

Themenspezifische Gruppierung deutscher Online-Zeitungen mit Natural Language Processing

Bachelorarbeit

Georg Donner
Matrikel-Nummer 553821

Betreuer	Prof. Dr. Gefei Zhang
Erstprüfer	Prof. Dr. Gefei Zhang
Zweitprüfer	Prof. Dr. Barne Kleinen

Inhaltsverzeichnis

Abbildungsverzeichnis	IV
1 Einleitung	1
2 Grundlagen	2
2.1 Natural Language Processing	2
2.1.1 Pipeline	3
2.2 Machine Learning	5
2.2.1 Feature Engineering	5
2.2.2 Dimensionsionalitätsreduktion	6
2.2.3 Klassifizierung	6
3 Textverarbeitung	7
3.1 Verwendete Tools	7
3.1.1 Python	7
3.1.2 SpaCy	7
3.1.3 Scikit-learn	7
3.2 Datenselektion	8
3.2.1 Datensatz	8
3.2.2 Normalisierung	8
3.3 Textaufbereitung	8
3.4 Featuregenerierung	9
3.4.1 Token-basiert	9
3.4.2 Lexikalisch	9
3.4.3 Morphologisch	9
3.4.4 Syntaktisch	10

Inhaltsverzeichnis

4	Datenauswertung	11
4.1	Überblick	11
4.2	Klassifizierung	11
4.2.1	Vorbereitung des Datensatzes	11
4.2.2	Feature Selection	12
4.2.3	Messung der Performance	12
4.2.4	Verwendete Verfahren	12
4.3	Clustering	13
4.3.1	Feature Extraction	13
5	Ergebnis	15
	Literaturverzeichnis	16

Abbildungsverzeichnis

2.1	Dependency Relations	4
-----	--------------------------------	---

1 Einleitung

Natural Language Processing ist ein großes Feld, welches besonders in der letzten Zeit im Zuge der Digitalisierung viel an Aufmerksamkeit und Wichtigkeit gewonnen hat. Es ermöglicht uns Informationen schneller zu finden, Systeme durch gesprochene Sprache zu steuern oder ganze Texte zu generieren. Eine weitere Aufgabe ist es, eine große Menge an Texten in Kategorien einzuteilen, um die Daten auf eine gewünschte Teilmenge für eine spezifischere Suche oder Analyse zu reduzieren. Die Kategorisierung der Dokumente nach ihrem Inhalt ist hier der häufigste Anwendungsfall, es ist aber auch möglich Texte nach ihrem generellen Genre oder Schreibstil zu vergleichen.

Diese Arbeit wird am Beispiel deutscher Online-Zeitungen untersuchen, welche Möglichkeiten es gibt Texte unabhängig von ihrem Inhalt zu vergleichen. Dabei werden lexikalische, morphologische und syntaktische Merkmale, aber auch die Verwendung inhaltlich irrelevanter Wörter als Features verwendet. Es wird überprüft, inwiefern die Artikel gruppiert werden können und Rückschlüsse auf Unterschiede im Schreibstil ganzer Zeitungen statt nur einzelner Artikel zulassen.

Des Weiteren wird untersucht, ob und wie sich der Schreibstil einer Zeitung je nach Thema, wie z.B. Politik und Sport, unterscheidet.

2 Grundlagen

Für die Verarbeitung und Analyse von Zeitungsartikeln sind zwei Teilgebiete der Informatik besonders wichtig: Natural Language Processing und Machine Learning. Im Folgenden werden die für die Arbeit relevantesten Konzepte der beiden Bereiche genauer erklärt.

2.1 Natural Language Processing

Natural Language Processing ist ein Teilgebiet der Informatik, das Konzepte und Techniken der künstlichen Intelligenz und des Machine Learning verwendet, um natürliche Sprache zu verarbeiten. Der Begriff natürliche Sprache wird verwendet, um menschliche Sprachen zu beschreiben, die im Gegensatz zu künstlich entwickelten Plansprachen eine historische Entwicklung durchlebt haben. Diese Sprachen befinden sich in einem dauerhaften Entwicklungsprozess und sind häufig sehr variabel in ihrer Verwendung durch den Menschen. Die Analyse der Semantik eines Wortes oder Satzes ist für Computer besonders schwierig, da sich die Bedeutung häufig erst durch den Kontext ergibt.

Mit den schnellen Fortschritten im Bereich des Machine Learning in den letzten Jahrzehnten, eröffneten sich für die Verarbeitung natürlicher Sprache jedoch völlig neue Möglichkeiten. Die Erkennung von Syntax und Semantik wurde damit immer präziser und das Teilgebiet immer relevanter. Gegenwärtig basiert dies hauptsächlich auf Algorithmen des Supervised Learning, für die die Texte vorher manuell mit relevanten Markierungen versehen werden müssen. Ein bekanntes Beispiel für einen Korpus deutscher Sprache mit solchen Annotationen ist der TIGER Corpus ¹.

¹ <http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html>

TODO: Verwendung nicht annotierter Korpora

2.1.1 Pipeline

Bei der Analyse eines Textes werden in der Regel verschiedene Schritte abgearbeitet, die jeweils eigene Merkmale der Sprache untersuchen. Es entsteht eine sogenannte Pipeline, die je nach Anwendungsfall unterschiedlich aussieht. Das sequenzielle Ausführen dieser einzelnen Vorgänge ist notwendig, da beispielsweise die Analyse der Syntax voraussetzt, dass das Dokument bereits in Token zerlegt wurde. Im Folgenden werden die für diese Arbeit relevanten Schritte beschrieben.

TODO: eventuell noch andere Modelle beschreiben?

Tokenisierung

Tokenisierung beschreibt den Prozess, einen gegebenen Text in gleiche Einheiten, sogenannte Token, zu zerlegen. Meistens handelt es sich dabei um Wörter und Satzzeichen, je nach Anwendungsfall können es aber auch Wortgruppen oder Sätze sein. Die Tokenisierung ist häufig eine Voraussetzung einer tiefgehenderen Analyse des Textes, daher ist eine hohe Genauigkeit hier besonders wichtig. Eine Teilung an jedem Satzzeichen um eine Liste von Sätzen zu erhalten, ist zum Beispiel eine triviale Lösung, die bereits gute Ergebnisse erzielt. Es müssen jedoch viele Sonderregeln wie Abkürzungen und Zahlen beachtet werden, sodass das Problem wesentlich komplexer ist, als es zunächst scheint. Ein Ansatz um höhere Genauigkeit zu erreichen, ist ein Modell auf Basis bereits mit Annotationen versehener Korpora zu trainieren, welches die Regeln automatisch erlernt.

Part-of-speech-Tagging

Das Part-of-speech-Tagging, auch POS-Tagging, ist ein Verfahren, bei dem jedem Wort oder Satzzeichen die jeweilige Wortart zugeordnet wird. Bei der Analyse ist hierbei vor allem der Kontext in dem das Wort erscheint wichtig, da sich daraus häufig erst die Bedeutung ergibt. Die Informationen über die Wortart und oft auch weitere Details, geben sogenannte

2 Grundlagen

Tags, die meist aus einem festen Tagset stammen. In dieser Arbeit werden die Tags aus dem Stuttgart-Tübingen-Tagset (STTS) ² verwendet. Der Satz „Martin findet eine grüne Blechdose.“ sieht nach dem POS-Tagging beispielsweise so aus:

Martin/NE findet/VVFIN eine/ART grüne/ADJA Blechdose/NN ./.\$.

Die Tags geben Auskunft über mehr als nur die Wortart. Zum Beispiel sind die beiden Wörter *Martin* und *Blechdose* jeweils Nomen. Da jedoch beim POS-Tagging auch die Definition des Wortes überprüft wird, kann *Martin* korrekt als Eigennamen mit dem Tag NE identifiziert werden. Weiterhin wurde in diesem Satz die Verbform und der Adjektiv-Typ korrekt erkannt.

Dependency Parsing

Die Analyse der syntaktischen Struktur eines Satzes ist ein weiterer wichtiger Schritt in der Verarbeitung natürlicher Sprache. Beim Dependency Parsing wird zunächst jeder Satz nur auf seine Wörter reduziert, um dann die Beziehungen, sogenannte Dependency Relations, der Wörter innerhalb dieses Satzes zu bestimmen.

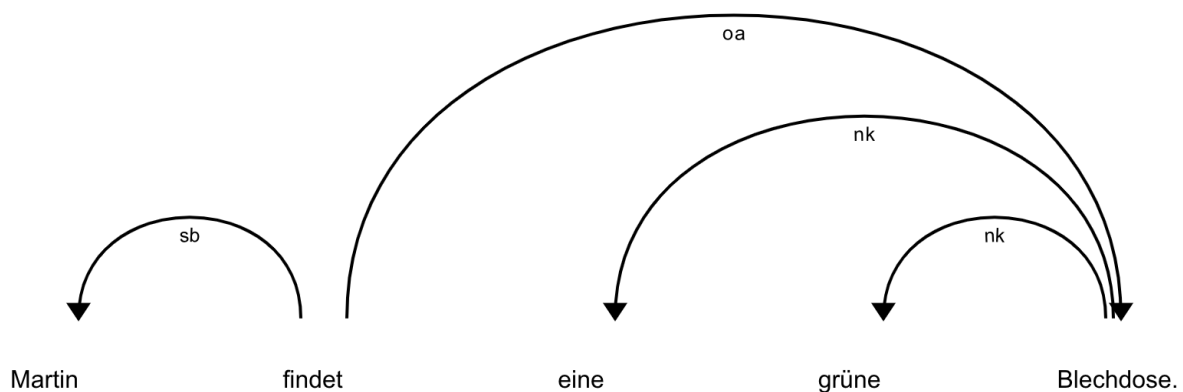


Abbildung 2.1: Visualisierung der Beziehungen mit Pfeilen

Diese Beziehungen ergeben letztendlich einen Baum, der navigiert werden kann und auch Aufschlüsse über die Komplexität eines Satzes zulässt. Die Ergebnisse des Dependency Parsing finden neben der Erkennung semantischer Beziehungen zwischen Wörtern auch noch

² <http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html>

2 Grundlagen

weitere Anwendungen. So werden sie z.B. von dem Natural Language Processing Tool SpaCy für die Erkennung der Satzenden verwendet [Exp19].

Lemmatisierung

Für viele Analysen eines Textes ist es von Vorteil oder sogar notwendig, dass nicht jedes Wort welches unterschiedlich geschrieben wird, als ein anderes Wort behandelt wird. Um dies zu erreichen, werden alle Wörter auf ihre Grundform reduziert. Dieser Prozess heißt Lemmatisierung. Dies ist besonders hilfreich für die Feststellung der Häufigkeit eines Wortes in einem Dokument oder Korpus. So wäre die Frequenz der Wörter *findet*, *fand* und *finden* zunächst jeweils eins. Nach der Lemmatisierung gibt es nur noch das Lemma *finden* mit einer Frequenz von drei. Dies hilft dabei, das Rauschen innerhalb eines Textes zu reduzieren und das tatsächliche Vokabular genauer zu beurteilen.

2.2 Machine Learning

2.2.1 Feature Engineering

Alle Beobachtungen oder Erfahrungen aus welchen ein zu trainierendes Modell lernt, haben immer eine bestimmte Anzahl an Features, die jede Beobachtung so gut wie möglich repräsentieren sollen. Die Auswahl dieser Features bildet die Grundlage des Trainings und ist ausschlaggebend für das letztendliche Resultat. Die Verwendung voneinander unabhängiger Features, die mit der vorherzusagenden Klasse in Beziehung stehen, wird dabei höchstwahrscheinlich zu sehr guten Resultaten führen [Dom12]. Die Auswahl der richtigen Features ist also der Schlüssel für das erfolgreiche Lernen, aber es ist auch der schwierigste Teil. „Coming up with features is difficult, time-consuming, requires expert knowledge.“ [Ng13] Häufig werden viele Features generiert bzw. bestimmt, die möglicherweise nicht alle zielführend oder unabhängig sind. Hier gibt es verschiedene Prozesse, um die Anzahl an Features zu reduzieren und gleichzeitig kaum an Informationsgehalt zu verlieren. Die Auswahl von Features aus der großen Menge heißt *Feature Selection*. Die Kombination von Features, um daraus neue zu generieren, wird *Feature Extraction* genannt. Diese Reduktion ist oftmals erforderlich um das statistische Rauschen in den Daten

2 Grundlagen

zu minimieren und beim Training Zeit zu sparen. Besonders bei sehr rechenintensiven Algorithmen ist dies absolut notwendig.

2.2.2 Dimensionsionalitätsreduktion

Bei der Generierung von d verschiedenen Features bzw. Variablen für n Beobachtungen, entsteht ein Raum der Dimension d . Um so viele Informationen wie möglich zu jeder Beobachtung zu speichern und für z.B. eine Klassifizierung zu verwenden, werden sehr viele Features generiert, welche wiederum einen hoch-dimensionalen Raum ergeben. Dabei entsteht das Problem, dass die Datenverarbeitung immer mehr Zeit beansprucht, da auch wesentlich mehr Beobachtungen notwendig sind. Weiterhin führt es zu Schwierigkeiten bei der Bestimmung von Distanzen zwischen einzelnen Datenpunkten. In solch einem Raum ist es unvermeidbar, dass die Verteilung der Punkte sehr dünn ist. Darauf basierend wurde festgestellt, dass in einem hoch-dimensionalen Raum alle Punkte annähernd die gleiche Entfernung haben und die Suche des nächsten Nachbarn nur sehr ungenau ist [HAK00]. Eine häufige Annahme bei der Betrachtung dieser Räume ist, dass die Variablen nicht alle voneinander unabhängig sind und die Anzahl an Dimensionen reduziert werden kann, ohne dabei einen signifikanten Anteil an Informationen zu verlieren. Ein häufig verwendetes Verfahren ist die *Hauptkomponentenanalyse* oder englisch *Principal Component Analysis (PCA)*. Hierbei werden iterativ Linearkombinationen der Variablen erstellt, die durch Maximierung der Varianz der Daten in die entsprechende Richtung bestimmt werden und so jeweils die höchste Aussagekraft besitzen. Bei einer Reduktion auf zwei oder drei Dimensionen können die Daten dann gut visualisiert werden, dabei gehen aber meistens zu viele Informationen verloren.

2.2.3 Klassifizierung

Was ist Klassifizierung? Zuordnung einer neuen Beobachtung zu einer Kategorie aus einem vordefinierten Set. Wurde vorher auf Basis eines Trainingssets trainiert, bei denen die Kategorie bekannt ist (supervised learning) (aus Wikipedia en). Wie funktioniert Klassifizierung grob (One vs All, Multinomial...). Welche Klassifizierungsalgorithmen gibt es? Auf jeden Fall keinen besten.

3 Textverarbeitung

Welche Schritte sind erforderlich, was muss besonders berücksichtigt werden?

3.1 Verwendete Tools

3.1.1 Python

Python ist die am häufigsten für Machine Learning verwendete Sprache und es gibt eine Vielzahl an packages die für nlp/ml optimiert sind. Welche packages wurden für diese Arbeit primär genutzt und warum so gut? Zwei sehr wichtige im Detail, sonst noch Numpy, Pandas und Matplotlib? Anaconda?

3.1.2 SpaCy

Was kann Spacy alles? Warum für Spacy entschieden? (ist für die deutsche Sprache schon sehr ausgereift). Vergleich zu anderen NLP tools und Überblick über die tatsächliche Schnelligkeit auf meiner Maschine beim processen?

3.1.3 Scikit-learn

Was kann sklearn alles? Warum für sklearn entschieden? Wie funktioniert es genau? Vergleiche zu anderen ML tools?

3.2 Datenselektion

Die Auswahl der Daten ist Grundlage für alles, blabla

3.2.1 Datensatz

Woher kommt der Datensatz? Wie groß? Welche Zeitungen? Überblick der Zeitspanne. Illustration wie verschieden die Artikel sind?

3.2.2 Normalisierung

Welche Normalisierungen mussten vor der Aufbereitung durchgeführt werden: Herausfiltern von viel zu kurzen Artikeln/Artikeln ohne Inhalt, Zuordnung von Kategorien zu jedem Artikel, Vereinheitlichung des Formats des Datums (ISO). Oft veröffentlichen Zeitungen auch Artikel einer Newsagentur wie z.B. der dpa, diese müssen herausgefiltert werden, um den Schreibstil einer Zeitung ermitteln zu können (nur Artikel von Autoren der Zeitung sollen berücksichtigt werden). Illustration zum Anteil der dpa Artikel.

3.3 Textaufbereitung

Analyse der Texten mit einem Natural Language Processing Tool, das die zuvor beschriebene Pipeline durchläuft (SpaCy). Vorherige Ansätze?

Nachbereitung der Texte: Entfernen von vermeintlichen Sätzen mit weniger als 4 Wörtern. Musste dort sehr rigoros sein, da jeglicher Inhalt mit enthalten war, so auch Kürzel wie (dpa), der Ort, Autor, Quellen oder Verweise.

3.4 Featuregenerierung

Es gibt eine Vielzahl an Features, die man für einen Text generieren bzw. auswählen kann, ohne dass dabei der Inhalt des Artikels einen Einfluss hat. Verweis zu Studien, die analysiert haben, welche Kombination an Features die besten Resultate liefern.

3.4.1 Token-basiert

Durchschnittliche Satzlänge, durchschnittliche Wortlänge?

Frequenz der n häufigsten Wörter (bei mir 30, 50, 100). Sind dann Wörter wie ['der', 'ich', 'dieser']

3.4.2 Lexikalisch

Type-token ratio: Beurteilt die Reichhaltigkeit des Vokabulars, dafür gibt es noch mehr präzisere Indizes (in der Studie von 2000 sind einige Beispiele). Nachteil: funktioniert besser, je länger die Texte sind und die Artikel sind durchschnittlich nicht besonders lang. Illustration dazu?

Readability: Wieder gibt es hier viele verschiedene Indizes. In dieser Arbeit verwendet wurde der Flesch-Reading Ease Index unter Berücksichtigung dass die Sprache Deutsch ist (gibt noch viel mehr für Englisch). Erforderte die Berechnung, wieviele Silben ein Wort hat (war in der Pipeline nicht mit drin)

3.4.3 Morphologisch

POS unigrams: Die Frequenz der einzelnen Wortarten für jeden Artikel, ist teilweise nur sehr gering und auch hier erhöht sich die Aussagekraft mit der Länge des Artikels.

3 Textverarbeitung

Lexical Density: Wieviele Wörter des Artikels tragen zum Inhalt bei (Nomen, Verben, Adjektive, Adverbien) im Verhältnis zur Gesamtanzahl an Wörtern? Ähnlich zur Stopword Frequency, die auch berechnet wurde, aber zu sehr mit der lexikalischen Dichte korreliert.

3.4.4 Syntaktisch

Dependency relation unigrams: Die Frequenz der einzelnen Dependency relations für jeden Artikel, ist teilweise nur sehr gering und auch hier erhöht sich die Aussagekraft mit der Länge des Artikels.

4 Datenauswertung

Wie können die Daten überhaupt ausgewertet werden? Es müssen gleichzeitig die Performance der Features, als auch des Algorithmus zur Klassifizierung analysiert und ausgewertet werden.

4.1 Überblick

Ein paar Plots zeigen, die die Korrelation ausgewählter Features zeigen?

4.2 Klassifizierung

Warum versuche ich die Klassifizierung? Weil wenn Klassifizierung nicht wirklich funktioniert und kein Model trainiert werden kann, welches die Zeitung auf Basis der Daten predicten kann, dann ist auch keine Gruppierung möglich.

Ist zudem eine sehr gute Methode dafür festzustellen, an welchen Features man die Zeitungen am besten voneinander unterscheiden kann und welche eventuell völlig unnötig sind und so bei der Gruppierung nur stören würden.

4.2.1 Vorbereitung des Datensatzes

Für das Training muss der Datensatz in Train/Testset aufgeteilt werden. Welche Aufteilung habe ich hier gewählt und warum?

4 Datenauswertung

Die Daten müssen vorher normalisiert oder standardisiert werden. Kommt allerdings auch auf die ausgewählten Features und den gewählten Algorithmus an.

4.2.2 Feature Selection

Welche Verfahren wurden zur Feature Selection verwendet? Warum? Warum war das überhaupt nötig. Vergleich zu bzw. Verweis auf bisherige(n) Studien zur Feature Selection zur Genre/ Author detection.

(evtl. welche Features sind pro Zeitung relevant gewesen? weil unterschiedliche thetas pro Kategorie bei One vs Rest Klassifizierung)

4.2.3 Messung der Performance

Welche verschiedenen Ansätze gibt es hier?

F-Maß

Wahrheitsmatrix

Beurteilung der Wahrscheinlichkeiten

Wahrheitsmatrix kann auch mit den probabilities gemacht werden. Sagt noch mehr über die Sicherheit der Vorhersagen aus.

4.2.4 Verwendete Verfahren

Warum müssen verschiedene Verfahren getestet werden?

Linear Discriminant Analysis

Logistische Regression

Random Forest

Support Vector Machines

4.3 Clustering

Zuerst wurde untersucht, ob beim Clustering der einzelnen Artikeln jeder Zeitung Cluster entstehen, welche die einzelnen Zeitungen repräsentieren. Weiterhin wird überprüft, ob sich dabei Cluster ergeben, die Artikel verschiedener Zeitung haben. Das wäre dann schon ein sehr guter Indikator dafür, dass zwei oder mehr Zeitungen einen ähnlichen Schreibstil haben.

Ein weiterer Ansatz ist es, den Durchschnitt jeder Zeitung zu berechnen und anschließend die Zeitungen zu clustern. Dies hat jedoch Nachteile: Der Durchschnitt einer Zeitung ist nicht besonders repräsentativ, besonders wenn die Standardabweichung hoch ist. Zudem gibt es hier je nach Kategorie nur etwa 8 oder weniger "Beobachtungen" die geclustert werden können.

4.3.1 Feature Extraction

Wie in den Grundlagen bereits beschrieben, ist es beim Clustering besonders wichtig, dass die Dimensionalität nicht hoch ist. Vor allem für die Visualisierung ist es notwendig, die Features auf zwei Dimensionen zu reduzieren.

PCA

Hauptkomponentenanalyse: welche Ergebnisse gibt es hier?

4 Datenauswertung

t-SNE

Ein Verfahren, was genau dafür gedacht ist visuell Cluster zu zeigen, bei denen die Distanz untereinander im Plot keine Aussagekraft hat. Wie sehen hier die Ergebnisse aus? Spoiler: KACKEE

5 Ergebnis

Literaturverzeichnis

- [Dom12] Domingos, Pedro: A few useful things to know about machine learning. In: *Communications of the ACM* 55 (2012), Nr. 10, S. 78–87
- [Exp19] ExplosionAI: *Linguistic Features - spaCy Usage Documentation*. <https://spacy.io/usage/linguistic-features#dependency-parse>, 2019. – letzter Zugriff: 12.01.2019
- [HAK00] Hinneburg, Alexander; Aggarwal, Charu C.; Keim, Daniel A.: What is the nearest neighbor in high dimensional spaces? In: *26th Internat. Conference on Very Large Databases*, 2000, S. 506–515
- [Ng13] Ng, Andrew: *Machine Learning and AI via Brain simulations*. 2013

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Ort, Datum

Unterschrift