 prämiert waren und 312 keinen Preis erhalten hatten. Erste Tests des AutoML-Modells haben gezeigt, dass das Modell nicht gut generalisiert und zu Überanpassung neigt. So konnte bei

[79] siehe auch: https://de.wikipedia.org/wiki/Zipfsches_Gesetz

zehn Versuchen keine unbekannte Perlen korrekt klassifiziert werden. Es wäre daher ratsam, den Korpus mit weiteren Reportagen aufzufüllen. Dies könnte beispielsweise durch das Crawlen von menschlich kuratierten Reportagen auf piqd[80] und liesmich[81] oder in den entsprechenden Rubriken[82] der Online-Nachrichtenportale geschehen. Im Zuge dessen ließe sich auch die Textqualität granularer aufschlüsseln. So könnte beispielweise der Status der Einreichung – z. B. Shortlist, Nominierung und Gewinner – für eine feinere Textqualitäts-Abstufung verwendet werden. Allerdings ist es vor dem Hintergrund noch schwieriger genügend Reportagen für die unterschiedlichen Qualitätsgrade zusammenzustellen. Ein weiterer naheliegender Optimierungsschritt ist die Verbesserung des Vorhersagemodells. Da bei der Klassifikation mit Standardparametern gearbeitet wurde, kann man vermuten, dass sich die Klassifikationsgenauigkeit durch Parameter-Tuning noch weiter verbessern lässt. Das Testen verschiedener Einstellungen könnte beispielsweise über GridSearchCV automatisiert werden.

Das Ergebnis der Experimente war, dass die meisten grammatikalischen und syntaktischen Merkmale von der Wort- und Satzanzahl abhängig sind. Um diese Abhängigkeit zu eliminieren könnte die Anzahl der Tags anhand ihrer Häufigkeit im Gesamtkorpus gewichtet werden, wie das beispielsweise beim TF-IDF[83]-Verfahren geschieht. Ein weiterer Ansatz wäre es, die Texte in kleinere Abschnitte, beispielsweise mit jeweils 100 Wörtern, zu zerlegen und die Mediane der Tag-Anzahlen für die Klassifikation zu verwenden. Mit diesen Ansätzen lässt sich voraussichtlich eine Wort- und Satzanzahl unabhängige Metrik für die meisten Tags berechnen.

Die Studien der US-amerikanischen Forscherinnen Pitler, Nenkova und Louis für die englische Sprache bieten einen umfangreichen Fundus an Ideen für komplexere Merkmalsextraktion. So könnte zur Bestimmung des Textzusammenhangs als globale Metrik die Anzahl von bestimmten Artikeln[84], Pronomen[85] und Demonstrativ-pronomen[86] herangezogen werden (Pitler/Louis/Nenkova 2010). Ebenfalls denkbar

[80] https://www.piqd.de/reportagen

[81] http://www.liesmich.me/

[82] http://www.faz.net/aktuell/feuilleton/reportagen/,
http://www.spiegel.de/thema/reportagen/,
https://sz-magazin.sueddeutsche.de/tag/reportage

[83] Die Kennzahl TF-IDF wird zur Beurteilung der Relevanz von Termen für die Indexierung von Dokumenten eine Korpus eingesetzt. Anhand der berechneten Gewichtung können Dokumente besser hinsichtlich ihrer Relevant angeordnet werden, als es beispielsweise über Worthäufigkeit alleine möglich wäre (siehe auch: https://de.wikipedia.org/wiki/Tf-idf-Maß).

[84] Bestimmte Artikel: der, die, das, ...

[85] Pronomen: er, sie, es, wir, ihr, sie

[86] Demonstrativpronomen: dieser, diese, jenes, jener, jene, jenes, ...

wäre das Zählen von überlappenden Wörtern oder Lemmata oder die Bestimmung der Cosinus Similarity in angrenzenden Sätzen. Minima, Maxima oder Median könnten als Merkmale in die Klassifikation einbezogen werden. Pitler und Nenkova haben außerdem erfolgreich das „Brown Coherence Toolkit“ verwendet, um den Übergang von Entitäten in angrenzenden Sätzen zu erfassen (Pitler/Nenkova 2008: S. 8). Es könnte daher versucht werden, den Algorithmus für die deutsche Sprache zu implementieren.

Es konnte vielfach belegt werden, dass sich die Verknüpfung von Textpassagen positiv auf die Textqualität auswirken (Lin/Ng/Kan 2011; Pitler/Nenkova 2008). Anhand von annotierten Korpora wurden hierfür vergleichende, kausale und temporale Beziehungen zwischen Sätzen erlernt. Die erlernten Zusammenhänge wurden verwendet, um den Grad der Beziehung für unbekannte Sätze vorherzusagen. Da es auch für die deutsche Sprache entsprechend annotierte Korpora[87] gibt, könnten diese ebenfalls für das Training eines entsprechenden Klassifikators verwendet werden.

Die US-amerikanische Sprachwissenschaftlerin Louis schlägt außerdem vor, die Abfolge von spezifischen und unspezifischen Sätzen, also ihren Informationsgehalt bezogen auf die Satzlänge, zu untersuchen (Louis 2012). Als Maßzahl dient die Wortentropie: Viele unterschiedliche Wörter entsprechen einer hohen Informationsdichte, während wenige unterschiedliche Wörter für wenig Informationen sprechen. Die Informationsdichte soll über angrenzende Sätze ausgewogen sein und kann sich je nach Textabschnitt unterscheiden.

Eine weitere Idee, um den Textzusammenhang zu erfassen, ist der Vergleich der Word Embeddings angrenzender Sätze. Word Embeddings stellen Vektoren der häufigsten Wörter dar, die über große Korpora in ihrem Zusammenhang auftauchen. Sie erfassen somit den Kontext eines Wortes, Satzes oder einer Passage. Mit ihrer Hilfe könnte man die Ähnlichkeit aufeinanderfolgender Sequenzen untersuchen, um thematische Sprünge zu erkennen, die die Lesbarkeit negativ beeinflussen. Vorteilhaft erscheint bei der Verwendung von Word Embeddings, dass Ähnlichkeit auch dann festgestellt werden kann, wenn die Wörter nicht exakt gleich sind.

Zur Verbesserung der Klassifikation von Textqualität wurden Sprachmodelle erfolgreich eingesetzt (Pitler/Louis/ Nenkova 2010). Um Eigenheiten der Sprache von Perlen und Nicht-Perlen festzustellen, könnten Uni-, Bi- und Trigramme extrahiert und anhand ihrer Wahrscheinlichkeiten verglichen werden. In einer späteren Studie untersuchten Louis und Nenkova den Einsatz syntaktischer Muster in Sprachmodellen (Louis/Nenkova 2013). Ihre These war, dass die Art der Formulierung gute Texte auszeichnet, da der geübte Autor durch gezielte Verwendung von Syntax bestimmte Ziele verfolgt. Da sie ihre These für die englische Sprache, im Hinblick auf

[87] u. a. Potsdam Commentary Corpus: http://angcl.ling.uni-potsdam.de/resources/pcc.html

verschiedene Textpassagen wie Einleitung und Schlusswort, belegen konnten, liegt die Vermutung nahe, dass ähnliche Zusammenhänge auch für die deutsche Sprache gelten. Eher experimenteller Natur ist die Untersuchung sprachlicher Merkmale wie die Verwendung von Wortwitz oder Metaphern. Man kann vermuten, dass es bei der Textgattung der Reportagen Überschneidung mit der Textgattung der Lyrik gibt. Insbesondere da es das Ziel der Reportagen ist, den Leser in ihren Bann zu ziehen und mitfühlen zu lassen. Eine Idee ist es daher, Substantive und Verben – insbesondere Metaphern – aus Gedichten zu extrahieren und die bildhafte Sprache mit der in Reportagen zu vergleichen. Es wäre gut möglich, dass sich gute Reportagen der bekannten und bildhaften Sprache der Lyrik bedienen.
Da gute Reportagen außerdem der Perspektivwechsel von objektiv und sachlich zu persönlich und subjektiv auszeichnet, könnte versucht werden, diese Wechsel durch Sentiment-Analyse zu erkennen und zu quantifizieren. Bei der Sentiment-Analyse wird versucht, eine positive oder negative Haltung gegenüber einem Thema, meist anhand von Wortlisten, zu erfassen. Anhand dieser Perspektivwechsel lassen sich möglicherweise Muster für die Klassifizierung finden. Besonders reißerische Headlines werden oftmals bei Nachrichten-Teasern verwendet, um die Nutzer zum Lesen der Texte zu bewegen (O A 2015). Denkbar wäre es daher, dass sich schockierend negative Stimmung, optimistisch positive oder auch eine Mischung aus beiden Lagern positiv auf den Lesegenuss auswirken. Darüber hinaus machen insbesondere Texte mit Wortwitz Spaß beim Lesen. Es gibt einige Studien die sich damit beschäftigen, den Humor bzw. Sarkasmus in Texten zu erfassen und zu klassifizieren (de Oliveira/Rodrigo 2017). Es wäre spannend zu untersuchen, ob sich z. B. anhand von Nutzerbewertungen trainierte Modelle auf „normale“ Sätze in Reportagen anwenden lassen. Allerdings wird es vermutlich schwierig, zu diesem Zweck einen annotierten, deutschsprachigen Korpus zu finden.
Neuronale Netze wurden bei der Klassifikation mittels AutoML beiläufig untersucht. Es gibt aktuelle Studien, die erfolgreich, neben den Word Embeddings, grammatikalische (Trask/Michalak/Liu 2015) oder syntaktische (Liu u. a. 2017) Muster als Eingaben für neuronale Netze verwenden. Im Vergleich zu den Ergebnissen der klassischen Machine Learning-Verfahren könnte geprüft werden, ob die extrahierten Tags in Kombination mit Word Embeddings bessere Klassifikatios-ergebnisse liefern.

4.9 Fazit

Die Ergebnisse des experimentellen Teils haben gezeigt, dass eine Klassifikation von Reportagen basierend auf ihren oberflächlichen Texteigenschaften möglich ist und legen nahe, sich weiter mit dem Thema auseinanderzusetzen. Im vorherigen Abschnitt wurden zahlreiche Optimierungen aufgeführt, die zur Verbesserung der Klassifikationsergebnisse umgesetzt werden könnten. Bisher wurden für die Experimente, neben einfach zu berechnenden Kennzahlen, lediglich die relativen

Häufigkeiten der Annotationen von Standard-Taggern[88] verwendet. Insgesamt kann davon ausgegangen werden, dass sich die Vorhersagegenauigkeit durch die Berechnung von komplexeren Merkmalen weiter verbessern lässt. Darüber hinaus hat das Modell aktuell zwei Schwachstellen. Die erste Einschränkung ist, dass die Textqualität aktuell lediglich für Reportagen vorhergesagt werden kann. Es wäre daher wünschenswert, die Analyse auf weitere Textsorten auszudehnen. Allerdings besteht gegebenenfalls die Notwendigkeit, unterschiedliche Textsorten zu erkennen, um die Qualität isoliert vorherzusagen. Hierzu wäre ein weiterer Klassifikator von Nöten und es ergibt sich wiederholt die Herausforderung, genügend Texte für die unterschiedlichen Textgenres zusammenzustellen. Die zweite Einschränkung ist, dass das Modell nur die Werte wahr und falsch als Ausgabe zurückliefert. Besser wäre es, einen kontinuierlichen, numerischen Wert entsprechend der Qualität auszugeben. Auf Basis einer solchen kontinuierlichen Kennzahl wäre es einfacher, Verbesserungen oder Verschlechterungen bei Textänderungen nachzuvollziehen. Auch hier ergibt sich die Notwendigkeit den Korpus zu vergrößern, um genügend Texte als Beispiele für unterschiedliche Qualitätsstufen hinzuzufügen.

Insgesamt hat sich die Beschaffung und Aufbereitung geeigneter Texte im Rahmen der Arbeit als größte Herausforderung herausgestellt. Das Zusammenstellen eines Korpus aus diversen Onlinequellen ist ein aufwändiger Prozess, insbesondere da für jede Quelle der Extraktionsansatz für den Rohtextes angepasst werden muss. Zudem lagen nicht für alle Texte Informationen über die Textkategorie vor, weswegen einige der Texte nach der Extraktion nicht verwendet werden konnten.

Es herrscht aktuell auch Unsicherheit darüber, ob die Nutzung von fremden Texten zukünftig weiterhin im rechtlichen Graubereich liegt. Der vorläufige Gesetzesentwurf zum europäischen Leistungsschutzrecht sieht vor, dass „jede ‚Digitale Nutzung', also auch das Crawlen und Abspeichern von Artikeln in einer Datenbank", (O A 2018a) zukünftig eine Lizensierung der Inhalte erfordert. Das würde voraussichtlich einen hohen Administrationsaufwand und hohe Kosten bedeuten und in der Folge weitere Entwicklung für dem Bereich ausbremsen. Grundsätzlich wäre wünschenswert, dass es mehr frei nutzbare Korpora für die wissenschaftliche Forschung gibt, um die Forschungslücke zu Textqualität für die deutsche Sprache rasch zu schließen.

[88] textacy (spacy), Standford Parser, RDRPOSTagger

5 Ausblick

Die Klassifikationsergebnisse des experimentellen Teils belegen, dass es oberflächliche Muster in Reportagen gibt, die prämierte von den normalen, nicht-prämierten unterscheiden. Diese Resultate decken sich mit Forschungsergebnissen zu Textqualität in englischer Sprache (Louis/Nenkova 2013) und können vielfältig genutzt werden. Dennoch ist das Fachgebiet insbesondere für die deutsche Sprache wenig erforscht und weiterführende Forschung ist von Nöten, um die Ergebnisse tatsächlich anwendbar zu machen. Im redaktionellen Alltag können die Ansätze zukünftig verwendet werden, um innerhalb eines Redaktionssystems Vorschläge für Textoptimierung zu geben. Es gibt bereits Webservices[89] für die englische Sprache, die einfache Textkennzahlen ausgeben und Autoren auf zu lange Sätze, zu viele Adverben und ein negatives Sentiment hinweisen. In ähnlicher Form könnten die Hinweise im Redaktionssystem ausgestaltet werden.

Ein weiterer Anwendungsbereich im redaktionellen Umfeld ist die Textqualitäts-messung von maschinell generierten Texten. Wie eingangs geschrieben hält Roboterjournalismus in diversen redaktionellen Sparten Einzug. Auf Basis der ermittelten Textqualitätsparameter könnten die Maschinen selbst Qualitätssicherung betreiben und ihre Schreibfertigkeiten weiter optimieren. So könnte eine Maschine beispielsweise diverse Varianten des selben Textes erzeugen und am Ende berechnen, welche Version vom Textqualitätsmodell am besten bewertet wird. Diese Textvariante wäre dann diejenige, die weiter verwendet wird.

Da jeder Leser über einen spezifischen Geschmack hinsichtlich der oberflächlichen Textmerkmale verfügt, könnten die Ergebnisse ebenfalls verwendet werden, um Lese-empfehlungen besser an die Nutzerbedürfnisse anzupassen. Bei dieser Idee geht es darum, die Texteigenschaften an einen bestimmten Nutzer und seine jeweilige Nutzungssituation anzupassen. Gerade im Hinblick auf die Nutzungskontext bieten sich interessante Anwendungsszenarien, da Leser unterwegs eher zu kurzen Texten neigen, während bei der Nutzung zu Hause auch mal ein ausführlicher Bericht konsumiert wird. So könnten Headlines und Texte angepasst werden, um dem Leser für Ausgabekanal und Nutzungssituation ein optimales Leseerlebnis zu bieten. Außerdem könnte ein Textqualitätsindex für das Ranking innerhalb der Treffer von

[89] z. B. https://readable.io/

Suchmaschinen-Ergebnislisten angeboten werden. Eine weitere Produktidee stellt ein Browser-Plugin dar, welches die oberflächlichen Texteigenschaften einer besuchten Webseite auswirft. Somit könnte sich der Textkonsument vor dem Lesen einen objektiven Eindruck der oberflächlichen Qualitätsparameter verschaffen.
Die skizzierten Anwendungsszenarien verdeutlichen, dass es vielfältige Einsatzbereiche für die maschinelle Analyse von Texten gibt. Nicht alle Ansätze zielen auf die Bestimmung von Textqualität ab, könnten sich aber die Ansätze der Merkmalsextraktion zu Nutze machen, um das individuelle Leseerlebnis optimal zu gestalten.

6 Literaturverzeichnis

Alikaniotis, Dimitrios/Yannakoudakis, Helen/Rei, Marek (2016): Automatic Text Scoring Using Neural Networks, 2016, S. 715–725.

Aluisio, Sandra/Specia, Lucia/Gasperin, Caroline/Scarton, Carolina (2010): Readability assessment for text simplification, in: *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, 2010, S. 1–9.

Bambrick, Noel (2016): Support Vector Machines for dummies; A Simple Explanation, in: *AYLIEN*, http://blog.aylien.com/support-vector-machines-for-dummies-a-simple/ [23.05.2018].

Barbaresi, Adrien (2011): A note on Amazon's text readability stats, online: *Bits of Language*, http://adrien.barbaresi.eu/blog/a-note-on-amazons-text-readability-stats.html [08.05.2017].

Barbaresi, Adrien (2012): Amazon's readability statistics by example, online: *Bits of Language*, http://adrien.barbaresi.eu/blog/amazons-readability-statistics-by-example.html [08.05.2017].

Barzilay, Regina/Lapata, Mirella (2008): Modeling local coherence: An entity-based approach, in: *Computational Linguistics*, 2008, S. 1–34.

BDZV (2017): Theodor-Wolff-Preis: Die Ausschreibung, online: *Theodor Wolff Preis*, http://www.bdzv.de/twp/ausschreibung/ausschreibung2017/ [23.06.2017].

BDZV (2018): Nominierungen 2018, online: *Theodor Wolff Preis*, http://www.bdzv.de/twp/nominierte-texte/2018/ [09.08.2018].

vor der Brück, Tim/Hartrumpf, Sven (2007a): A Readability Checker Based on Deep Semantic Indicators., in: *LTC*, 2007, S. 232–244.

vor der Brück, Tim/Hartrumpf, Sven (2007b): A semantically oriented readability checker for German, in: *Proceedings of the 3rd Language & Technology Conference*, 2007, S. 270–274.

vor der Brück, Tim/Hartrumpf, Sven/Helbig, Hermann (2008a): A readability checker with supervised learning using deep indicators, in: *Informatica*, 2008.

vor der Brück, Tim/Hartrumpf, Sven/Helbig, Hermann (2008b): A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators, http://www.vimistschlimm.de/papers/brueck_hartrumpf_helbig08.pdf [17.11.2018].

vor der Brück, Tim/Helbig, Hermann/Leveling, Johannes (2008): The Readability Checker Delite: Technical Report, 2008.

Callan, Kevyn/Collins-Thompson, Jamie (2004): A language modeling approach to predicting reading difficulty, 2004.

Cleve, Jürgen/Lämmel, Uwe (2014): *Data Mining*, München: De Gruyter Oldenbourg.

Dell'Orletta, Felice/Montemagni, Simonetta/Venturi, Giulia (2011): Read-it: Assessing readability of italian texts with a view to text simplification, in: *Proceedings of the second workshop on speech and language processing for assistive technologies*, 2011, S. 73–83.

DuBay, William H. (2004): The Principles of Readability.

DuBay, William H. (2006): *The Classic Readability Studies*, Costa Mesa: Impact Information.

Feng, Lijun (2010): *Automatic readability assessment,* City University of New York.

François, Thomas/Fairon, Cédrick (2012): An AI readability formula for French as a foreign language, in: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, S. 466–477.

Frens, John (2017): *Lexical Diversity. Repository for two methods of scoring lexical diversity, MTLD and HD-D.*

Góes, Anderson Roges Teixeira/Steiner, Maria Teresinha Arns/Peniche, Rodrigo Antonio/Góes, Anderson Roges Teixeira/Steiner, Maria Teresinha Arns/Peniche, Rodrigo Antonio (2015): Classification of power quality considering voltage sags in distribution systems using KDD process, in: *Pesquisa Operacional* 35, 2, S. 329–352.

Hancke, Julia/Vajjala, Sowmya/Meurers, Detmar (2012): Readability Classification for German using Lexical, Syntactic, and Morphological Features., in: *COLING*, 2012, S. 1063–1080.

Heilman, Michael/Collins-Thompson, Kevyn/Callan, Jamie/Eskenazi, Maxine (2007): Combining lexical and grammatical features to improve readability measures for first and second language texts, in: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 2007, S. 460–467.

Heilman, Michael/Collins-Thompson, Kevyn/Eskenazi, Maxine (2008): An analysis of statistical models and features for reading difficulty prediction, in: *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, 2008, S. 71–79.

Hong, Kai/Conroy, John M/Favre, Benoit/Kulesza, Alex/Lin, Hui/Nenkova, Ani (2014): A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization

Johansson, Victoria (2008): Lexical diversity and lexical density in speech and writing: a developmental perspective

Karamanis, Nikiforos/Mellish, Chris/Poesio, Massimo/Oberlander, Jon (2008): Evaluating centering for information ordering using corpora, in: *Computational Linguistics*, 2008, S. 29–46.

Krug, Christian/Petzold, Andreas (2016): Nannen-Preis-2017 Satzung, https://www.nannen-preis.de/download/Nannen-Preis-2017_Satzung.pdf [08.05.2017].

Larsson, Patrik (2006): *Classification into Readability Levels.*

Li, Junyi Jessy/Nenkova, Anna (2015): Fast and Accurate Prediction of Sentence Specificity, https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9941/9554 [09.08.2018].

Lin, Ziheng/Ng, Hwee Tou/Kan, Min-Yen (2011): Automatically evaluating text coherence using discourse relations, in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, S. 997–1006.

Liu, Rui/Hu, Junjie/Wei, Wei/Yang, Zi/Nyberg, Eric (2017): Structural Embedding of Syntactic Trees for Machine Comprehension, in: *arXiv:1703.00572 [cs]*.

Louis, Annie (2012): Automatic metrics for genre-specific text quality, in: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 2012, S. 54–59.

Louis, Annie/Nenkova, Ani (2011a): Automatic identification of general and specific sentences by leveraging discourse annotations, vom 8. November 2011.

Louis, Annie/Nenkova, Ani (2011b): Text specificity and impact on quality of news summaries, in: *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, 2011, S. 34–42.

Louis, Annie/Nenkova, Ani (2013): A corpus of science journalism for analyzing writing quality, in: *Dialogue & Discourse* 3, 2, S. 87–117.

McCarthy, Philip M./Jarvis, Scott (2010): MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment, in: *Behavior Research Methods* 42, 2, S. 381–392.

Nenkova, Ani (2012): Automatic Text Understanding of Content and Text Quality, 2012.

Nenkova, Ani/Chae, Jieun/Louis, Annie/Pitler, Emily (2010): Structural Features for Predicting the Linguistic Quality of Text, in: *Empirical Methods in Natural Language Generation*, 2010, S. 222–241.

O A (2015): Das große Klick-Dilemma: Leitmedien im digitalen Reichweiten-Rattenrennen › Meedia, https://meedia.de/2015/12/10/das-grosse-klick-dilemma-leitmedien-im-digitalen-reichweiten-rattenrennen/ [05.08.2018].

O A (2016a): Einen guten Bericht und eine gute Reportage schreiben, http://www.online-lernen.levrai.de/deutsch-uebungen/bericht-reportage/01_bericht_reportage_aufsatz.htm [02.03.2018].

O A (2016b): Future News, online: *Digital News Initiative*, https://newsinitiative.withgoogle.com/dnifund/dni-projects/future-news/ [06.08.2018].

O A (2016c): G+J Pressedatenbank - Leistungen, http://www.pressedatenbank.guj.de/PDB/Leistungen.htm [02.03.2018].

O A (2016d): Lexical Density, http://www.analyzemywriting.com/lexical_density.html [26.03.2018].

O A (2017a): Offizielle Webseite | NANNEN PREIS, https://www.nannen-preis.de/ [08.01.2017].

O A (2017b): Reporter-Forum: 2017, http://www.reporter-forum.de/index.php?id=231 [07.08.2018].

O A (2018a): Leistungsschutzrecht/Uploadfilter: Worüber das Europaparlament wirklich abstimmt - Golem.de, https://www.golem.de/news/leistungsschutzrecht-uploadfilter-worueber-das-europaparlament-wirklich-abstimmt-1807-135322.html [06.08.2018].

de Oliveira, Luke/Rodrigo, Alfredo Lainez (2017): Humor Detection in Yelp reviews

Östling, Robert/Grigonyte, Gintare (2017): Transparent text quality assessment with convolutional neural networks, 2017, S. 282–286.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Ergänzungen zu Korpuslinguistik, http://corpora.ids-mannheim.de/libac/doc/libac-addOn-LexikalVielfalt.pdf [26.03.2018].

Petersen, Sarah E./Ostendorf, Mari (2006): A Machine Learning Approach to Reading Level Assessment, vom 6. Juni 2006.

Pitler, Emily/Louis, Annie/Nenkova, Ani (2009): Automatic sense prediction for implicit discourse relations in text, in: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, 2009, S. 683–691.

Pitler, Emily/Louis, Annie/Nenkova, Ani (2010): Automatic Evaluation of Linguistic Quality in Multi-Document Summarization, in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Juli 2010, S. 544–554.

Pitler, Emily/Nenkova, Ani (2008): Revisiting readability: A unified framework for predicting text quality, in: *Proceedings of the conference on empirical methods in natural language processing*, 2008, S. 186–195.

Redaktion alpha Lernen, Prof Dr Juliane Köster (Fachberatung) (2017a): Journalistische Textsorten: Feature und Reportage

Redaktion alpha Lernen, Prof Dr Juliane Köster (Fachberatung) (2017b): Journalistische Textsorten: Meldung, Nachricht, Bericht

Reporter Forum e.V. (2018): Reporter-Forum: Neuer Journalismus, http://www.reporter-forum.de/ [05.11.2018].

ryte.com (2016): Flesch-Reading-Ease, online: *Ryte.com*, https://de.ryte.com/wiki/Flesch-Reading-Ease [27.09.2017].

Salchert, Monka (2012): *Verständliches Schreiben – Mehr Erfolg durch gute Texte*, Brühl: Bundesakademie für öffentliche Verwaltung im Bundesministerium des Innern.

Schneider, Wolf/Raue, Paul-Josef (2012): *Das neue Handbuch des Journalismus und des Online-Journalismus*, Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Schreibhaus, Serpils (2015): Schreiber der Götter – Von der Höhlenmalerei zur Schrift in 32000 Jahren, in: *Serpils Schreibhaus*, https://schreibhaus.wordpress.com/2015/08/18/von-der-hoehlenmalerei-zur-schrift-in-32000-jahren/ [05.11.2018].

Schwarm, Sarah E./Ostendorf, Mari (2005): Reading level assessment using support vector machines and statistical language models, in: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, S. 523–530.

Serrao, Marc Felix (2016): Algorithmen über Bücher: Der Bestseller-Code, online: *Frankfurter Allgemeine Zeitung*, http://www.faz.net/aktuell/wirtschaft/bestsellercode-fuer-bestseller-jodie-archer-und-matthew-l-jockers-14493720.html [08.05.2017].

Si, Luo/Callan, Jamie (2001): A statistical model for scientific readability, in: *Proceedings of the tenth international conference on Information and knowledge management*, 2001, S. 574–576.

Steendam, Elke van/Tillema, Marion/Rijlaarsdam, Gert/Bergh, Huub van den (2012): *Measuring Writing: Recent Insights into Theory, Methodology and Practice,* BRILL.

Keynote (TensorFlow Dev Summit 2018), LAND 2018, .

Trask, Andrew/Michalak, Phil/Liu, John (2015): sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings, in: *arXiv:1511.06388 [cs]*.

Tufts, Chris (2018): *Cheat_Sheets: Cheat sheets for stats/ML/SP/DM*.

Vajjala, Sowmya/Meurers, Detmar (2012): On improving the accuracy of readability classification using insights from second language acquisition, in: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 2012, S. 163–173.

Waldron, Mike (2015): Naive Bayes for Dummies; A Simple Explanation, https://www.datasciencecentral.com/profiles/blogs/naive-bayes-for-dummies-a-simple-explanation [23.05.2018].

Yang, Yinfei/Nenkova, Anna (2014): Detecting Information-Dense Texts in Multiple News Domains, https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8430/8622 [09.08.2018].

7 Tabellenverzeichnis

8 Abbildungsverzeichnis

9 Anhang

9.1 Beschreibung zu Tagger/Parser-Ausgaben

Universal Part-of-Speech Tags[90] (spaCy)

POS	Description	Examples
ADJ	Adjective	[der] schlaue [Mitarbeiter]
ADP	Adposition	in [der Bäckerei], auf [dem Bahnhof]
ADV	Adverb	[Das Auto fährt] schnell
AUX	Auxiliary	[Er] hat [gelernt], [sie] sollte [aufstehen]
CCONJ	Coordinating conjunction	[Ich habe viel geschwitzt,] weil [es die Tage sehr warm war.]
DET	Determiner	dieser, jener, welcher, ein, der, die, das,…
INTJ	Interjection	Huch, Ah
NOUN	Noun	[Die] Arbeiter [machen Pause.]
NUM	Numeral	Fünf, 27
PART	Particle	[Sie stehen] auf
PRON	Pronoun	Er [mag die Pizza], ihm [gefällt das Lied]
PROPN	Proper noun	London, Jupiter, Microsoft, Sarah
PUNCT	Punctuation	.;?*. „
SCONJ	Subordinating conjunction	[Ich nutze das Rad,] um [zur Arbeit zu kommen]
SYM	Symbol	@☺
VERB	Verb	[Wir] gehen [durch die Nacht.]
X	Other	Good morning! I love you

[90] http://universaldependencies.org/u/pos/

STTS Part-of-Speech Tags[91] (spaCy)

POS	Description	Examples
ADJA	attributives Adjektiv	[der] schlaue [Mitarbeiter]
ADJD	adverbiales ODER	[er spricht] schnell
	prädikatives Adjektiv	[Sein Sprechen ist] schnell
ADV	Adverb	Bald schon [kommt sie] wohl
APPR	Präposition; Zirkumposition links	nach [Berlin]; ohne [Hund]
APPRART	Präposition mit Artikel	zum [Streichen]; zur [Sache]
APPO	Postposition	[ihm] zuliebe; [der Sache] wegen
APZR	Zirkumposition rechts	[von mir] aus
ART	bestimmter ODER unbestimmter Artikel	Der [Mann schenkte] eine [Rose] einer [unerwarteten Frau]
CARD	Kardinalzahl	zwei [Männer im Jahre] 1994
FM	Fremdsprachliches Material	[Er sagte:"] Hasta luego [,] amigos [."]
ITJ	Interjektion	Mhm; ach; tja; dann halt nicht
KOUI	unterordnende Konjunktion mit (zu-)Infinitiv	[Sie kommt,] um [zu arbeiten]; anstatt [anzufangen, geht sie wieder]
KOUS	unterordnende Konjunktion	[Emma wartet,] weil/ ob/ solange/ dass [sie stiehlt]
KON	nebenordnende Konjunktion und, oder, aber	[sie] und/ oder [Emma kommen] und streichen
KOKOM	Vergleichskonjunktion als, wie	[blauer] als [er]; [blau] wie [er]
NN	normales Nomen	[am] Tage [dem] Mann [den] Schlaf
NE	Eigennamen	[die] Emma; [dem] Hans [sein] HSV
PDAT	attribuierendes Demonstrativpronomen	Jene [Männer sprachen] dieses [lockere Spanisch]
PDS	substituierendes Demonstrativpronomen	Denen [war] dies [nicht übelzunehmen]
PIAT	attribuierendes Indefinitpronomen	Manche [Rose währt] einige [Tage]

[91] https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/STTS_Tagset_Tiger; http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

PIS	substituierendes Indefinitpronomen	Manche [verzeihen] niemandem
PPER	(nicht-reflexives) Personalpronomen	Er [schenkt] sie der Freundin
PPOSAT	attribuierendes Possessivpronomen	Unsere [Wand ist rosa]
PPOSS	substituierendes Possessivpronomen	Meiner [schlägt] deinen
PRELS	substituierendes Relativpronomen	[die Mannschaft,] der [du nacheiferst]
PRELAT	attribuierendes Relativpronomen	[die Mannschaft,] deren [Aura du verehrst]
PRF	Reflexivpronomen	[Erinnere] dich [, wie er] sich [ereiferte]
PWS	substituierendes Interrogativpronomen	Wer [hat] wen [bestohlen?]
PWAT	attribuierendes Interrogativpronomen	Wessen [Hund wurde gestohlen?]
PWAV	adverbiales Interrogativpronomen	Warum [schneidest du die Rose?]
PROAV	Pronominaladverb	Deswegen [sprechen wir] darüber
PTKZU	zu vor Infinitiv	[Ich versuchte] zu [schlafen]
PTKNEG	Negationspartikel nicht	Nicht [schlecht, wie du] nicht [hinsiehst]
PTKVZ	abgetrennter Verbzusatz/Verbpartikel	[Pass] auf [und hör] zu
PTKANT	Antwortpartikel	ja; nein; danke; bitte
PTKA	Partikel "am" o. "zu" vor Adjektiv o. Adverb	Zu [teure Rosen welken] am [schnellsten]
TRUNC	abgetrenntes Kompositionserstglied	[Mallorca liegt zwischen] An- [und Abreise]
VVFIN	finites Vollverb	[Wir] passen [auf und] hören
VAFIN	finites Voll- oder Kopulaverb	[Sie] ist [blumig.]; [Du] hast [weggehört]
VMFIN	finites Modalverb	[Sie sollte] passen
VVINF	infinites Vollverb	[Wir wollen] weghören
VAINF	infinites Hilfsverb oder Kopulaverb	[Sie soll rot geworden] sein
VMINF	infinites Modalverb	[Er hat nicht schlafen] können
VVIMP	Vollverb im Imperativ	Pass [auf und] hör [zu!]
VAIMP	Kopulaverb im Imperativ	Hör [zu!]
VVPP	partizipiales Vollverb (Partizip II)	[Wir haben] verschlafen
VAPP	partizipiales Hilfs-/Kopulaverb (Partizip II)	[Das ist verdrängt] worden
VMPP	partizipiales Modalverb (Partizip II)	[Sie hat spielen] gedurft
VVIZU	Vollverb/Partikelverb im "zu"-Infinitiv	[Wir planen] wegzuhören

XY	Nichtwort, Sonderzeichen, Kürzel	[Es enthält viel] D2XW3
$,	Komma	,
$(	sonstige satzinterne Interpunktion	( ) u.a.
$.	satzbeendende Interpunktion	. ? ! ;

Syntactical Tags[92] (Standford Parser)

Tag	Meaning	Description	Example
AA	superlative phrase with "am"		[Karl lachte] am lautesten; [der] am lautesten [lachende Mann]
AP	adjektive phrase	adjective (ADJA, ADJD, MTA, also CARD) + its dependents	[das] vom Repertoire her unerschrockene [Ensemble]; [die] an sich netten [Melodien]; Peter ist [an sich ganz nett]; [der] nach Hamburg fahrende [Mann]; gar keine [Freude]; ca. 10 [Jugendliche]; rund 10.000 Jahre
AVP	adverbial phrase	phrase headed by an adverb	gar nicht so wichtig, dass
CAC	coordinated adposition		[die Züge] von und nach [Hamburg]
CAP	coordinated adjektive phrase	the conjuncts are APs, CAPs and adjectives (ADJA, ADJD, CARD)	[die] alten und neuen [Ideen]
CAVP	coordinated adverbial phrase	coordination of AVPs, CAVPs and adverbs (ADV)	nur heute [oder nie]; heute und morgen
CCP	coordinated complementiser		ob und wann [er kommt]
CH	chunk		
CNP	coordinated noun phrase	possible conjuncts: NPs, NNs, NEs, MPNs	ein Mann und eine Frau; Peter und sein Onkel; die Jungen und Mädchen

[92] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/knoten.html; http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/annotation/tiger_introduction.pdf

CO	Coordination	coordination of different categories (S+NP, NP+PP, etc.)	jeden Tag oder zumindest am kommenden Freitag
CPP	coordinated adpositional phrase	coordination of PPs and prepositional adverbs (PP+PP, PP+PROAV,...)	in der Stadt und auf dem Lande; für die Reformen oder dagegen
CS	coordinated sentence	coordination of sentences (or sentence chunks)	Peter kommt und Paul geht; er kam, sah und siegte
CVP	coordinated verb phrase (non-finite)	coordination of (possibly incomplete) non-finite VPs	[Er wollte] Peter besuchen und Hans anrufen; [Er wollte] uns besuchen oder anrufen
CVZ	coordinated zu-marked infinitive		zu vergessen [und] zu vergeben; einzusteigen [und] zu bleiben
DL	discourse level constituent	discourse unit without explicit syntactic dependencies between its components	["Lass mich in Ruhe!"] ärgerte sich Peter
ISU	idiosyncratic unit		Löwenzahn
MPN	multi-word proper noun	name consisting of several words (1st name + family name, etc.)	Karl Schulz; Bad Münstereifel
MTA	multi-token adjective	adjective related to a multi-token proper noun	[die] Bad Godesberger [Bürger]
NM	multi-token number		eine Million Menschen; 10.000 Menschen
NP	noun phrase	head noun + its modifiers, complements, determiners.	der Mann; der alte Mann; der Mann aus Hamburg
PP	adpositional phrase	basically, an NP containing a pre- post- or circumposition. Note that there is no embedded NP in the P	in der Stadt; meiner Meinung nach; um die Stadt herum; dagegen, dass er kommt
QL	quasi-language		
S	sentence	in most cases, a finite verb + its dependents	Hans liebt Maria. [S: Gut, [S: dass du kommst]]
VP	verb phrase (non-finite)	a non-finite verb form with its dependents	[Peter will] uns besuchen

VZ	zu-marked infinitive	only if "zu" is written as a separate word	[Er versucht, uns] zu täuschen

Syntactic Dependency Parsing[93] (spaCy)

Tag	Meaning	Description	Example
AC	Adpositional case marker	Preposition/ postposition in a PP, annotated as a sister constituent of the determiner, adjectives, noun etc.	auf [dem Dach]
ADC	Adjective component	Component of a multi-token adjective (MTA)	[die] Bad Godesberger [Bürger]
AG	Genitive attribute		[die Mutter] des Freundes
AMS	Measure argument of adjective	Accusative-marked measure arguments	zwanzig Jahre [alt]
APP	Apposition	"inserted" phrase, further specifying/modifying the entity described by the matrix NP.	[der Mann,] ein Mittvierziger
AVC	Adverbial phrase component	Component of a head-less AVP	immer wieder
CC	Comparative complement	Argument of comparative adjectives/adverbs, sometimes of nouns	[jünger] als er; älter [als wir dachten]; [einer] wie du
CD	Coordinating conjunction		[Peter] und [Anna]
CJ	Conjunct	Constituent participating in coordination (cf. CD)	Peter [und] Anna
CM	Comparative conjunction	‘wie', 'als' and 'denn' in comparative constructions	besser [als Peter]
CP	Complementizer	Complementizer introducing a subordinate clause or a VP	dass [er uns kennt]; um [ihn anzurufen]
CVC	Collocational verb construction	Dative object/ 'free dative'	[der Student muss sich] einer Prüfung unterziehen
DA	Dative		[Ich gebe] dem Hund [das Futter]
DH	Discourse-level head	Head of a discourse-level constituent (DL)	["Lass mich in Ruhe!"] ärgerte sich Peter

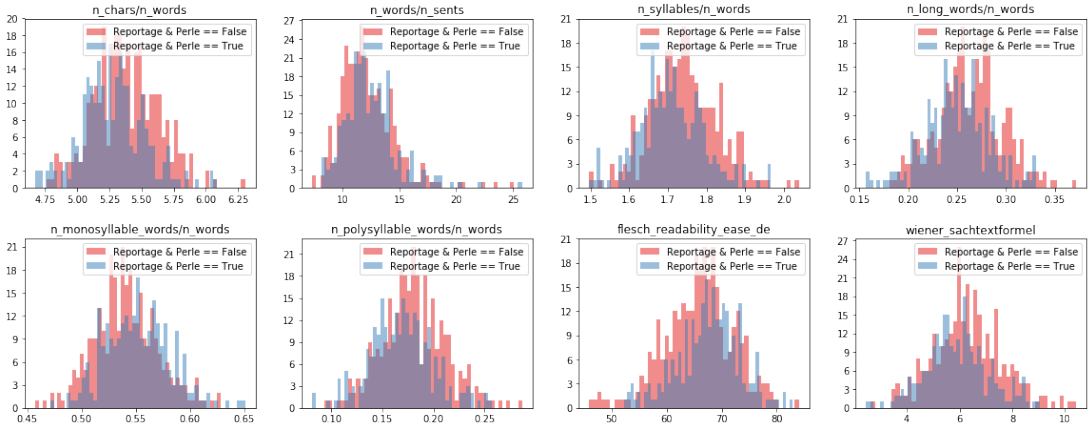
[93] https://spacy.io/api/annotation; http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/kanten.html; https://nats-www.informatik.uni-hamburg.de/CDG/NegraCorpusEdges

DM	Discourse marker	Basically, 'ja', 'nein', etc.	Ja [, das funktioniert.]
EP	Expletive es		[Ihr Geld ist ja nicht weg.] Es [haben jetzt nur andere.]
HD	Head	There is no explicitly marked head in NPs, PPs, DLs and coordinated constituents.	[der] Hut [des Mannes]
JU	Junctor	like CD, but with only one conjunct	Und [Peter ging weg.]
MNR	Postnominal modifier	Postnominal NP/PP modifier (or complement).	[das Haus] in Hamburg; [der Verzicht] auf Atomwaffen; [das Haus,] 1888 errichtet
MO	Modifier	MO denotes different functions in different phrases: 1. in S/VPs/APs: - modifiers (adjuncts) - prepositional objects 2. in NPs/PPs: - focus adverbs such as: 'nur', `auch', 'sogar', 'vor allem'	1. [er wartet] hier [auf mich]; 2. nur [am kommenden Freitag]
NG	Negation	the negation particle 'nicht' (also modified)	[er schläft] nicht; [er schläft] gar nicht.
NK	Noun kernel element		die große schwarze Katze
NMC	Numerical component	Part of a multi-token number (NM)	10.000 Leute; eine Million] Leute
OA	Accusative object	Accusative objects of verbs, participles and certain adjectives	[ich sehe] ihn; [ich bin] es [gewohnt]
OA2	Second accusative object	second accusative objects of verbs like 'lehren'	[er lehrte ihn] Deutsch
OC	Clausal object	VP/S subcategorised by a verb, adjective or a noun.	[er hat] geschlafen; [er verspricht,] uns zu besuchen; [der Beschluss,] ein Haus zu bauen
OG	Genitive object	Genitive objects of verbs, participles and certain adjectives	[es bedarf] großer Anstrengungen; [er ist] des Englischen [mächtig]
OP	Prepositional Object		[Die Feier ist nach der] Hochzeit
PAR	Parenthetical element		[John,] der Freund von Carlos [, ist ein exzellenter Schwimmer]

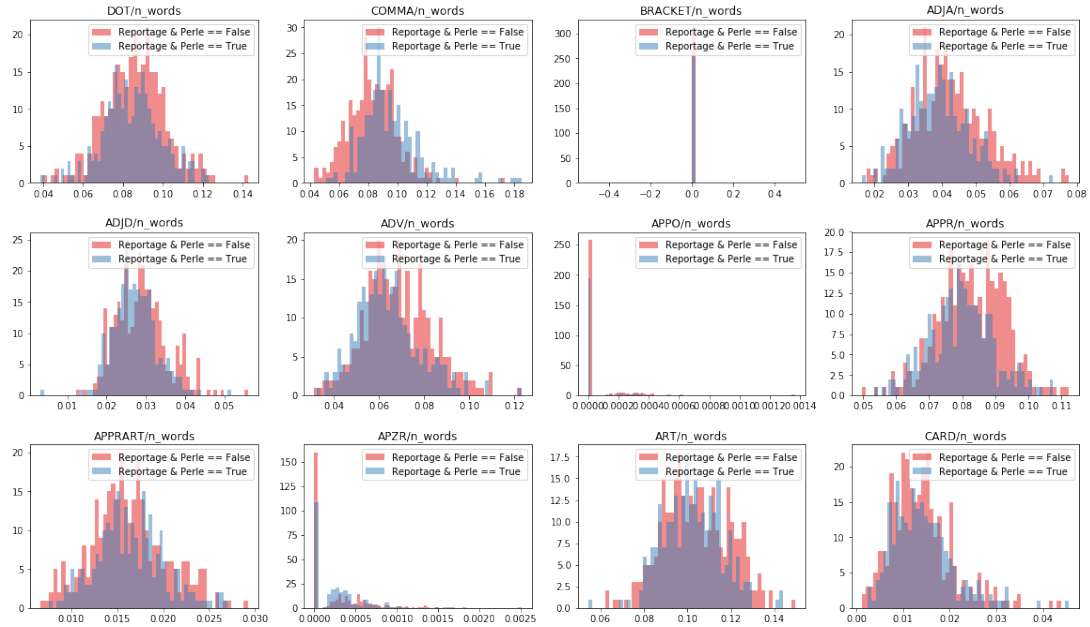
PD	Predicate	Predicative AP/NP/PP, typically in a copular construction	[Peter ist] krank; [die Tür ist] offen; [Karl ist] Lehrer; [Hans ist in] Berlin
PG	Phrasal genitive	A 'von'-PP used instead of an adnominal genitive	[der Hut] von Peter
PH	Placeholder	A pronominal adverb or pronoun ('es') correlating with an extraposed constituent.	[Ich glaube nicht] daran [, dass er kommt]; [er liebt] es [, ein Auto zu fahren]
PM	Morphological particle	Just two cases: the infinitival 'zu' (zu gehen) the adjectival (superlative) 'am' (am besten)	zu [singen]; am [schlechtesten]
PNC	Proper noun component	every daughter node of a multi-token proper name (MPN)	Peter J. Mueller
RC	Relative clause	A relative clause is a relative clause	[ein Mann,] den wir nicht kennen; [er schläft,] was mir nicht gefällt
RE	Repeated element	An extraposed constituent replaced in situ by a correlate (placeholder, PH)	cf. PH
RS	Reported speech	The complementary function to DH (discourse head) in a DL phrase.	[Sie sagte, sie könne tanzen,] wenn sie wolle
SB	Subject	A subject is a subject. Possible only within S.	[Der] Mann [kennt uns] Was er sagt [, interessiert mich nicht]
SBP	passivised subject (PP)		Peter [wurde von jemanden gesehen]
SP	Subject or predicate		Peter [war] Paul
SVP	Separable verb prefix		[Das Kind schläft] ein
UC	(Idiosyncratic) unit component		Löwenzahn
VO	Vocative		Hans [, wo gehst du jetzt hin?]

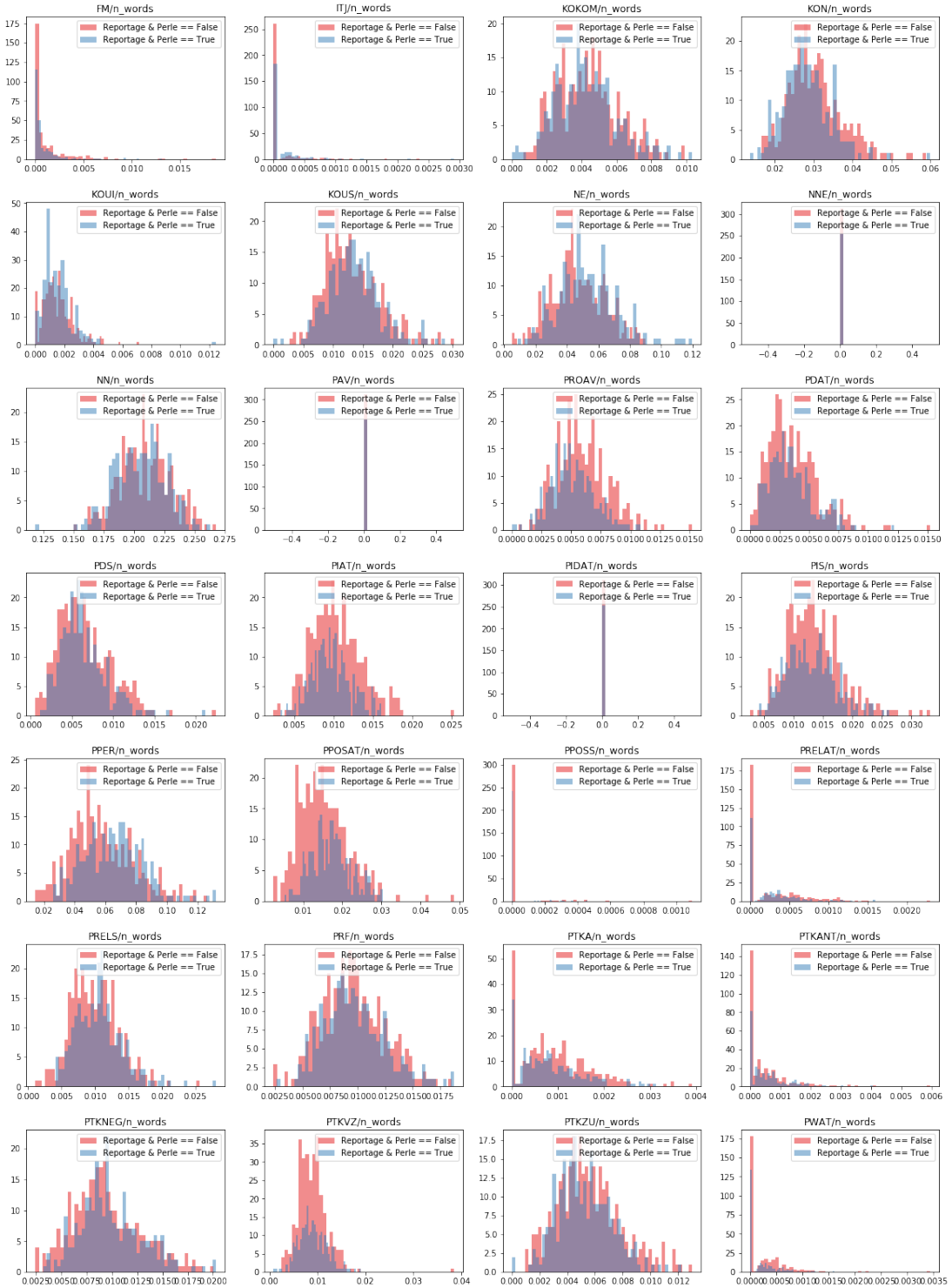
9.2 Histogramme der berechneten Feature-Gruppen

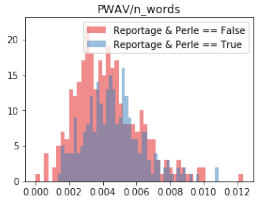
Histogramme der klassischen Features

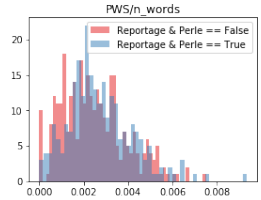


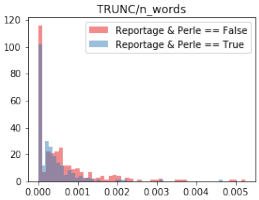
Histogramme der STTS POS-Tags

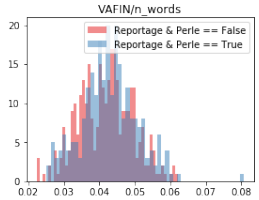


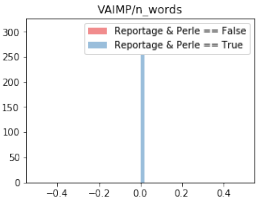

FM/n_words
ITJ/n_words
KOKOM/n_words
KON/n_words
KOUI/n_words
KOUS/n_words
NE/n_words
NNE/n_words
NN/n_words
PAV/n_words
PROAV/n_words
PDAT/n_words
PDS/n_words
PIAT/n_words
PIDAT/n_words
PIS/n_words
PPER/n_words
PPOSAT/n_words
PPOSS/n_words
PRELAT/n_words
PRELS/n_words
PRF/n_words
PTKA/n_words
PTKANT/n_words
PTKNEG/n_words
PTKVZ/n_words
PTKZU/n_words
PWAT/n_words
Reportage & Perle == False
Reportage & Perle == True

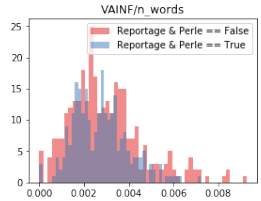


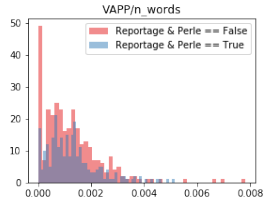


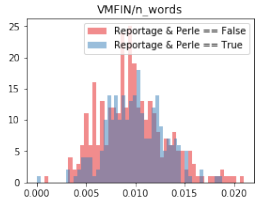


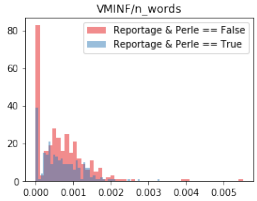


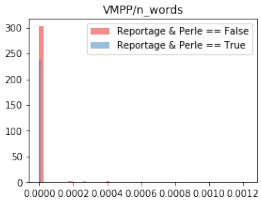


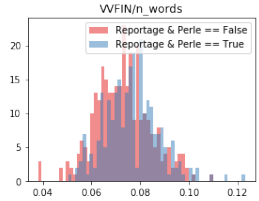


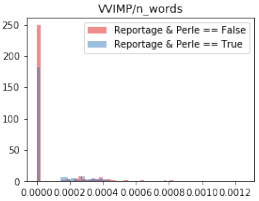


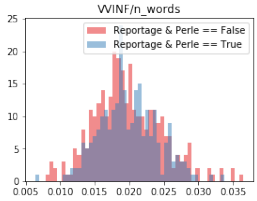


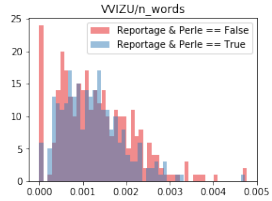


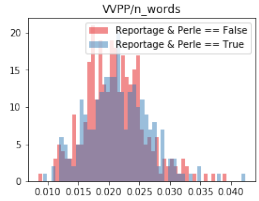


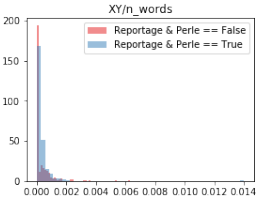


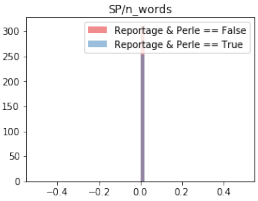




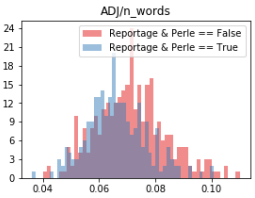


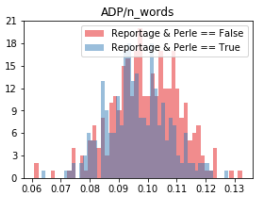


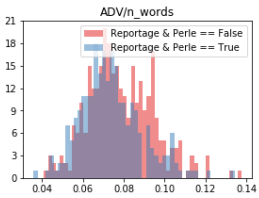


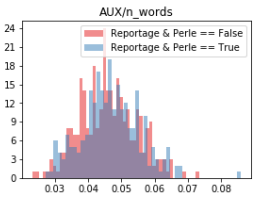


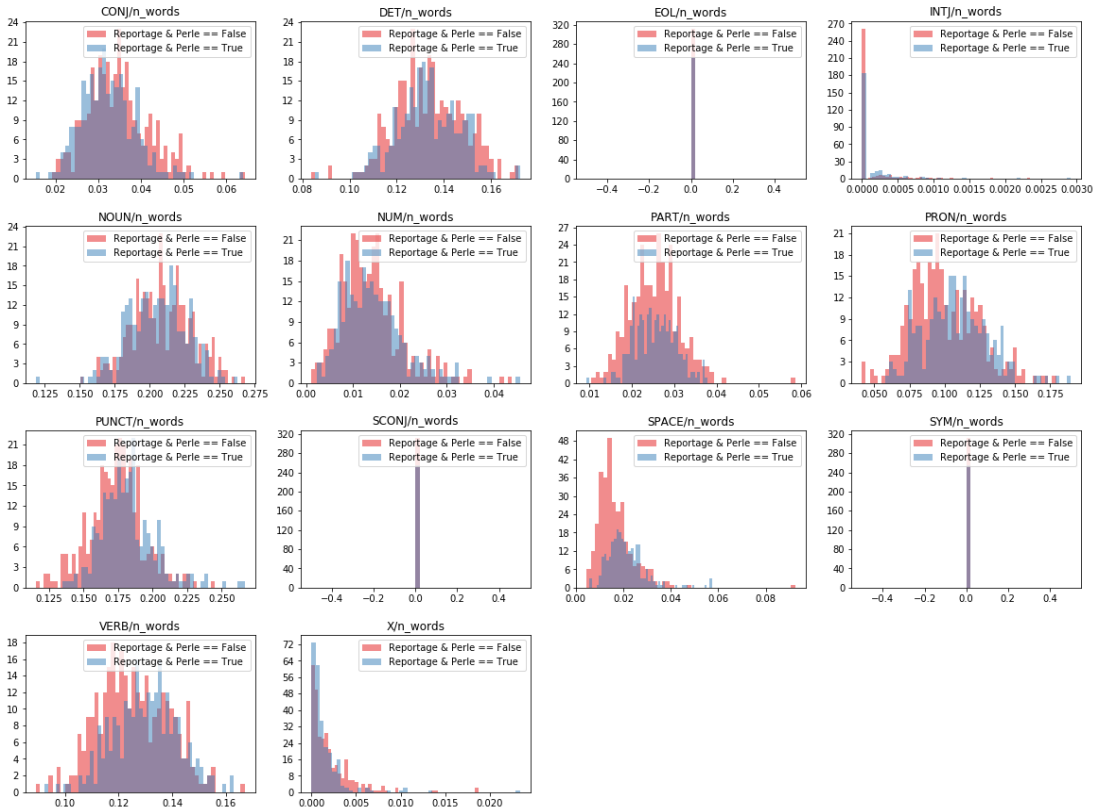
Histogramme der Universal POS-Tags







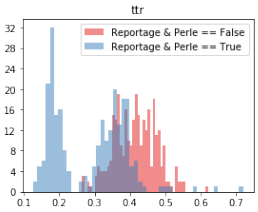


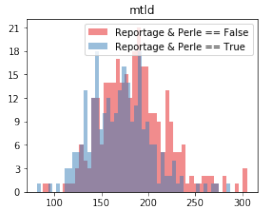


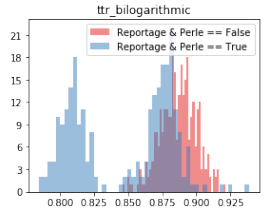
Histogramme der lexikalischen Features

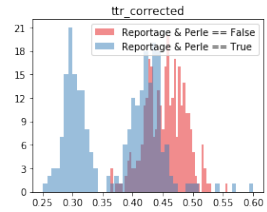
Die Anzahl verschiedener Wörter nimmt bei steigender Textlänge ab, wie in der Fachliteratur beschrieben (Perkuhn/Keibel/Kupietz 2012). Während der Klassifikation muss daher geprüft werden, ob die folgenden Feature mit bimodalen Histogramm einen negativen Einfluss auf das Ergebnis haben:

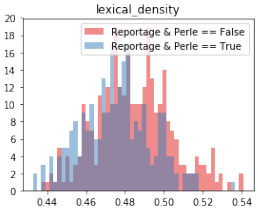
- ttr
- ttr_bilogarithmic
- ttr_corrected
- adjective_variation
- adjective_variation_lemma
- adverb_variation
- noun_variation
- noun_variation_lemma
- verb_variation
- verb_variation_lemma

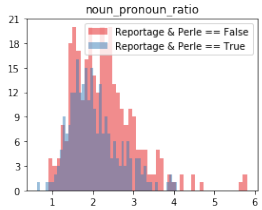

ttr
Reportage & Perle == False
Reportage & Perle == True

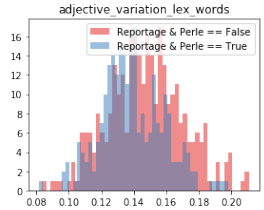

mtld
Reportage & Perle == False
Reportage & Perle == True

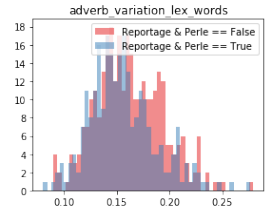

ttr_bilogarithmic
Reportage & Perle == False
Reportage & Perle == True

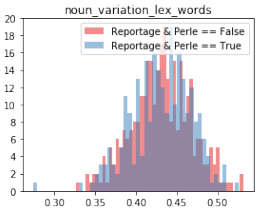

ttr_corrected
Reportage & Perle == False
Reportage & Perle == True

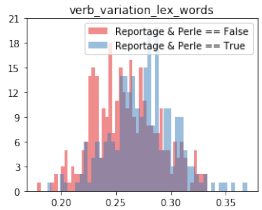

lexical_density
Reportage & Perle == False
Reportage & Perle == True

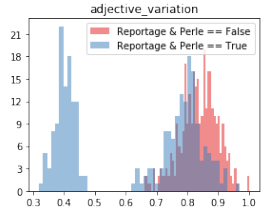

noun_pronoun_ratio
Reportage & Perle == False
Reportage & Perle == True

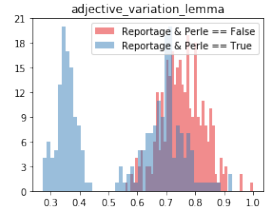

adjective_variation_lex_words
Reportage & Perle == False
Reportage & Perle == True

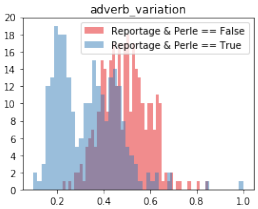

adverb_variation_lex_words
Reportage & Perle == False
Reportage & Perle == True

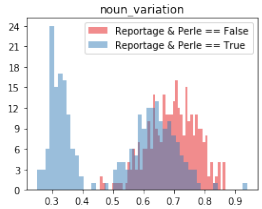

noun_variation_lex_words
Reportage & Perle == False
Reportage & Perle == True

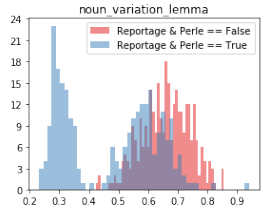

verb_variation_lex_words
Reportage & Perle == False
Reportage & Perle == True

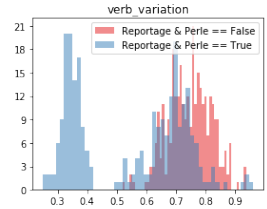

adjective_variation
Reportage & Perle == False
Reportage & Perle == True

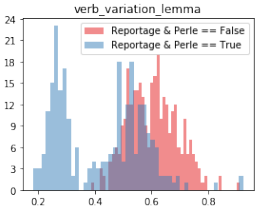

adjective_variation_lemma
Reportage & Perle == False
Reportage & Perle == True


adverb_variation
Reportage & Perle == False
Reportage & Perle == True

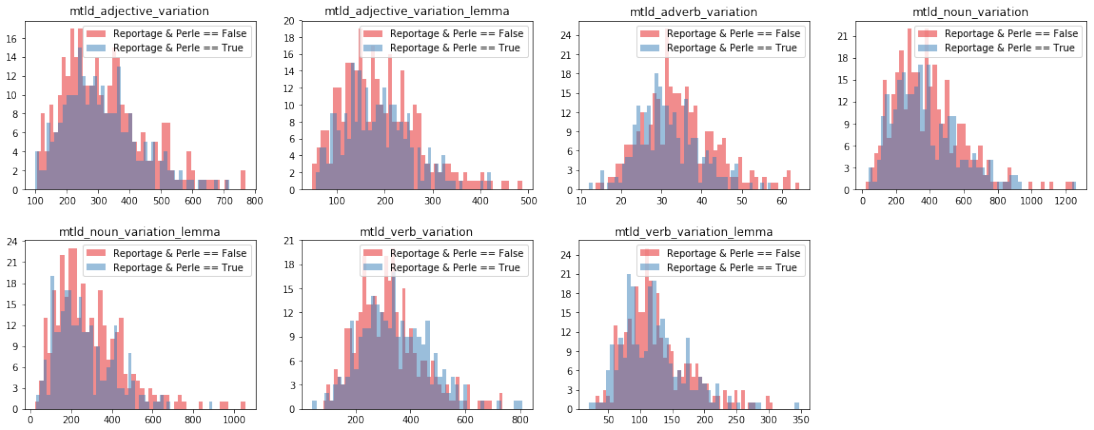

noun_variation
Reportage & Perle == False
Reportage & Perle == True


noun_variation_lemma
Reportage & Perle == False
Reportage & Perle == True


verb_variation
Reportage & Perle == False
Reportage & Perle == True

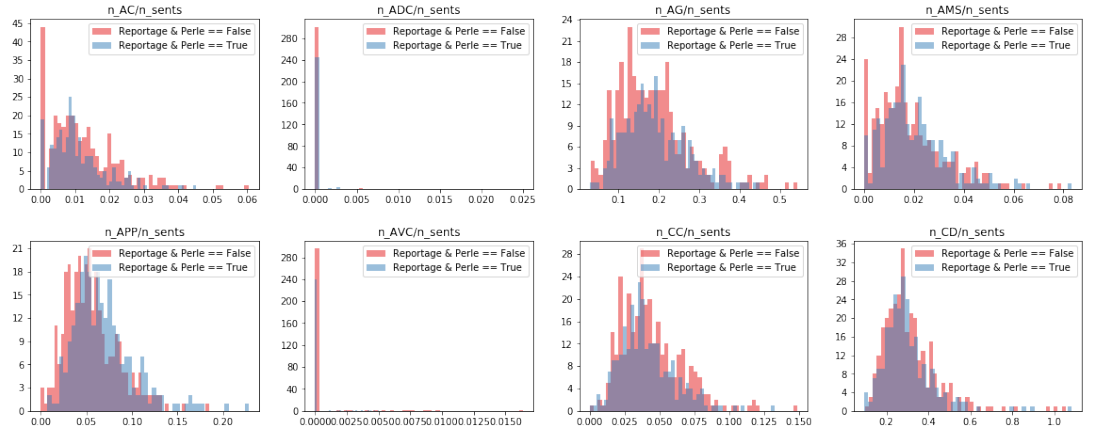

verb_variation_lemma
Reportage & Perle == False
Reportage & Perle == True

Die folgenden lexikalischen Features wurden mit Hilfe der Python-Bibliothek von Frens[94] berechnet (Frens 2017) , da die TTR-Berechnungen keine plausiblen Daten geliefert hat.

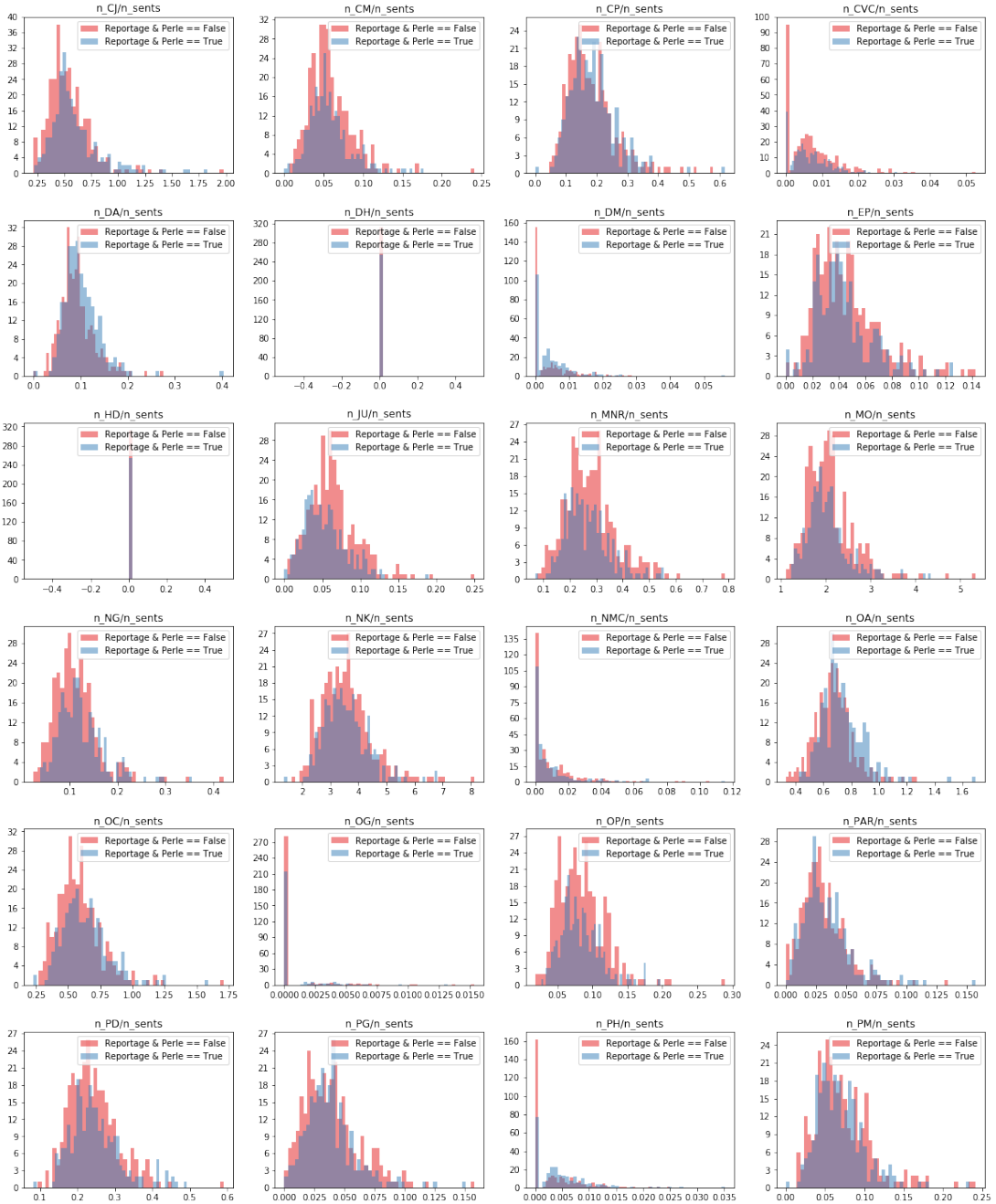


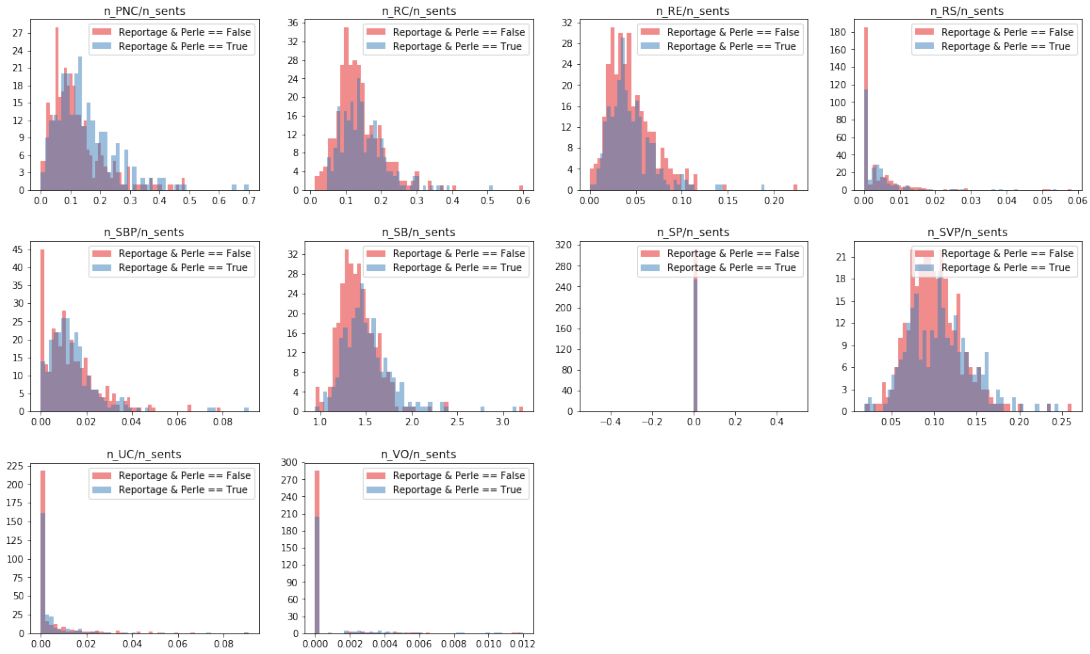
Histogramme der syntaktischen Features (spaCy)

Dieser Teil der Features wurde mit textacy prozessiert. textacy verwendet den spaCy-Tagger für das Syntactic Dependency Parsing. Eine Beschreibung der Tags und ihre Bedeutung findet sich im Anhang ab Seite 83ff.



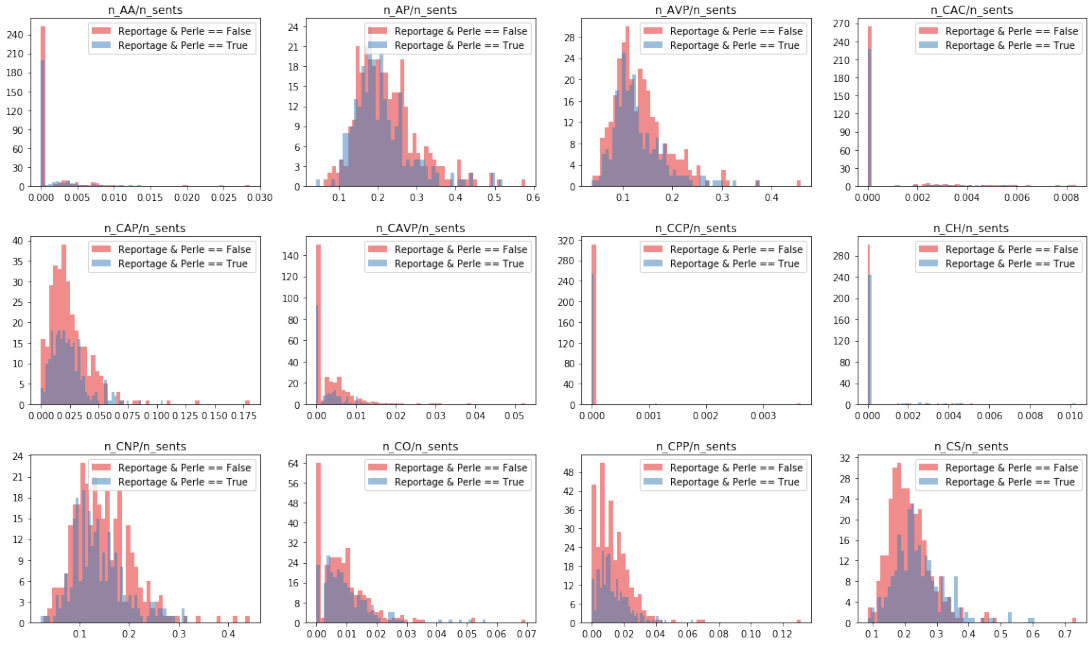
[94] https://github.com/jfrens/lexical_diversity

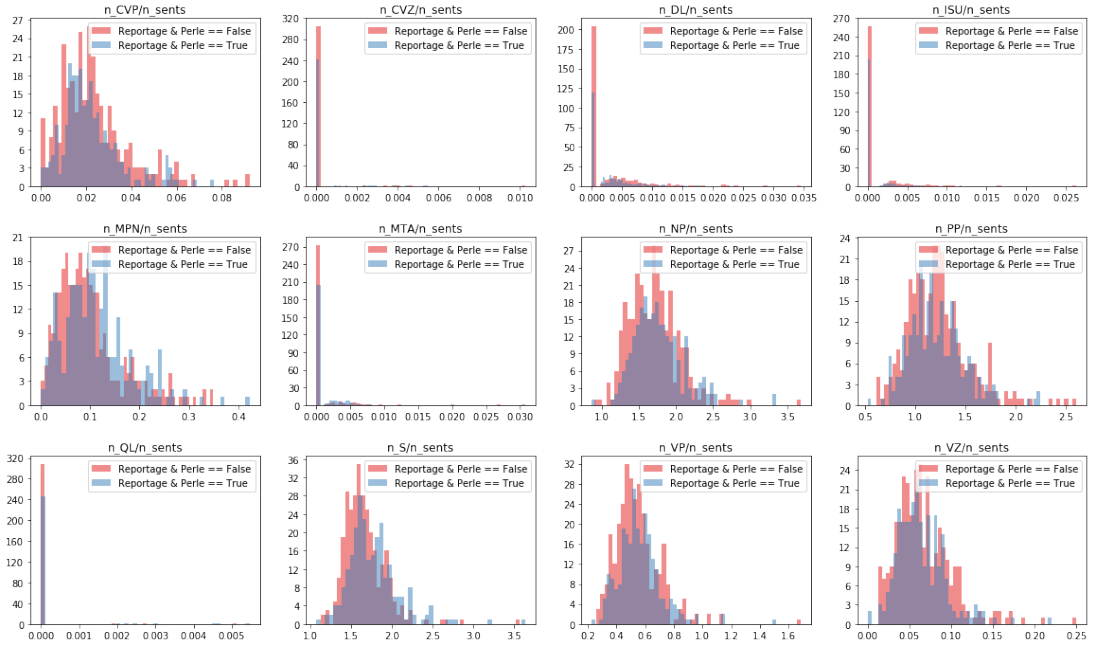

n_CJ/n_sents
n_CM/n_sents
n_CP/n_sents
n_CVC/n_sents
n_DA/n_sents
n_DH/n_sents
n_DM/n_sents
n_EP/n_sents
n_HD/n_sents
n_JU/n_sents
n_MNR/n_sents
n_MO/n_sents
n_NG/n_sents
n_NK/n_sents
n_NMC/n_sents
n_OA/n_sents
n_OC/n_sents
n_OG/n_sents
n_OP/n_sents
n_PAR/n_sents
n_PD/n_sents
n_PG/n_sents
n_PH/n_sents
n_PM/n_sents
Reportage & Perle == False
Reportage & Perle == True



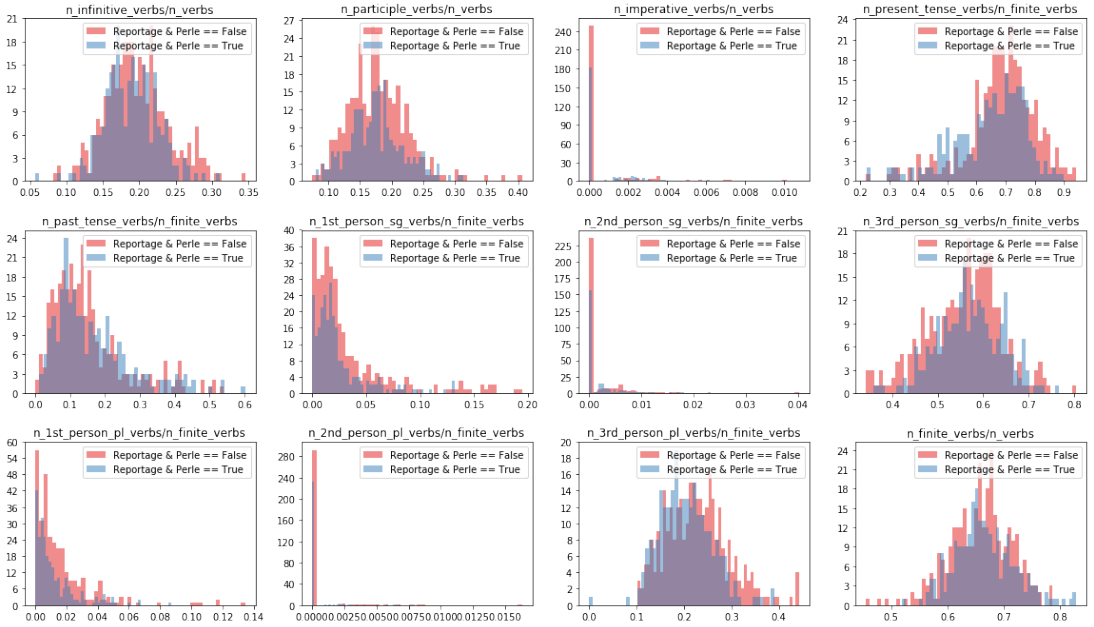
Histogramme der syntaktischen Features (Stanford)

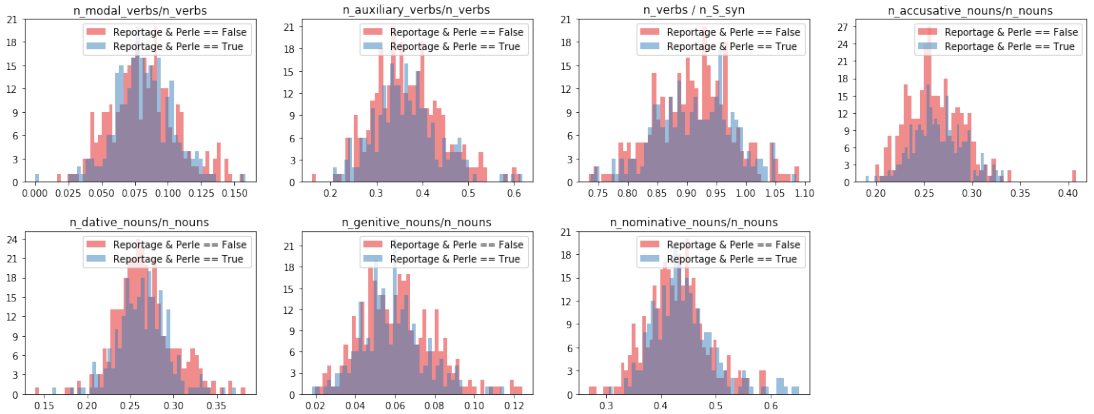
Dieser Teil der Features wurde mit dem Stanford Parsers extrahiert. Eine Beschreibung der Tags und ihre Bedeutung findet sich im Anhang ab Seite 83ff.





Histogramme der morphologischen Features





9.3 Pearson-Korrelation mit Wortanzahl

Die Tabelle zeigt die Ergebnisse der Korrelationsuntersuchung zwischen sämtlichen Features und der Wortanzahl n_words. Die nicht vorhandenen Werte für einige wenige Merkmale am Ende der Tabelle resultieren daraus, dass in den untersuchten Texten keine entsprechenden Treffer gefunden wurden.

feature	**Corr n_words**
ttr_corrected	-0,84
ttr	-0,83
adverb_variation	-0,81
verb_variation_lemma	-0,81
adjective_variation_lemma	-0,80
verb_variation	-0,80
adjective_variation	-0,78
noun_variation_lemma	-0,78
noun_variation	-0,77
ttr_bilogarithmic	-0,76
n_TRUNC / n_words	-0,19
lexical_density	-0,19
noun_pronoun_ratio	-0,18
mtld	-0,16
n_ADJ / n_words	-0,14
n_ADJA / n_words	-0,14
n_ADP / n_words	-0,14

n_APPR / n_words	-0,13
n_long_words / n_words	-0,13
n_chars / n_words	-0,12
mtld_adverb_variation	-0,12
n_3rd_person_pl_verbs / n_finite_verbs	-0,11
n_ART / n_words	-0,11
n_PROAV / n_words	-0,11
n_NOUN / n_words	-0,11
n_NN / n_words	-0,11
adjective_variation_lex_words	-0,11
n_ADV /n_words	-0,11
n_KON / n_words	-0,11
n_ADV / n_words	-0,11
n_AA / n_sents	-0,10
n_syllables / n_words	-0,10
n_CCONJ / n_words	-0,09
n_PTKA / n_words	-0,09
n_X / n_words	-0,09

n_MNR / n_sents	-0,08
n_AVC / n_sents	-0,08
n_present_tense_verbs / n_finite_verbs	-0,08
wiener_sachtextformel	-0,08
adverb_variation_lex_words	-0,07
n_AP / n_sents	-0,06
n_APPRART / n_words	-0,06
n_DET / n_words	-0,06
n_XY / n_words	-0,06
n_JU / n_sents	-0,06
n_finite_verbs / n_verbs	-0,05
n_CNP / n_sents	-0,05
n_ADJD / n_words	-0,05
noun_variation_lex_words	-0,04
n_ISU / n_sents	-0,04
n_imperative_verbs / n_verbs	-0,04
n_AC / n_sents	-0,04
n_PTKVZ / n_words	-0,03
n_ADC / n_sents	-0,03
n_PP / n_sents	-0,03
n_VVIMP / n_words	-0,03
n_genitive_nouns / n_nouns	-0,03
n_OG / n_sents	-0,03
n_PIS / n_words	-0,03
n_MO / n_sents	-0,03
n_CD / n_sents	-0,03
n_dative_nouns / n_nouns	-0,03
n_VAIMP / n_words	-0,02
n_CAC / n_sents	-0,02
n_CPP / n_sents	-0,02
n_NK / n_sents	-0,02
n_EP / n_sents	-0,02
n_AVP / n_sents	-0,02

n_PIAT / n_words	-0,02
n_1st_person_pl_verbs / n_finite_verbs	-0,02
n_KOUI / n_words	-0,01
n_FM / n_words	-0,01
n_1st_person_sg_verbs / n_finite_verbs	-0,01
n_NUM / n_words	-0,01
n_CARD / n_words	-0,01
mtld_verb_variation_lemma	-0,01
n_INTJ / n_words	-0,01
n_ITJ / n_words	-0,01
n_CAP / n_sents	-0,01
n_MTA / n_sents	-0,01
n_auxiliary_verbs / n_verbs	0,00
n_CVP / n_sents	0,00
n_2nd_person_pl_verbs / n_finite_verbs	0,00
n_CCP / n_sents	0,00
n_DM / n_sents	0,00
n_CAVP / n_sents	0,00
n_PDS / n_words	0,00
n_NMC / n_sents	0,00
n_PPOSS / n_words	0,01
n_CVZ / n_sents	0,01
n_3rd_person_sg_verbs / n_finite_verbs	0,01
n_APPO / n_words	0,01
n_UC / n_sents	0,01
n_APZR / n_words	0,01
n_VAINF / n_words	0,01
n_SBP / n_sents	0,01
n_QL / n_sents	0,02
n_DL / n_sents	0,02
n_VVIZU / n_words	0,02
n_PWAT / n_words	0,02

n_participle_verbs / n_verbs	0,02
mtld_noun_variation_lemma	0,02
n_CC / n_sents	0,02
n_MPN / n_sents	0,03
n_past_tense_verbs / n_finite_verbs	0,03
n_VMPP / n_words	0,03
n_AMS / n_sents	0,03
n_CO / n_sents	0,03
n_PRF / n_words	0,03
n_PART / n_words	0,03
n_CH / n_sents	0,04
n_SVP / n_sents	0,04
n_OP / n_sents	0,04
n_VVFIN / n_words	0,04
n_PM / n_sents	0,04
n_KOKOM / n_words	0,05
n_PTKZU / n_words	0,05
n_infinitive_verbs / n_verbs	0,05
n_RE / n_sents	0,05
n_VMINF / n_words	0,05
n_verbs / n_S	0,05
n_AG / n_sents	0,05
n_words / n_sents	0,05
n_VZ / n_sents	0,05
n_PWAV / n_words	0,05
n_PDAT / n_words	0,06
n_accusative_nouns / n_nouns	0,06
n_RS / n_sents	0,06
n_PTKNEG / n_words	0,06
n_CM / n_sents	0,06
n_VAPP / n_words	0,06
flesch_readability_ease_DE	0,06
n_PAR / n_sents	0,06

n_PG / n_sents	0,07
n_NP / n_sents	0,07
n_CVC / n_sents	0,07
monosyllable_words / n_words	0,08
n_PRELS / n_words	0,08
n_PTKANT / n_words	0,08
n_PD / n_sents	0,08
n_PRELAT / n_words	0,09
n_NG / n_sents	0,09
n_PWS / n_words	0,09
n_CP / n_sents	0,09
n_NE / n_words	0,09
mtld_adjective_variation_lemma	0,09
n_2nd_person_sg_verbs / n_finite_verbs	0,09
n_PNC / n_sents	0,09
n_VO / n_sents	0,09
n_AUX / n_words	0,09
n_VAFIN / n_words	0,09
mtld_noun_variation	0,10
n_RC / n_sents	0,10
n_KOUS / n_words	0,10
n_nominative_nouns / n_nouns	0,10
n_VVPP / n_words	0,10
n_modal_verbs / n_verbs	0,10
n_CJ / n_sents	0,10
n_PPOSAT	0,11
n_VP / n_sents	0,12
n_VVINF / n_words	0,13
n_DA / n_sents	0,13
n_VMFIN / n_words	0,14
n_CS / n_sents	0,14
n_APP / n_sents	0,15

n_S / n_sents	0,15
n_SB / n_sents	0,15
n_VERB / n_words	0,16
n_PRON / n_words	0,16
n_OC / n_sents	0,16
mtld_adjective_variation	0,17
n_PPER / n_words	0,17
n_PH / n_sents	0,17
n_OA / n_sents	0,19
verb_variation_lex_words	0,20
mtld_verb_variation	0,20

n_NNE / n_words	nan
n_PAV / n_words	nan
n_PIDAT / n_words	nan
n_SP / n_words	nan
n_EOL / n_words	nan
n_SCONJ / n_words	nan
n_SYM / n_words	nan
n_DH / n_sents	nan
n_HD / n_sents	nan
n_SP / n_sent	nan

9.4 Pearson-Korrelation mit Wort- und Satzanzahl

	Pearson correlation with n_words		**Pearson correlation with n_sents**	
feature	**corr coeff**	**p-value**	**corr coeff**	**p-value**
n_chars / n_words	-0,12	0,00	-0,17	0,00
n_words / n_sents	0,05	0,20	-0,20	0,00
n_syllables / n_words	-0,10	0,02	-0,15	0,00
n_long_words / n_words	-0,13	0,00	-0,17	0,00
n_monosyllable_words / n_words	0,08	0,07	0,12	0,00
wiener_sachtextformel	-0,08	0,07	-0,20	0,00
flesch_readability_ease_DE	0,06	0,15	0,20	0,00
n_ADJA / n_words	-0,14	0,00	-0,21	0,00
n_ADJD / n_words	-0,05	0,28	-0,05	0,28
n_ADV / n_words	-0,11	0,01	-0,10	0,01
n_APPO / n_words	0,01	0,85	0,01	0,72
n_APPR / n_words	-0,13	0,00	-0,19	0,00
n_APPRART / n_words	-0,06	0,13	-0,07	0,10
n_APZR / n_words	0,01	0,83	0,00	0,97
n_ART / n_words	-0,11	0,01	-0,13	0,00
n_CARD / n_words	-0,01	0,81	0,00	0,93

n_FM /n_words	-0,01	0,77	-0,04	0,40
n_ITJ / n_words	-0,01	0,86	0,02	0,58
n_KOKOM / n_words	0,05	0,28	0,01	0,77
n_KON / n_words	-0,11	0,01	-0,18	0,00
n_KOUI / n_words	-0,01	0,74	-0,03	0,50
n_KOUS / n_words	0,10	0,02	0,05	0,23
n_NE / n_words	0,09	0,04	0,08	0,05
n_NNE / n_words	nan	1,00	nan	1,00
n_NN / n_words	-0,11	0,01	-0,11	0,01
n_PAV / n_words	nan	1,00	nan	1,00
n_PROAV / n_words	-0,11	0,01	-0,14	0,00
n_PDAT / n_words	0,06	0,18	0,03	0,54
n_PDS / n_words	0,00	0,92	0,04	0,37
n_PIAT / n_words	-0,02	0,67	-0,02	0,57
n_PIDAT / n_words	nan	1,00	nan	1,00
n_PIS / n_words	-0,03	0,50	-0,02	0,59
n_PPER / n_words	0,17	0,00	0,24	0,00
n_PPOSAT / n_words	0,11	0,01	0,14	0,00
n_PPOSS / n_words	0,01	0,90	0,02	0,63
n_PRELAT / n_words	0,09	0,04	0,02	0,65
n_PRELS / n_words	0,08	0,07	-0,04	0,30
n_PRF / n_words	0,03	0,46	0,02	0,64
n_PTKA / n_words	-0,09	0,03	-0,10	0,02
n_PTKANT / n_words	0,08	0,07	0,13	0,00
n_PTKNEG / n_words	0,06	0,16	0,08	0,05
n_PTKVZ / n_words	-0,03	0,43	0,04	0,39
n_PTKZU / n_words	0,05	0,26	0,01	0,73
n_PWAT / n_words	0,02	0,64	0,02	0,71
n_PWAV / n_words	0,05	0,20	0,07	0,11
n_PWS / n_words	0,09	0,04	0,12	0,01
n_TRUNC / n_words	-0,19	0,00	-0,19	0,00
n_VAFIN / n_words	0,09	0,03	0,15	0,00
n_VAIMP / n_words	-0,02	0,57	-0,02	0,63
n_VAINF / n_words	0,01	0,79	-0,01	0,77

n_VAPP / n_words	0,06	0,15	0,02	0,57
n_VMFIN / n_words	0,14	0,00	0,17	0,00
n_VMINF / n_words	0,05	0,23	0,03	0,41
n_VMPP / n_words	0,03	0,51	0,02	0,61
n_VVFIN / n_words	0,04	0,30	0,11	0,01
n_VVIMP / n_words	-0,03	0,48	0,00	0,98
n_VVINF / n_words	0,13	0,00	0,13	0,00
n_VVIZU / n_words	0,02	0,65	0,00	0,95
n_VVPP / n_words	0,10	0,02	0,09	0,04
n_XY / n_words	-0,06	0,17	-0,03	0,44
n_SP / n_words	nan	1,00	nan	1,00
n_ADJ / n_words	-0,14	0,00	-0,20	0,00
n_ADP / n_words	-0,14	0,00	-0,20	0,00
n_ADV / n_words	-0,11	0,01	-0,11	0,01
n_AUX / n_words	0,09	0,03	0,13	0,00
n_CCONJ / n_words	-0,09	0,03	-0,17	0,00
n_DET / n_words	-0,06	0,15	-0,08	0,06
n_EOL / n_words	nan	1,00	nan	1,00
n_INTJ / n_words	-0,01	0,86	0,02	0,58
n_NOUN / n_words	-0,11	0,01	-0,11	0,01
n_NUM / n_words	-0,01	0,81	0,00	0,93
n_PART / n_words	0,03	0,43	0,08	0,06
n_PRON / n_words	0,16	0,00	0,21	0,00
n_SCONJ / n_words	nan	1,00	nan	1,00
n_SYM / n_words	nan	1,00	nan	1,00
n_VERB / n_words	0,16	0,00	0,22	0,00
n_X / n_words	-0,09	0,04	-0,10	0,02
mtld	-0,16	0,00	-0,18	0,00
lexical_density	-0,19	0,00	-0,18	0,00
noun_variation_lex_words	-0,04	0,32	-0,04	0,30
adjective_variation_lex_words	-0,11	0,01	-0,17	0,00
adverb_variation_lex_words	-0,07	0,10	-0,07	0,10
verb_variation_lex_words	0,20	0,00	0,25	0,00
noun_pronoun_ratio	-0,18	0,00	-0,21	0,00

mtld_adjective_variation	0,19	0,00	0,15	0,00
mtld_adjective_variation_lemma	0,11	0,01	0,07	0,09
mtld_adverb_variation	-0,10	0,01	-0,12	0,00
mtld_noun_variation	0,10	0,02	0,06	0,15
mtld_noun_variation_lemma	0,03	0,51	-0,01	0,87
mtld_verb_variation	0,21	0,00	0,17	0,00
mtld_verb_variation_lemma	0,00	0,95	-0,03	0,51
n_AA / n_sents	-0,10	0,01	-0,12	0,00
n_AP / n_sents	-0,06	0,13	-0,24	0,00
n_AVP / n_sents	-0,02	0,64	-0,14	0,00
n_CAC / n_sents	-0,02	0,58	-0,04	0,39
n_CAP / n_sents	-0,01	0,86	-0,11	0,01
n_CAVP / n_sents	0,00	0,92	-0,04	0,30
n_CCP / n_sents	0,00	0,95	0,02	0,66
n_CH / n_sents	0,04	0,40	0,05	0,28
n_CNP / n_sents	-0,05	0,27	-0,19	0,00
n_CO / n_sents	0,03	0,48	-0,07	0,10
n_CPP / n_sents	-0,02	0,62	-0,12	0,00
n_CS / n_sents	0,14	0,00	-0,05	0,21
n_CVP / n_sents	0,00	0,98	-0,12	0,00
n_CVZ / n_sents	0,01	0,87	0,00	0,96
n_DL / n_sents	0,02	0,70	0,03	0,48
n_ISU / n_sents	-0,04	0,32	-0,06	0,13
n_MPN / n_sents	0,03	0,54	-0,05	0,22
n_MTA / n_sents	-0,01	0,89	-0,03	0,46
n_NP / n_sents	0,07	0,11	-0,16	0,00
n_PP / n_sents	-0,03	0,48	-0,25	0,00
n_QL / n_sents	0,02	0,72	0,02	0,58
n_S / n_sents	0,15	0,00	-0,08	0,06
n_VP / n_sents	0,12	0,00	-0,07	0,12
n_VZ / n_sents	0,05	0,20	-0,08	0,06
n_AC / n_sents	-0,04	0,39	-0,12	0,01
n_ADC / n_sents	-0,03	0,47	-0,05	0,25
n_AG / n_sents	0,05	0,23	-0,10	0,02

n_AMS / n_sents	0,03	0,51	-0,03	0,42
n_APP / n_sents	0,15	0,00	0,03	0,52
n_AVC / n_sents	-0,08	0,06	-0,08	0,05
n_CC / n_sents	0,02	0,58	-0,11	0,01
n_CD / n_sents	-0,03	0,51	-0,22	0,00
n_CJ / n_sents	0,10	0,02	-0,10	0,01
n_CM / n_sents	0,06	0,15	-0,08	0,06
n_CP / n_sents	0,09	0,04	-0,08	0,07
n_CVC / n_sents	0,07	0,08	0,00	0,94
n_DA / n_sents	0,13	0,00	0,01	0,76
n_DH / n_sents	nan	1,00	nan	1,00
n_DM / n_sents	0,00	0,94	0,02	0,68
n_EP / n_sents	-0,02	0,63	-0,14	0,00
n_HD / n_sents	nan	1,00	nan	1,00
n_JU / n_sents	-0,06	0,18	-0,09	0,04
n_MNR / n_sents	-0,08	0,05	-0,25	0,00
n_MO / n_sents	-0,03	0,50	-0,24	0,00
n_NG / n_sents	0,09	0,04	-0,03	0,45
n_NK / n_sents	-0,02	0,62	-0,24	0,00
n_NMC / n_sents	0,00	0,91	-0,04	0,31
n_OA / n_sents	0,19	0,00	-0,02	0,57
n_OC / n_sents	0,16	0,00	-0,01	0,73
n_OG / n_sents	-0,03	0,50	-0,05	0,21
n_OP / n_sents	0,04	0,34	-0,10	0,02
n_PAR / n_sents	0,06	0,13	-0,02	0,64
n_PD / n_sents	0,08	0,05	-0,06	0,19
n_PG./ n_sents	0,07	0,12	-0,05	0,21
n_PH / n_sents	0,17	0,00	0,09	0,03
n_PM / n_sents	0,04	0,29	-0,09	0,04
n_PNC / n_sents	0,09	0,03	0,00	0,96
n_RC / n_sents	0,10	0,02	-0,11	0,01
n_RE / n_sents	0,05	0,24	-0,08	0,05
n_RS / n_sents	0,06	0,17	0,10	0,02
n_SBP / n_sents	0,01	0,76	-0,09	0,03

n_SB / n_sents	0,15	0,00	-0,07	0,09
n_SP / n_sents	nan	1,00	nan	1,00
n_SVP / n_sents	0,04	0,37	-0,03	0,41
n_UC / n_sents	0,01	0,83	-0,03	0,46
n_VO / n_sents	0,09	0,03	0,10	0,01
n_infinitive_verbs / n_verbs	0,05	0,25	0,01	0,84
n_participle_verbs / n_verbs	0,02	0,64	-0,02	0,60
n_imperative_verbs / n_verbs	-0,04	0,33	-0,01	0,73
n_present_tense_verbs / n_finite_verbs	-0,08	0,06	-0,06	0,14
n_past_tense_verbs / n_finite_verbs	0,03	0,52	0,00	0,98
n_1st_person_sg_verbs / n_finite_verbs	-0,01	0,80	0,05	0,22
n_2nd_person_sg_verbs / n_finite_verbs	0,09	0,04	0,12	0,00
n_3rd_person_sg_verbs / n_finite_verbs	0,01	0,86	0,00	0,98
n_1st_person_pl_verbs / n_finite_verbs	-0,02	0,72	0,00	0,96
n_2nd_person_pl_verbs / n_finite_verbs	0,00	0,97	0,00	0,97
n_3rd_person_pl_verbs / n_finite_verbs	-0,11	0,01	-0,16	0,00
n_finite_verb s/n_verbs	-0,05	0,20	0,00	0,94
n_modal_verbs / n_verbs	0,10	0,02	0,11	0,01
n_auxiliary_verbs / n_verbs	0,00	0,91	0,00	0,98
n_verbs / n_S	0,05	0,23	-0,02	0,69
n_accusative_nouns / n_nouns	0,06	0,18	0,03	0,43
n_dative_nouns / n_nouns	-0,03	0,52	-0,09	0,03
n_genitive_nouns / n_nouns	-0,03	0,49	-0,10	0,01
n_nominative_nouns / n_nouns	0,10	0,02	0,18	0,00

9.5 scikit-learn Classifier-Konfiguration

Die nachfolgende Tabelle führt die verwendeten scikit-learn Classifier-Einstellungen auf, die für alle Klassifikationsdurchläufe verwendet wurde.

Algorithmus	**scikit-Konfiguration**
Naives Bayes	GaussianNB(priors=None)
SVM	SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma=0.001, kernel='linear', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)
Logistische Regression	LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=1000, multi_class='ovr', n_jobs=1, penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0, warm_start=False)
Decision Tree	DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=25, max_features=0.2, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=20, min_samples_split=20, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best')
Random Forest	RandomForestClassifier(bootstrap=True, class_weight=None, criterion='entropy', max_depth=None, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=1000, n_jobs=1, oob_score=False, random_state=None, verbose=0, warm_start=False)
Gradient Boosting	GradientBoostingClassifier(criterion='friedman_mse', init=None, learning_rate=1.0, loss='deviance', max_depth=3, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None,min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=1000, presort='auto', random_state=0, subsample=1.0, verbose=0, warm_start=False)

Versicherung über Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, den _____ _____