UNIVERSITÄT LEIPZIG

FAKULTÄT FÜR MATHEMATIK UND INFORMATIK
(INSTITUT FÜR INFORMATIK)

Eine computergestützte Analyse über die Grenzen und Dynamik des Overton-Fensters

BACHELORARBEIT
(DIGITAL HUMANITIES)
ERSTELLT VON:

David Hanslischeck

Betreuender Hochschullehrer:
Prof. Dr. Martin POTTHAST
Dr. Christian KAHMANN

4. April 2022

Inhaltsverzeichnis

1 Abstract **3**

2 Einleitung **3**

2.1 Overton-Fenster 4

2.2 Fragestellung 4

2.3 Motivation 6

2.3.1 Technisch 6

2.3.2 Wissenschaftlich 6

2.3.3 Gesellschaftlich 6

3 Vorarbeiten **8**

3.1 Context Change 8

3.2 BERT 8

3.3 GSBERT 9

3.4 SentiWS 10

3.5 SBERT 11

4 Experimente **12**

4.1 Vorbereitungen 12

4.2 Experiment 1 - Diachrone Sentiment-Analyse 13

4.2.1 Ablauf 13

4.2.2 Ergebnisse 16

4.2.3 Limitationen 24

4.3 Experiment 2 - Sentiment-Reaktion-Korrelation-Analyse 25

4.3.1 Ablauf 25

4.3.2 Ergebnisse 26

4.3.3 Limitationen 31

4.4 Experiment 3 - Diachrone Satz-Ähnlichkeit-Analyse 33

4.4.1 Ablauf 33

4.4.2 Ergebnisse 37

4.4.3 Limitationen 41

5 Zukünftige Arbeiten **42**

6 Diskussion **43**

7 Fazit **44**

8 Anhang **44**

Abbildungsverzeichnis

1 GSBERT: ‘taz’ Artikel 16
2 GSBERT: ‘taz’ Titel 17
3 GSBERT: ‘taz’ Sätze 18
4 SentiWS: ‘taz’ Artikel 19
5 SentiWS: ‘taz’ Sätze 20
6 SentiWS: Gewichtsverteilung Sätze 21
7 SentiWS: Gewichtsverteilung Artikel 22
8 Sentiment ‘taz’ Boxplot 23
9 Hyperpartisan Hand-Label 35
10 Sankey Diagramm: ‘refugee’ 37
11 Sankey Diagramm: ‘vaccine’ 38
12 Sankey Diagramm: ‘healthcare’ 39
13 Sankey Diagramm: ‘gunlaw’ 40

Tabellenverzeichnis

1 Datensatz von GSBERT 9
2 Erweiterungen für Python 12
3 Tweet Korrelation-Analyse GSBERT 26
4 Top Reaktionen Tweets 27
5 Negative Tweets GSBERT 28
6 Tweet Korrelation-Analyse SentiWS 29
7 Negative Tweets SentiWS 30
8 Beispielsätze für Hyperpartisan Hand-Label 36

1 Abstract

Das Overton-Fenster stellt ein potenziell mächtiges Modell dar, mit dessen Hilfe viele wichtige gesellschaftliche Fragen besser beantwortet werden könnten. Dieses ist bislang jedoch relativ unerforscht und bis heute gibt es keine standardisierten Verfahren, die das Fenster greifbar oder messbar werden lassen. In dieser Arbeit wird sowohl auf die Dynamik des Overton-Fensters eingegangen als auch Versuche vorgestellt, welche Veränderungen innerhalb des Fensters aufzeigen und die Grenzen des Fensters messbar werden lassen. Die zentrale Frage ist, in welcher Art und mit welchen Methoden das Fenster gemessen werden kann. Dafür werden drei verschiedene Ansätze mit verschiedenen Methoden verwendet. Die Grenzen des Fensters werden mithilfe einer diachronen Sentiment-Analyse, einer Sentiment-Reaktion-Korrelation-Analyse und einer Satz-Ähnlichkeit-Analyse ertastet. Dazu werden auch drei verschiedene Datensätze verwendet. Es werden deutsche und US-amerikanische Zeitungsartikel sowie deutsche Tweets aus der Plattform Twitter analysiert. Mithilfe dieser Datensätze und den angewendeten Analyseverfahren ist es möglich, diverse Verschiebungen der Grenzen des Overton-Fensters zu messen. Außerdem wird durch die Vielfalt an Versuchen ermittelt, welches Verfahren sich am besten zur Messung des Overton-Fensters eignen könnte. Schlussendlich macht die Arbeit auch darauf aufmerksam, dass das Overton-Fenster schon alleine in seiner Definition schwer zu fassen ist und dass für ein klares Bild des Fensters interdisziplinäre Zusammenarbeit von Experten dringend notwendig ist.

2 Einleitung

Im Lauf der Entwicklung der Menschheit entstanden, beeinflusst durch unterschiedliche Lebensumstände, viele verschiedene Deutungen und somit Ansichten, wie die Welt und das Leben funktioniert. Dies führte zu einer Vielzahl von unterschiedlichen Lebensweisen und Meinungen. So unterschiedlich diese Ansichten auch sein mögen, es vereint sie eine Eigenschaft: Sie werden von mindestens einem Menschen als 'universell richtig' anerkannt. Das bedeutet, dass diese Ansicht als seine Ideologie dient.

Durch den Zusammenschluss der Menschen und die Entstehung der ersten Gesellschaften trafen verschiedene Ideologien aufeinander. Unausweichlich mussten sich diese Ansichten anpassen, um eine stabile Gesellschaft zu ermöglichen. Dass solche Anpassungen meist nicht gewaltfrei verliefen – das lässt sich aus unseren Geschichtsbüchern erschließen – ist nicht verwunderlich, stellt Gewalt doch den 'einfachen' Weg dar. Aber nicht immer muss eine solche Anpassung gewaltvoll stattfinden.

Schauen wir auf unsere heutige Gesellschaft, können wir sowohl die gewaltvollen als auch die gewaltfreien Prozesse der Anpassung beobachten. Nehmen wir das Beispiel in der politischen Haltung zwischen Rechts und Links. Als Ergebnis der Anpassung sehen wir einen ewigen Diskurs. Doch beide Ansichten koexistieren miteinander, obwohl sie in vielen Bereichen widersprüchlich sind. Damit diese koexistieren können, muss es einen Bereich geben, welcher von beiden Ansichten geteilt wird. Dieser Bereich kann als Overton-Fenster bezeichnet werden.

2.1 Overton-Fenster

Das Overton-Fenster ist ein theoretisches Gesellschaftsmodell, benannt nach Joseph. P. Overton (*4. Januar 1960; †30 Juni 2003). Overton war ein US-amerikanischer Anwalt und Vizepräsident des 'Mackinac Center for Public Policy'. Sein Fenster gilt als ein Konzept, an das sich Politiker halten können, um keine radikalen oder extremen Aussagen zu treffen. Dabei bedient sich das Modell bestimmter Grundsätze, die in der Politik und Gesellschaft als allgemein 'richtig' und 'wahr' angesehen werden. Am besten kann man sich das Overton-Fenster als einen abstrakten Rahmen vorstellen, in dem alle Ideen und Aussagen, die von unserer Gesellschaft akzeptiert werden, zusammengefasst sind. Im Kern des Rahmens befinden sich alle grundsätzlichen Ideen und Aussagen, die von einer Mehrheit als richtig und zumindest vorübergehend als unumstößlich erachtet werden. Außerhalb des Rahmens liegen alle Ideen und Aussagen, welche als radikal, extrem oder undenkbar gelten (Jacobsen, 2018).
Overton hatte die Theorie, dass die Lebensdauer einer politischen Idee davon abhinge, ob sie in das Modell – oder anders gesagt in das Fenster – passt oder nicht. Das ist zunächst leicht nachzuvollziehen, versucht doch fast jeder Mensch sein Ansehen innerhalb einer Gruppe zu bewahren und nicht mit einer extremen Idee oder Aussage aufzufallen. Inwiefern jedoch eine über den Randbereich hinaus platzierte Aussage eine Veränderung der Grenzen des Overton-Fensters bewirkt, ist nicht nachvollziehbar. Um das beispielhaft zu erklären, gehen wir in der Zeit zurück. Bis einschließlich 1918 gab es in Deutschland kein Wahlrecht für Frauen. Erst 1919 durften sich Frauen aktiv an den Wahlen in Deutschland beteiligen. Damit diese 'neue' Idee in der Gesellschaft und vor allem im Gesetz fest verankert werden konnte, muss es eine Veränderung innerhalb des Overtonmodells gegeben haben. Die festgesetzten Grundsätze in dem Overton-Fenster scheinen damit nicht statisch, sondern dynamisch zu sein und das gleiche gilt für seine Grenzen. Um die Gestalt des Fensters zu begreifen, muss also nicht nur der inhaltliche, sondern auch der zeitliche Faktor beachtet werden. Das steigert die Komplexität dieses Fensters erheblich. Es wäre somit spannend herauszufinden, wie sich diese Dynamik auswirkt, welche Formen das Overton-Fenster annehmen kann und inwiefern es manipulierbar ist.

2.2 Fragestellung

Die theoretische Eigenschaft des Overton-Fensters könnte es ermöglichen, spannenden Gesellschaftsfragen auf den Grund zu gehen. Zum Beispiel: Wie verbreitet sich Meinung innerhalb einer Gesellschaft? Oder: Wie erkennt man wachsenden 'Extremismus'? Mit Extremismus ist eine Meinung oder Aussage gemeint, welche von der Öffentlichkeit weitgehend nicht akzeptiert ist. Das Problem ist jedoch, dass dieses Fenster nicht einfach zu fassen oder zu messen ist. Es existieren keine klaren Grenzen oder Konstanten und die dynamischen Verhältnisse innerhalb des Fensters sind bisweilen unbekannt. Diese Arbeit wird sich darauf konzentrieren, das Overton-Fenster aus verschiedenen Richtungen zu ertasten, um eine Idee zu bekommen, mit welchen Mitteln es messbar werden könnte. Die grundsätzliche Frage lautet also: Ist es möglich, das Overton-Fenster oder zumindest Teile des Overton-Fensters zu messen? Das Overton-Fenster an sich ist ein abstraktes Konstrukt, das keine Möglichkeit bietet, direkte Messungen vorzunehmen. Es könnte aber mög-

lich sein, dass große Textanalysen gewisse Muster preisgeben, die eine Veränderung innerhalb des Fensters andeuten. So habe ich in dieser Arbeit mit drei Experimenten versucht, das Overton-Fenster durch Textanalysen messbar und damit greifbar zu bekommen. Jedes der Experimente wird sich mit einer anderen Herangehensweise und Beobachtung dem Fenster nähern.

Das erste Experiment wird mithilfe von deutschen Zeitungsartikeln und einer diachronen Sentiment-Analyse versuchen, Veränderungen innerhalb des Fensters zu einem gegebenen Schlüsselwort zu entdecken. Die Fragestellung lautet: Lässt sich innerhalb eines Zeitraums eine messbare 'Gefühlsschwankung' zu dem Thema 'Flüchtling' nachweisen? Die Idee dahinter wäre, dass diese 'Gefühlsschwankung' als eine Bewegung innerhalb des Overton-Fensters interpretiert werden kann. Somit wäre eine Veränderung der Grundsätze nachweisbar. Zudem ergibt sich aus der Idee des ersten Experiments eine weitere Frage: Wenn man Veränderungen nachweisen kann, können Veränderungen dann auch prognostiziert werden, beziehungsweise kann die Ursache einer Veränderung entdeckt werden?

Das nächste Experiment wird mithilfe von gecrawlten deutschen Tweets aus der Plattform Twitter eine Sentiment-Reaktion-Korrelation-Analyse erstellen. Das bedeutet, dass eine Sentiment-Analyse zu den jeweiligen Tweets mit der Reaktionsbereitschaft anderer Nutzer verglichen wird. Reaktionsbereitschaft bedeutet in diesem Kontext das Kommentieren von Tweets durch andere Nutzer. Die Fragestellung lautet hier: Gibt es unterschiedliche Reaktionsbereitschaften bei Tweets, die 'positiv', 'negativ' oder 'neutral' gelabelt wurden? Die Idee wäre, dass eine Aussage, die viele Reaktionen bekommen hat, eine besondere Rolle innerhalb des Overton-Fensters erhält. Sprich, solche Aussagen könnten aktiv an Veränderungen innerhalb des Overton-Fensters beteiligt sein, was wiederum bedeutet, dass die Identifizierung solcher Aussagen ein besseres Verständnis über die Dynamik des Overton-Fensters bereitstellen könnte. Darüber hinaus wäre es auch spannend herauszufinden, ob und welche Nutzer das Overton-Fenster gezielt beeinflussen.

Mein drittes und letztes Experiment wird sich mithilfe von US-amerikanischen Zeitungsartikeln und einer Satz-Vektorisierung-Methode ein Bild darüber machen, ob gewisse Sätze oder Aussagen im Laufe der Zeit wiederverwertet wurden. Das ist besonders interessant, da alle Aussagen und Sätze im verwendeten Datensatz nach ihrer politischen Haltung gelabelt wurden. Damit lautet die letzte Fragestellung: Lässt es sich nachweisen, dass Sätze und Aussagen innerhalb eines gegebenen Zeitraums von einer anderen politischen Haltung übernommen wurden? Die Idee hinter dieser Frage scheint recht intuitiv. Wenn eine Aussage erst von der einen und später von der anderen politischen Seite verwendet wird, hat es offensichtlich einen Meinungswechsel gegeben. Die Wiederverwendung deutet also auf einen Wandel innerhalb des Overton-Fensters hin. Theoretisch müsste sich dieser Wandel auch auf die Grenzen und Grundsätze des Fensters ausgewirkt haben.

2.3 Motivation

2.3.1 Technisch

Bis vor ein paar Jahrzehnten wäre es noch unmöglich gewesen einen Überblick über die riesige Masse an Texten, Aussagen, Ideen und Meinungen zu bekommen. Doch in der heutigen Zeit, in welcher soziale Netzwerke einen öffentlichen bis halböffentlichen Raum darstellen und Meinungsverbreitung gerne und viel über diese Netzwerke praktiziert wird, ist es nicht mehr unvorstellbar einen groben Überblick über das landes- oder gar weltweit Gesagte beziehungsweise Geschriebene zu erhalten. Digital basierte Kommunikationssysteme, soziale Netzwerke, Content-Communities, Foren, Blogs und Kollektivprojekte wie Wikipedia sind allesamt maschinell lesbar und können ohne großen Aufwand von einem Computer verarbeitet werden. Das macht die Nutzung von computergesteuerten Text-Analysen um einiges einfacher und effizienter. Es gibt viele öffentlich zugängliche Datenbanken mit umfangreichen Kollektionen an Algorithmen für die verschiedensten Aufgabenbereiche der Computerlinguistik. Die Erforschung des Overton-Fensters bietet eine gute Möglichkeit, ein paar dieser Algorithmen auszutesten.

2.3.2 Wissenschaftlich

Da es ein zentraler Aufgabenbereich der Digital Humanities ist, konventionelle geisteswissenschaftliche Fragen in ein digitales Format zu konvertieren, liegt es nahe, das Overton-Fenster mithilfe computergesteuerten Textanalysen zu erforschen. Ein zusätzlicher Spannungsfaktor ist, ob ein abstraktes Konstrukt wie das Overton-Fenster durch Technik gemessen, beziehungsweise in Bezug auf den Rahmen, vermessen werden kann. Auch wenn in den letzten Jahren enorme Fortschritte im Bereich von computergesteuerten Textanalysen gemacht wurden, ist es dennoch eine technische Herausforderung, Millionen von Aussagen und Sätzen effizient zu analysieren und miteinander zu vergleichen. Es ist spannend herauszufinden, ob mit diesen neuen Errungenschaften der Technik tiefere Einblicke in komplexe Systeme wie das Overton-Fenster möglich sind. Mit immer präziser werdenden Sentiment-Analyse-Algorithmen und leistungsstarken Satz-Vektorisierung-Algorithmen im Bereich maschinelles Lernen, scheinen immer bessere Werkzeuge zur Verfügung zu stehen, um das Overton-Fenster zu erforschen.

2.3.3 Gesellschaftlich

An diesem Punkt kommt vielleicht die Frage auf, warum es – neben der technischen und wissenschaftlichen Herausforderung – generell interessant ist, etwas über das Overton-Fenster herauszufinden. Es gibt mehrere Eigenschaften des Fensters, die für unsere heutige Gesellschaft durchaus relevant sein könnten. Unabhängig vom technischen Aspekt ist der weitaus interessantere der gesellschaftliche Aspekt. Das Overton-Modell hat das Potenzial gesellschaftliche Probleme und Fragen besser zu veranschaulichen, was wiederum helfen könnte, Lösungen für diese Probleme und Fragen zu finden. Wie verbreitet sich eine Meinung innerhalb einer Gesellschaft? Welche Bedeutung haben gewisse Plattformen für eine Gesellschaft? Ab wann ist eine Aussage extrem und wo liegen die Grenzen? Wie und wo entstehen Gruppen mit extremen Meinungen? Lässt

sich Extremismus genauer definieren? Gibt es Möglichkeiten, heranwachsende Konflikte frühzeitig zu erkennen und zu verhindern? Ab wann ist eine Meinung keine Meinung mehr, sondern eine Ideologie? Gibt es Meinungen, die mehr 'wert' sind als andere? Wie viele Menschen stehen in Konflikt mit anderen Ideologien? Welche Ansichten sind ergänzend, welche widersprüchlich? Die Liste kann ewig fortgesetzt werden und die Möglichkeiten, solche Fragen zu stellen, scheinen unerschöpflich. Darüber hinaus könnte ein besseres Verständnis über das Overton-Fenster hilfreich sein, um bewusste Manipulationen einzudämmen und das Fenster vor gezielten Missbrauch zu schützen. Um jedoch das Overton-Fenster für solche Zwecke nutzen zu können, wird die Zusammenarbeit von vielen verschiedenen Experten und Wissenschaftlern aus Bereichen wie Politik, Soziologie, Jura, Geschichte und weiteren Bereichen benötigt. Als Bezugspunkte für die 'Grundsätze' gelten in dieser Arbeit also nur die Daten selbst beziehungsweise die jeweils verwendeten Algorithmen. Mehr dazu in der Diskussion auf Seite 43.

Zusammengefasst ist das Ziel dieser Arbeit eine Einschätzung, ob es realisierbar wäre, ein konkretes Bild von dem Overton-Fenster zu erhalten. Die Frage wäre damit inwiefern es technisch und methodisch möglich ist, einen groben Überblick und Einblick in das Overton-Fenster zu bekommen, sowie die Erforschung von Dynamiken innerhalb dieses Fensters.

3 Vorarbeiten

3.1 Context Change

Die zuvor erwähnten Experimente bedienen sich an Methoden und Techniken, die sich in gewisser Weise schon in anderen Disziplinen bewährt haben. Zu meiner Arbeit hat mich die Abhandlung 'Measuring Context Change to Detect Statements Violating the Overton Window' von Kahmann und Heyer (2019) inspiriert. Sie ist einer der ersten Versuche, sich über eine computergesteuerte Analyse dem Overton-Fenster zu nähern. Die Idee in dieser Arbeit ist es, einen Standard-Kontext zu einem Thema für eine gewisse Zeitspanne zu identifizieren. Dieser Standard-Kontext kann dann mit neuen Dokumenten verglichen werden und zeigt auf, wie hoch der Wortanteil ist, der noch nie oder noch nicht in diesem Kontext genutzt wurde. Fällt dieser Anteil groß aus, ist es wahrscheinlicher, dass das Dokument eine neue und möglicherweise extreme Meinung beinhaltet. Dies kann letztlich als Indiz dafür dienen, dass diese Meinung die Grenzen des Overton-Fensters verschoben hat. Als Datensatz verwendeten Kahmann und Heyer (2019) Zeitungsartikel aus der Tageszeitung 'taz' in deutscher Sprache. Dieser Datensatz umfasst einen Großteil der von 2010 bis 2018 publizierten Artikel. Da ich selbst diesen Datensatz verwende, werde ich später auf die Details dazu eingehen. Der Standard-Kontext bezieht sich in diesem Paper auf das Thema 'Flüchtling'. Mithilfe einer Satz-Term-Matrix wurde jeweils das Signifikanzmaß von Dice berechnet, das ein Wort zu diesem Thema hat. Das Signifikanzmaß wird genutzt, um in einem weiteren Messverfahren (ähnlich zu Context Volatility (Kahmann et al., 2017)) eine Distanz zu berechnen. Diese Distanz bezieht sich auf einen Satz und dessen Abweichung vom Standard-Kontext und wird als Menge der Unvorhersehbarkeit definiert. Damit ist es möglich, extreme Aussagen und Meinungen zu entdecken, welche scheinbar die Grenzen des Overton-Fensters überschreiten. Jedoch gibt es auch Limitationen die erwähnt werden. Zum einen muss die Verwendung von neuen Wörtern nicht zwingend eine Meinung außerhalb des Fensters repräsentieren. Zum anderen wurde dieser Vergleich nur zwischen Referenzpunkt und Testpunkt erstellt und konnte damit nicht auf das Verschieben der Grenzen oder allgemein auf die Dynamik des Fensters eingehen. Für zukünftige Arbeiten erhoffen Kahmann und Heyer (2019) sich sowohl eine diachrone Ansicht auf das Fenster, als auch die Einbettung von Sentiment.

3.2 BERT

Das im Jahr 2019 vorgestellte 'Bidirectional Encoder Representations from Transformers', kurz 'BERT' (Devlin et al., 2019) Modell, spielte eine Schlüsselrolle für den Fortschritt im Bereich der natürlichen Sprachverarbeitung (Saha et al., 2019). Seit seiner Veröffentlichung sind eine Vielzahl an verschiedenen vortrainierten Modellen und Methoden erschienen, die sich allen möglichen Aufgabenbereichen der Computerlinguistik widmen. BERT spielt hierbei eine zentrale Rolle für meine Arbeit. Alle drei Experimente verwenden ein Modell basierend auf BERT.
BERT ist designt worden, um bidirektionale Sprach-Repräsentationen vorzutrainieren, indem es den Kontext von beiden Seiten eines Tokens in allen Layern aufbereitet. Dieser Ansatz verbessert vorherige Modelle, die nur unidirektionale Sprach-Repräsentationen verwendeten. Die gezeigten Ergebnisse brachten BERT an die Ranking-Spitze in elf verschieden Bereichen der Computer-

linguistik. Damit erreichte BERT 'state-of-the-art' Status. Die Modellarchitektur ist ein 'bidirectional multi layer transformer' basierend auf der beschriebenen Implementierung von Vaswani et al. (2017). Zusätzlich verwendet das Modell eine bidirektionale 'self-attention'. Das bedeutet, dass jedes Token den Kontext seines rechten und linken Nachbarn einsehen kann. Mit diesen Eigenschaften ist es möglich, eine Vielzahl an unterschiedlichen Aufgaben zu bewältigen und das, indem nur wenige bis keine Anpassungen vorgenommen werden müssen.

Zwar scheint BERT in vielen Aufgabenbereichen universell verwendbar zu sein, dennoch wurden Hunderte von weiteren Sub-Modellen entwickelt, die alle auf BERT basieren. Auch für meine Experimente existieren Sub-Modelle, die sich besser eignen als das Basismodell. Ich habe mir für meine Aufgaben zwei Sub-Modelle ausgesucht. Das eine ist darauf spezialisiert, Sätze in gleich dimensionale Vektoren zu transformieren. Das andere kann das Sentiment für deutschsprachige Texte klassifizieren.

3.3 GSBERT

Es wurden zwar sehr viele Sentiment-Modelle seit der Veröffentlichung von BERT entwickelt, doch schränkt man die Suche auf den deutschsprachigen Raum ein, bleiben nicht mehr viele Sentiment-Analyse-Verfahren übrig. Eines der momentan präzisesten (und öffentlich verfügbaren) Modelle ist das 'German-Sentiment-BERT' Modell von Guhr et al. (2020). Diese Arbeit hatte sich als Ziel gesetzt, eine Methode zu finden, die Serviceroboter dazu befähigen soll, die Gefühle ihrer Nutzer zu interpretieren. Tatsächlich wurden in dieser Abhandlung zwei Modelle trainiert, eines mit 'FastText' (Joulin et al., 2016) und eines mit BERT. Da das FastText Modell jedoch in fast allen Ergebnissen schlechter abschneidet, werde ich hier nicht weiter auf dieses Modell eingehen. Das BERT Modell wurde mit ungefähr 5,4 Millionen gelabelten Samples trainiert. Die Datensätze umfassen folgende Sammlungen in Tabelle 1.

Datensatz	Quelle	Samples	Neutral	Negativ	Positiv
PotS	(Sidarenka, 2016)	7.504	2.487	4.569	3.448
SB10k	(Cieliebak et al., 2017)	7.474	4.628	1.130	1.716
GermanEval-2017	(Wojatzki et al., 2017)	23000	16.309	5.845	1.371
Scare	(Sänger et al., 2016)	735.382	0	197.279	538.103
Filmstarts	filmstarts.de	55.659	0	15.610	40.049
Holidaycheck	holidaycheck.de	3.524.193	0	388.744	3.135.449
Leipzig-Wikipedia	(Goldhahn et al., 2012)	1.000.000	1.000.000	0	0
Emotions	(Guhr et al., 2020)	1.306	28	1.090	188
Total		5.355.043	1.023.452	611.267	3.720.324

Tabelle 1: Hier sehen wir die verwendeten Datensätze von Guhr et al.. PotS und SB10k sind Sammlungen von Tweets. Scare, Filmstarts und Holidaycheck sind allesamt Reviews. GermanEval-2017 besteht aus gemischten Textarten. Emotions ist ein von Hand zusammengesammelter Satz an Texten, welcher viele Beleidigungen beinhaltet.

Die Daten umfassen somit knapp 3,7 Millionen positive, etwa 1 Millionen neutrale und ungefähr 600.000 negative Samples. Da dieses Verhältnis nicht ausgeglichen ist, haben Guhr et al. (2020) das Modell nochmals mit einem ausgeglichenen Verhältnis trainiert. Die Scare Sammlung diente als Evaluierungsdatensatz. Sie wurde nicht in das Training integrierten und diente letztendlich als ungesehener Datensatz für die Leistungskontrolle. Mit einem F1-Score von 0,80 mit dem Scare Satz und einem allgemeinem F1-Score von 0,96 ist das vorgestellte Modell das momentan beste und umfangreichste, das für die Sentiment-Analyse der deutsche Sprache zur Verfügung steht. Da Guhr et al. ihrem Modell keinen Namen gegeben haben, werde ich es folgend als 'GSBERT' (German-Sentiment-BERT) bezeichnen. German BERT (GBERT) wurde mittlerweile schon verwendet (Scheible et al., 2020).

3.4 SentiWS

GSBERT ist aber nicht das einzige Sentiment-Modell das ich verwende. So gut GSBERT auch funktioniert, gibt es mehrere Limitationen, die mit dem Modell einhergehen. Ich werde später nochmals detailliert auf die Limitationen zu sprechen kommen, jedoch ist eine davon, dass es nicht möglich ist herauszufinden, wie stark ausgeprägt das Sentiment einer Aussage ist. Um dieser Limitation auszuweichen, verwende ich auch 'SentimentWortschatz' oder kurz 'SentiWS' (Remus et al., 2010), für eine Sentiment-Analyse.
SentiWS umfasst eine Liste an Sentiment behafteten Wörtern zusammen mit ihren Flexionen. Jedem Wort wird ein Gewicht in dem Intervall [-1; 1] zugeordnet. In meinen Versuchen verwende ich die SentiWS Version 2.0. Insgesamt stehen 1.828 negative und 1.645 positive Wörter gegenüber. Zusammen ergeben sich damit 34.126 gewichtete Flexionen. Die Wörterliste enthält sowohl Adjektive und Adverbien, welche explizit Sentiment ausdrücken, als auch Substantive und Verben, welche implizit Sentiment enthalten. Als erste Quelle für diese Liste wurde das General Inquirer Lexikon (Stone et al., 1967) verwendet. Die Kategorien 'positiv' und 'negativ' wurden dann mit Google-Translate ins Deutsche übersetzt und nachträglich per Hand bearbeitet. Die zweite Quelle entsprang aus einer Kookkurrenzanalyse bewerteter Produktrezensionen eines anonymen Geschäftspartners. Die dritte Quelle entstammt dem German Collocation Dictionary (Quasthoff, 2010), dass mithilfe einer Kollokationsanalyse Wörter mit semantischer Gleichheit gruppieren konnte. Die Gewichte zu all den Wörtern, die aus den drei Quellen zusammengetragen wurden, werden durch die sogenannte "Pointwise Mutual Information" (PMI) (Church & Hanks, 1989) Methode ermittelt und anschließend auf ein Intervall von -1 bis +1 skaliert. Ein Gewicht repräsentiert dementsprechend wie stark ein Wort mit einem Sentiment assoziiert wird. Dabei repräsentiert '-1' absolut negativ und '+1' absolut positiv. Die Verteilung der Gewichte folgt einer Zipfschen Verteilung (Zipf, 1972), das heißt nur wenige Wörter haben ein großes Gewicht, ein paar haben ein mittleres Gewicht und sehr viele haben ein kleines Gewicht.
SentiWS scheint immer noch weiterentwickelt zu werden und stellt keinen vollkommen fehlerfreien Wortschatz dar. Dennoch zeigt SentiWS seine Stärken. Es eignet sich hervorragend um eine Intensität eines Sentiments zu ermitteln, solange man beachtet, dass es nur Annäherungen sind. Zudem ist ein Verfahren, das SentiWS nutzt, auch auf älteren Computern gut ausführbar.

3.5 SBERT

Bei dem letzten Verfahren, welches ich vorstellen möchte, handelt es sich um kein Sentiment-Analyse-Verfahren, da ich für mein letztes Experiment kein weiteres Sentiment-Verfahren verwende. Dieses Mal versuche ich das Overton-Fenster mit Satz-Vektorisierung greifbar zu bekommen. Wie anfangs erwähnt, gibt es auch für diese Aufgabe ein Sub-Modell von BERT, das Sätze in Vektoren gleicher Länge transformiert. 'Sentence-BERT' oder kurz 'SBERT' (Reimers & Gurevych, 2019) stellt eine kombinierte Modifikation von BERT und dem 'RoBERTa' Modell (Liu et al., 2019) dar. RoBERTa steht für 'Robustly Optimized BERT Pretraining Approach' und ist – wie der Name schon verrät – eine weitere optimierte BERT Version. RoBERTa kann dementsprechend auch als Sub-Modell von BERT angesehen werden. Mit RoBERTa wurde gezeigt, dass die Leistung von BERT durch kleine Anpassungen im Vortrainingsprozess weiter gesteigert werden kann.
BERT ist auch ohne Modifikation in der Lage, Sätze zu vergleichen und ein Ähnlichkeitsmaß zu ermitteln, jedoch sorgt die Architektur von BERT dafür, dass solche Berechnungen sehr viel Rechenleistung benötigen. Will man zum Beispiel die zwei ähnlichsten Sätze unter 10.000 finden, benötigt BERT etwa 65 Stunden, SBERT hingegen nur 5 Sekunden, und das ohne dabei an Präzision zu verlieren. Dieser Effizienzsprung wird mit der Nutzung von siamesische- und Tripel Netzwerkarchitekturen (Schroff et al., 2015) erreicht. Die siamesische Netzwerkarchitektur ermöglicht, dass die Input-Sätze in Vektoren gleicher Länge umgewandelt werden können. Das heißt, dass die Vektoren mit einer gleichen Anzahl an Elementen auch den gleichen multidimensionalen Vektorraum besitzen. Zusätzlich wird dafür gesorgt, dass Sätze mit ähnlicher Semantik sich dementsprechend nah beieinander im gleichen Vektorraum befinden. Mit dieser Eigenschaft der Vektoren kann anschließend eine Ähnlichkeitsmessung durchgeführt werden, wie zum Beispiel die Kosinus-Ähnlichkeit oder der Euklidische Abstand. Das Maß repräsentiert wiederum, wie semantisch ähnlich sich zwei Sätze sind. Für den fine-tune Prozess werden sowohl siamesische als auch Tripel Netzwerke erstellt. Die Netzwerkarchitektur ist abhängig von der sogenannten 'objective function' oder Zielfunktion. Diese hilft, eine gute Aktualisierung der Gewichte innerhalb des Netzwerks zu realisieren. Insgesamt ist mit drei verschiedenen Netzwerkarchitekturen mit unterschiedlichen Zielfunktionen experimentiert worden. Zusätzlich wurde jeweils dem Output des SBERT Modells eine pooling Funktion hinzugefügt, um Satz-Embeddings fester Größe herzuleiten. Damit ist es Reimers und Gurevych (2019) gelungen ein Modell zu entwickeln, das alle bisherigen Satz-Embedding Methoden sowohl in der Qualität als auch in der benötigten Laufzeit übertrifft.

Es scheint so, dass BERT und seine Sub-Modelle herausragende Ergebnisse in fast allen Bereichen der Computerlinguistik erzielen. Damit liegt es nahe, dass ich diese Modelle auch in meine kommenden Experimente implementiere. Alle vorgestellten Modelle sind auch als Packages für Python erhältlich.

4 Experimente

4.1 Vorbereitungen

Für die Umsetzung meiner Experimente habe ich eine Pipeline in der Programmiersprache Python geschrieben. Ziel dieser Pipeline ist, eingespeiste Datensätze automatisch zu analysieren und entsprechend zugehörige Tabellen und Grafiken zu erstellen. Insgesamt sind dadurch mehr als 1900 Zeilen Code entstanden. Die Pipeline teilt sich in zehn Klassen ein und beinhaltet drei Mains. Der Code bedient sich an mehreren Modulen, Packages und Libraries. Die wichtigsten darunter sind in Tabelle 2 zu sehen.

Name	Beschreibung
'time'	Wird verwendet um Laufzeiten zu berechnen.
'math'	Bietet komplexe arithmetische Operationen an.
'numpy'	Ermöglicht effiziente Berechnungen von Operationen auf Vektoren.
're'	Bettet reguläre Ausdrücke (Regex) in Python ein.
'pickle'	Ermöglicht das direkte Speichern von Python Objekten auf der Festplatte.
'pandas'	Beinhaltet ein umfangreiches 'Data Framework' für Python.
'matplotlib'	Bibliothek für die Erstellung statistischer Grafiken.
'seaborn'	Bibliothek für die Erstellung statistischer Grafiken, basierend auf matplotlib.
'scipy'	Umfasst mehrere Algorithmen für die wissenschaftliche Informatik.
'sklearn'	Bibliothek für maschinelles Lernen in Python.
'nltk'	Das 'Natural Language Toolkit' für die Be- und Verarbeitung von Sprachen.
'sentence-splitter'	Beinhaltet Methoden für eine effiziente Trennung in Sätze.
'transformers'	Beinhaltet das Transformers Modell.
'germansentiment'	Beinhaltet das GSBERT Modell.
'sentence_transformers'	Beinhaltet das SBERT Modell.

Tabelle 2: Auflistung der wichtigsten verwendeten Modulen, Packages und Libraries der Pipeline.

Zusätzliche Erweiterungen sind 'ast', 'tqdm', 'sys', 'os', 'datetime', 'csv', 'plotly', 'logging' und 'typing'. Gegebenenfalls müssen die jeweiligen Packages zuerst durch den '`pip install <packagename>`' Konsolenbefehl auf dem Computer heruntergeladen und installiert werden. Alle folgenden Experimente wurden mit dieser Pipeline realisiert. Der gesamte Code wurde in ein Google Colab Notebook übertragen und lässt sich theoretisch auch dort ausführen. Aus Vorsicht, kein Urheberrecht zu verletzen, stehen die verwendeten Daten jedoch nicht zur Verfügung. Eine Ausführung des Codes ist somit nur mit eigenen Datensätzen möglich. Einen Link zu dem Notebook finden Sie in dem Anhang auf Seite 44.

4.2 Experiment 1 - Diachrone Sentiment-Analyse

4.2.1 Ablauf

In dem ersten Experiment möchte ich mich der Frage widmen, ob es in den Printmedien eine Veränderung in der 'Gefühlslage' zu dem Thema 'Flüchtling' gab. Gefühlslage bedeutet hier, mit welcher Haltung und Empfindung die Gesellschaft im Allgemeinen oder die Zeitung im Konkreten zu diesem Thema steht. Das kann auch als das Sentiment bezeichnet werden. Wenn sich die Empfindung verändert hat, spricht es dafür, dass eine Meinung oder Idee entweder mehr oder weniger akzeptiert wird. Die Hypothese ist, dass eine Veränderung im Sentiment als Indiz für eine Verschiebung der Grenzen des Overton-Fensters genommen werden kann. Zudem könnte man das Maß an Veränderungen gegebenenfalls als Referenz nehmen, um eine präzisere Entdeckung von radikalen und extremen Aussagen zu ermöglichen. Damit möchte ich an den vorgestellten Ansatz von Kahmann und Heyer (2019) anschließen.
Wie schon zu Beginn der Arbeit erwähnt, verwende ich bei diesem Experiment denselben Datensatz, den Kahmann und Heyer (2019) verwendet hat. Die Daten umfassen Zeitungsartikel aus der Tageszeitung 'taz' von 2010 bis 2018. Mit einer Größe von 2,2 GB stehen 339.367 Artikel zur Analyse zur Verfügung. Als Bezugsthema für diesen Versuch verwende auch ich das Stichwort 'Flüchtling'. Nach einer Schlüsselwortsuche bleiben mit 163 MB, 37.966 Artikel übrig. Für die Sentiment-Analyse verwende ich zuerst GSBERT. GSBERT setzt eine Liste aus Strings als Input voraus. Die Textbereinigung selbst wird von GSBERT übernommen und muss nicht vorher ausgeführt werden. Das vorhandene GSBERT Modell wurde insoweit von mir modifiziert, dass zusätzlich zu den Sentiment-Ausgaben 'positiv', 'negativ' und 'neutral' auch die entsprechenden 'Logits' ausgegeben werden. Meine Idee war, dass die Logits gegebenenfalls als Intensitätsmaß des Sentiments dienen können. Auf die Tatsache, dass Logits nicht als solches dienen können, komme ich gleich zu sprechen.
Es sollte zuerst erwähnt werden, dass dieser Schritt mit meinem PC praktisch nicht möglich war. Mein verfügbarer Computer beinhaltet als Prozessor einen AMD Phenom II X6 1090T mit 12GB RAM und als Grafikkarte eine AMD Radeon HD 5770 mit 1GB RAM. Die Leistung meines Computers ist dementsprechend niedrig und die Berechnungen, die BERT, SBERT und GSBERT betreiben müssen, überschreiten schnell die Limits meines Systems. Als Alternative verwendete ich das schon erwähnte Notebook von Google Colab. Leider ist die Ressourcenzuteilung von Colab nicht immer zuverlässig und auch oft wird die Verbindung willkürlich abgebrochen. Das ist besonders ärgerlich, wenn stundenlange Berechnungen direkt vor den Augen gelöscht werden. Nichtsdestotrotz unterstützt Colab die 'CUDA' Technologie, die von neueren Grafikkarten genutzt werden kann. Das sorgt für einen großen Leistungsschub. Im Vergleich benötigt eine Sentiment-Analyse für einen 5 MB großen Datensatz mit 1000 Artikeln ungefähr 55 Minuten ohne CUDA Unterstützung, und 67 Sekunden mit CUDA Unterstützung. Das entspricht etwa 2% der ursprünglichen Laufzeit. Die Nutzung von CUDA ist damit unverzichtbar, vor allem wenn große Mengen an Daten analysiert werden sollen.
Nachdem GSBERT via Google Colab zu jedem Artikel eine Sentiment-Klassifikation und die dazugehörigen Logits ausgegeben hat, speichere ich alle Ergebnisse samt relevanter Metadaten in

einer neuen ‘TSV’ (Tab Separated Value) Datei ab. Ich habe mich für TSV entschieden, damit die Datei auch gut für das menschliche Auge lesbar ist. Die neue TSV Datei kann auch von der Pipeline eingelesen werden, falls man nicht noch mal eine Sentiment-Berechnung vornehmen möchte. Anschließend werden die Ausgaben von GSBERT analysiert. Hier komme ich noch mal auf die Logits zu sprechen. Laut Google, werden Logits folgendermaßen definiert:

> “Ein Logits ist ein Vektor mit rohen, nicht-normalisierten Vorhersagen, die von einem Klassifikationsmodell generiert werden. Üblicherweise wird dieser Vektor zu einer Normalisierungsfunktion weiter geleitet. In einer Multi-Klassen-Klassifikation werden Logits einer softmax Funktion übergeben. Die softmax Funktion generiert daraufhin einen Vektor mit normalisierten Wahrscheinlichkeitswerten für jede Klasse.” (Google, 2022).

Dementsprechend sind Logits eigentlich nicht dazu geeignet, als ein Intensitätsmaß für ein Sentiment herangezogen zu werden. GSBERT verwendet jedoch keine softmax Funktion für seine Vorhersagen. Die Logits werden dagegen einer argmax Funktion übergeben, welche den Index des größten Eintrags ausgibt. Es gibt drei Einträge, wodurch die Ausgabe von 0 bis 2 reicht. 0 steht dabei für die Klasse ‘positiv’, 1 für die Klasse ‘negativ’ und 2 für die Klasse ‘neutral’. Jeder Eintrag kann dabei einen Wert im Intervall zwischen [-10; 10] annehmen. Durch die fehlende softmax Funktion kam mir der Gedanke, dass es gegebenenfalls möglich ist, explizit bei GSBERT die Logits als Intensitätsmaß heranzuziehen. Der Wert 10 wäre somit eine absolute Angehörigkeit zu einer Klasse, während -10 keinerlei Angehörigkeit repräsentiert. Da ich aber weder eine Bestätigung noch eine Widerlegung zu dieser konkreten Anwendung gefunden habe, habe ich mich dagegen entschieden, Logits als Intensitätsmaß heranzuziehen. GSBERT wird also nur für die Kategorisierung des Sentiments verwendet.

Als Alternative habe ich zusätzlich eine Methode implementiert die SentiWS verwendet. Dabei werden für jeden Text Gewichte berechnet, welche die Intensität des enthaltenen Sentiments spiegeln. Der Text kann ein oder mehrere Sätze beinhalten. Folgendes muss dabei beachtet werden: Nicht jedes Wort besitzt ein Gewicht und alle enthaltenen Wortgewichte werden pro Text miteinander addiert. Zwangsläufig werden Texte mit mehr Wörtern größere Gewichtungen erhalten. Damit kein ‘bias’ für lange Texte entsteht, müssen die Gewichte an die Textlänge angepasst werden.

Angepasstes Gewicht:

$$w_{angepasst} = w \cdot \left(1 - \left(\frac{W_{norm} + T_{norm}}{2}\right)\right) \tag{1}$$

In Formel (1) werden die Textgewichte entsprechend ihrer Anzahl an enthaltenen Wörtern W und enthaltenen gewichteten Wörtern T angepasst. W_{norm} und T_{norm} sind nach ‘MinMax-Scaling‘ normalisiert worden. Die innere Klammer nimmt einen maximalen Wert von 1 und einen minimalen Wert von 0 an. Mit steigender Satzlänge werden die Gewichte entsprechend prozentual verkleinert. Damit der längste Text im Datensatz nicht automatisch das Gewicht 0 erhält, erhöhe ich beim MinMax-Scaling den maximal Wert um den Faktor 1,1. Das entspricht einer Gewichtung von 10%

bei maximaler Wörteranzahl. Anschließend werden die Sätze anhand ihres Gewichts kategorisiert. Ein positives Gewicht entspricht einem positiven Sentiment, ein negatives Gewicht einem negativen Sentiment und kein Gewicht wäre ein neutrales Sentiment. Es kommt jedoch selten vor, dass ein Text kein Gewicht erhält, besonders wenn komplette Artikel analysiert werden. Zudem ist es fragwürdig, ob die Entdeckung von nur einem gewichteten Wort hinreichend genug ist, um das Sentiment des kompletten Textes zu identifizieren. Dementsprechend habe ich mich dafür entschieden, eine Toleranz für die neutrale Klassifizierung einzubauen. Die Toleranzgrenzen passen sich dabei dynamisch der Textlänge an und werden mit folgender Formel (2) berechnet.

$$\text{Toleranzgrenze:} \qquad \begin{cases} Toleranz_{min} = \dfrac{\Delta W_{\varnothing} + \Delta T_{\varnothing}}{2} \cdot \overline{w}_{negativ} \\ Toleranz_{max} = \dfrac{\Delta W_{\varnothing} + \Delta T_{\varnothing}}{2} \cdot \overline{w}_{positiv} \end{cases} \tag{2}$$

Die Variable $\Delta W_{\varnothing}$ entspricht dem Verhältnis der durchschnittlichen Gesamt-Wörteranzahl zwischen dem Text und einem Satz. $\Delta T_{\varnothing}$ entspricht dem Verhältnis der durchschnittlichen Wörteranzahl mit Gewicht zwischen dem Text und einem Satz. $\overline{w}$ ist das durchschnittliche Gewicht der positiven oder negativen Wörterliste von SentiWS. Diese liegen bei -0,16 und 0,07. Gewichte innerhalb dieser Werte entsprechen der einmaligen Verwendung eines durchschnittlich gewichteten Wortes. Bei einer Satzanalyse löst sich der Bruch auf und die durchschnittlichen Gewichte $\overline{w}_{positiv}$ und $\overline{w}_{negativ}$ dienen als Grenzen. Dementsprechend werden alle Gewichtungen, die im Intervall [-0,16; 0,07] mal dem Toleranzfaktor liegen, in diesem Versuch als neutral eingestuft.

4.2.2 Ergebnisse

Zuerst schauen wir uns die Ergebnisse von GSBERT zu den ‘taz’ Artikeln an.

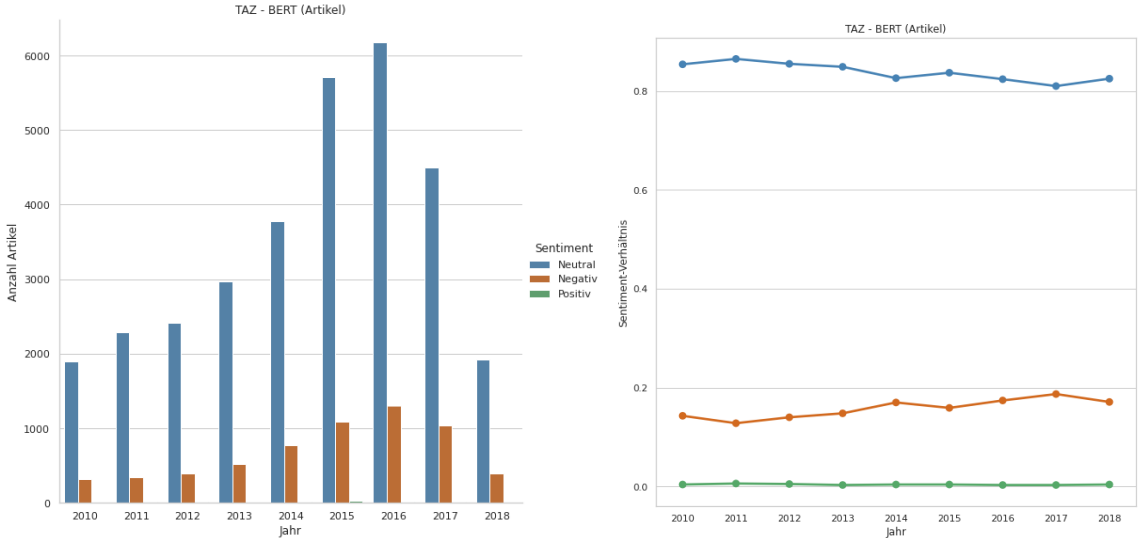

Abbildung 1: Links im Bild sehen wir die Anzahl an Artikeln pro Jahr. Jeder Artikel wurde einem Label zugeordnet. Neutral ist blau gekennzeichnet, negativ orange und positiv ist grün markiert. Auf der rechten Seite im Bild sehen wir die dazugehörigen Verhältnisse pro Jahr. Über die Jahre ist der Anteil der negativen Artikel gewachsen.

In Abbildung 1 sieht man die Anzahl an erschienenen Artikeln mit ihrem dazugehörigen Sentiment. In dem linken Diagramm lässt sich deutlich erkennen, dass das Thema ‘Flüchtling’ für die Zeitung ‘taz’ an Bedeutung gewann und bis 2016 mehr darüber berichtet wurde. In dem rechten Diagramm kann man sehen, dass im Laufe der Jahre im Verhältnis mehr negative Artikel erschienen sind. Bemerkenswert hierbei ist der kaum vorhandene positive Anteil. Die Gesamtzahl an Artikeln in 2018 ist deutlich geringer als zum Vorjahr, da nur Artikel aus der ersten Hälfte des Jahres im Datensatz vorhanden sind. Das Jahr 2018 ist somit nicht vollständig in dieser Analyse. Unabhängig davon ist über die Jahre ein leichter Trend beobachtbar. Das Thema ‘Flüchtling’ wurde immer präsenter und das assoziierte Sentiment negativer. Die Berichterstattung der ‘taz’ über das Thema ‘Flüchtling’ ist damit scheinbar unerfreulicher als früher. Das wiederum ist ein Indiz dafür, dass sich die Grenzen im Overton-Fenster verschoben haben. Da aber nur ein leichter Trend beobachtbar ist, ist das auch nur ein schwaches Indiz.

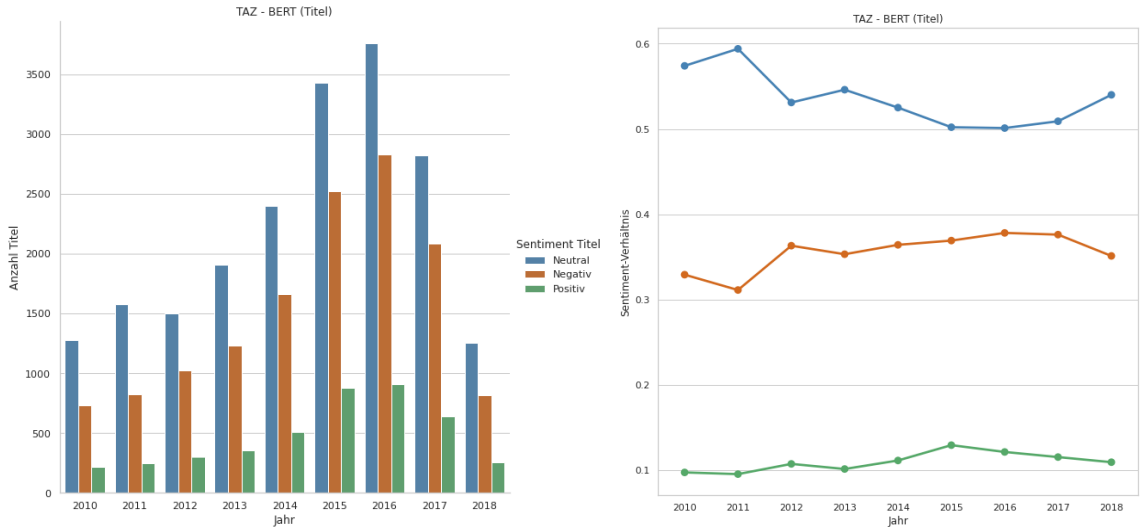

Abbildung 2: Links im Bild sehen wir die Anzahl an Artikeltitel pro Jahr. Jeder Artikel wurde einem Label zugeordnet. Neutral ist blau gekennzeichnet, negativ orange und positiv ist grün markiert. Auf der rechten Seite im Bild sehen wir die dazugehörigen Verhältnisse pro Jahr. Über die Jahre ist auch hier der Anteil der negativen Titel gewachsen.

In Abbildung 2 wurde das Sentiment der Schlagzeile eines Artikels berechnet. Auch diese sind wieder in ihrer Anzahl angegeben. Interessanterweise ist der negative Anteil weitaus größer und auch der positive Anteil ist deutlich größer als bei den gesamten Artikeln. Das lässt sich wahrscheinlich dadurch erklären, dass Zeitungen im Allgemeinen, beziehungsweise Autoren im Konkreten, die Leserschaft dazu animieren wollen, ihren Artikel zu lesen. Je sensationeller die Schlagzeilen einer Zeitung sind, desto interessanter wirken die Information die sie beinhalten und desto eher wird die Zeitung dann auch gekauft und gelesen. Das Gleiche könnte für die Autoren der 'taz' gelten. Angeblich setzt man Sensation mit Sentiment um. Das würde zumindest den erhöhten negativen und positiven Anteil erklären. Doch auch hier lässt sich über die Jahre ein leichter Trend erkennen: weniger neutrale und mehr negative Titel. 2018 bricht den Trend, aber ich erwähne nochmal dass 2018 nicht vollständig ist. Die diachrone Sentiment-Analyse wurde sowohl für komplette Artikel, als auch für einzelne Sätze aus der 'taz' vorgenommen. Die verwendeten Sätze für die Satz-Sentiment-Analyse werden anhand einer Schlüsselwörterliste selektiert. Die Schlüsselwörterliste erweitert das Stichwort 'Flüchtling' um andere Ausdrücke und Formulierungsmöglichkeiten, die mit 'Flüchtling' assoziiert werden, Wörter wie: 'geflüchtet' oder 'Syrer'.

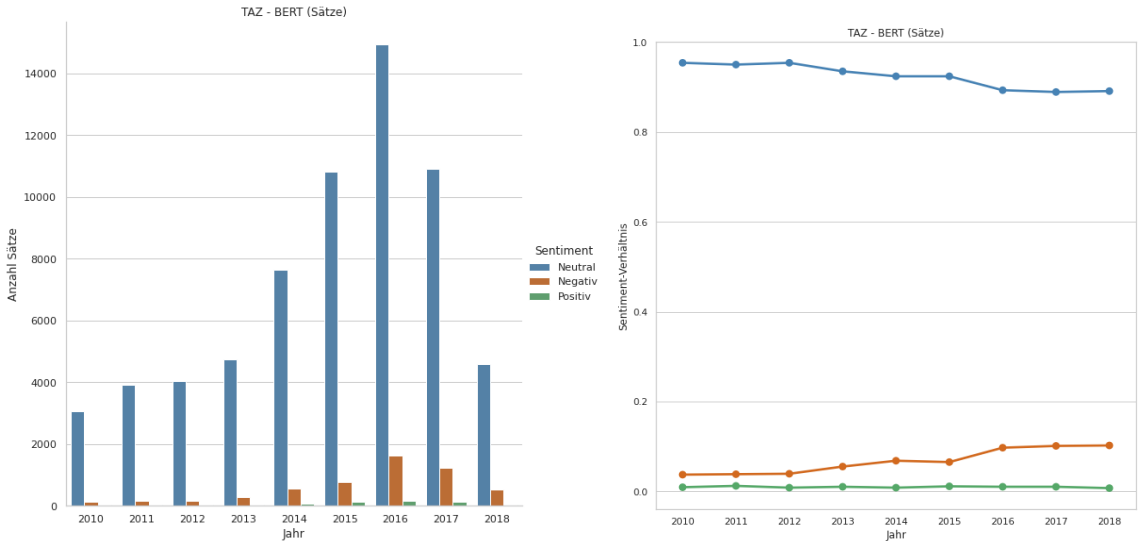

Abbildung 3: Links im Bild sehen wir die Anzahl an Sätzen pro Jahr. Jeder Artikel wurde einem Label zugeordnet. Neutral ist blau gekennzeichnet, negativ orange und positiv ist grün markiert. Auf der rechten Seite im Bild sehen wir die dazugehörigen Verhältnisse pro Jahr. Auch hier ist über die Jahre der Anteil der negativen Sätze gewachsen.

Abbildung 3 zeigt die Verhältnisse der Sätze auf. Die Verhältnisse lehnen sich eher an die Resultate der Artikel, als an deren Titel. Auch hier sehen wir einen leichten Trend für vermehrtes Auftauchen von negativen Sätzen im Laufe der Zeit. Mit den Ergebnissen lässt sich sagen, dass GSBERT insgesamt eine schwache Tendenz zum Negativen detektiert hat. Nun schauen wir uns die Ergebnisse von SentiWS an.

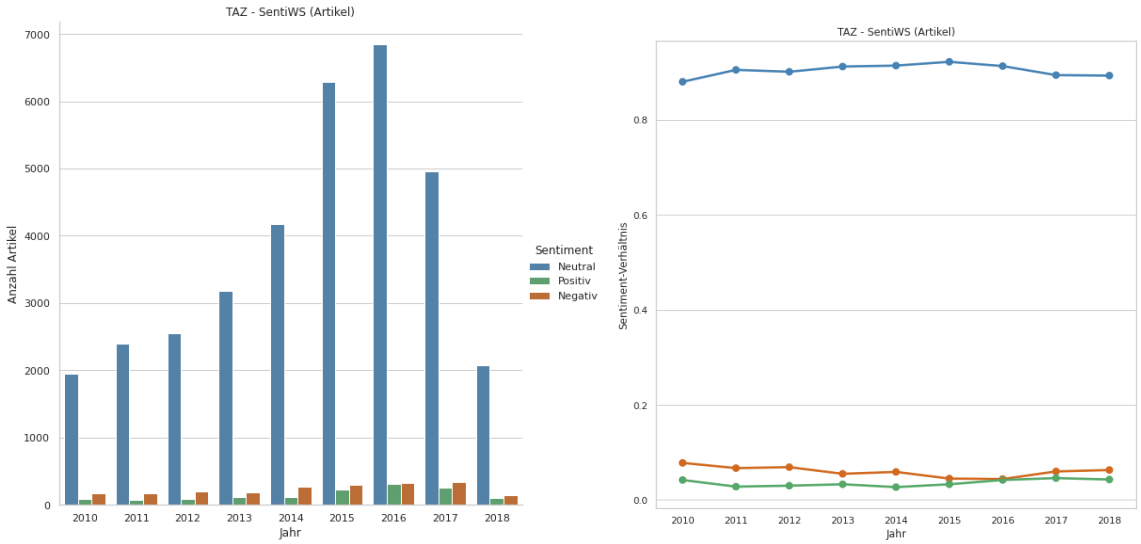

Abbildung 4: Links im Bild sehen wir die Anzahl an Artikeln pro Jahr. Jeder Artikel wurde einem Label zugeordnet. Neutral ist blau gekennzeichnet, negativ orange und positiv ist grün markiert. Auf der rechten Seite im Bild sehen wir die dazugehörigen Verhältnisse pro Jahr. Über die Jahre ist der Anteil der Labels relativ stabil geblieben.

Abbildung 4 zeigt die Verhältnisse der Artikel, deren Sentiment mit SentiWS berechnet wurde. Hier fallen zwei Sachen auf. Es gibt viel mehr positiv kategorisierte Artikel und deutlich weniger negativ kategorisierte Artikel als bei GSBERT. Zusätzlich lässt sich hier kein wirklicher Trend erkennen. Die Verhältnisse verändern sich nur minimal über den kompletten Zeitraum.

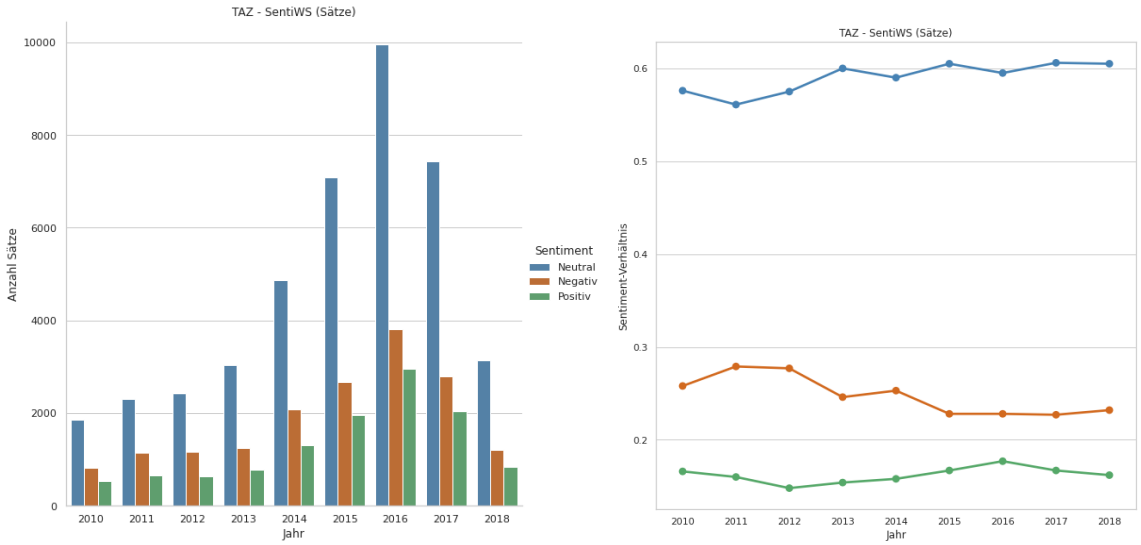

Abbildung 5: Links im Bild sehen wir die Anzahl an Sätze pro Jahr. Jeder Artikel wurde einem Label zugeordnet. Neutral ist blau gekennzeichnet, negativ orange und positiv ist grün markiert. Auf der rechten Seite im Bild sehen wir die dazugehörigen Verhältnisse pro Jahr. Über die Jahre ist der Anteil der negativen Artikel gesunken.

Schauen wir uns die Sentiment-Verhältnisse von Sätzen in Abbildung 5 an. Die Methode die SentiWS nutzt, scheint durchschnittlich weniger Sätze als neutral zu kategorisieren als GSBERT. Zudem scheint es im Gegensatz zu den Ergebnissen von GSBERT einen positiven Trend zu geben. Aber wieso sind die Resultate so unterschiedlich? Sind die angepassten Gewichte falsch berechnet worden? Um das zu prüfen, vergleiche ich auf der nächsten Seite die Anzahl an genutzten Wörtern pro Satz mit der jeweiligen Gewichtung und der angepassten Gewichtung.

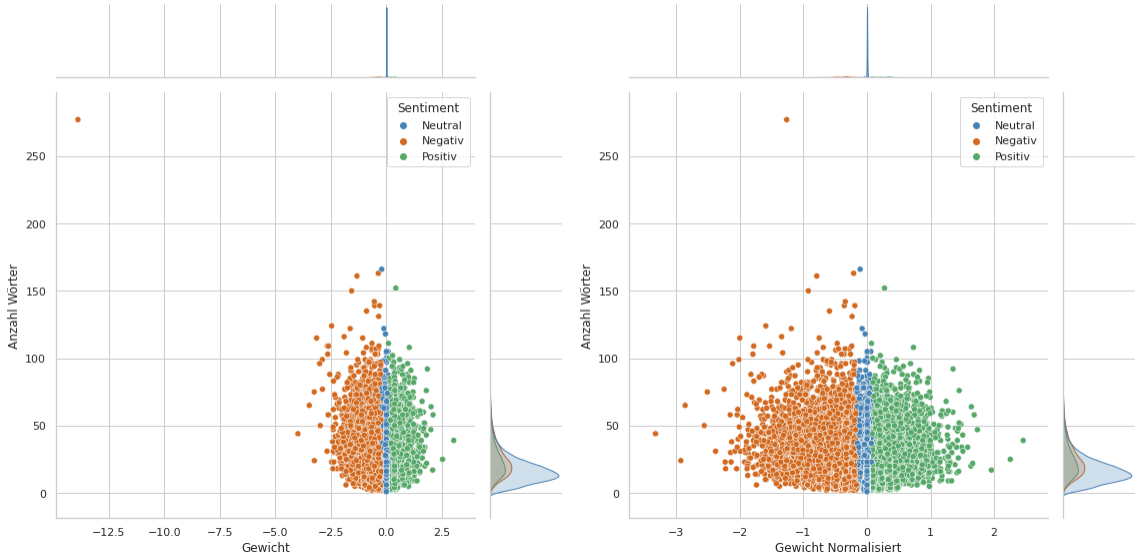

Abbildung 6: In beiden Bildern sehen wir die Verteilung der gelabelten Sätze. Links werden die Sätze mit der Anzahl an Wörtern ins Verhältnis zu ihrem ermittelten Gewicht gestellt. Rechts steht die Anzahl der Wörter im Verhältnis zu dem angepassten Gewicht. Die Kurven, die sich oben und rechts an den Grafiken befinden, zeigen die Dichte der Daten an.

In Abbildung 6 können wir beide Gewichte nebeneinander sehen. Der Grafik nach scheint die Normalisierung der Gewichte nicht fehlerhaft zu sein. Der längste Satz mit über 250 Wörtern hatte zuvor mit Abstand das kleinste Gewicht. Im angepassten Gewicht befindet er sich im 'negativen Mittelfeld'. Dass explizit dieser sehr lange Text nicht zureichend in seine Sätze zerlegt wurde, ist nicht auszuschließen. Die Anpassung scheint aber zu funktionieren, also muss es einen anderen Grund geben. Zuerst muss beachtet werden, dass SentiWS von sich aus nur eine Möglichkeit bietet, neutral zu kategorisieren. Wie anfangs erwähnt, habe ich eine Toleranz eingebaut. In Abbildung 6 lassen sich die neutralen Sätze anhand eines schmalen blauen Streifens erkennen. Dieser Streifen lässt sich mithilfe der Toleranzgrenze breiter oder schmaler machen, wie wir in Abbildung 7 auf der nächsten Seite sehen können. Hier wird auch die Anzahl der Wörter ins Verhältnis zu den Gewichten gesetzt, jedoch für komplette Artikel.

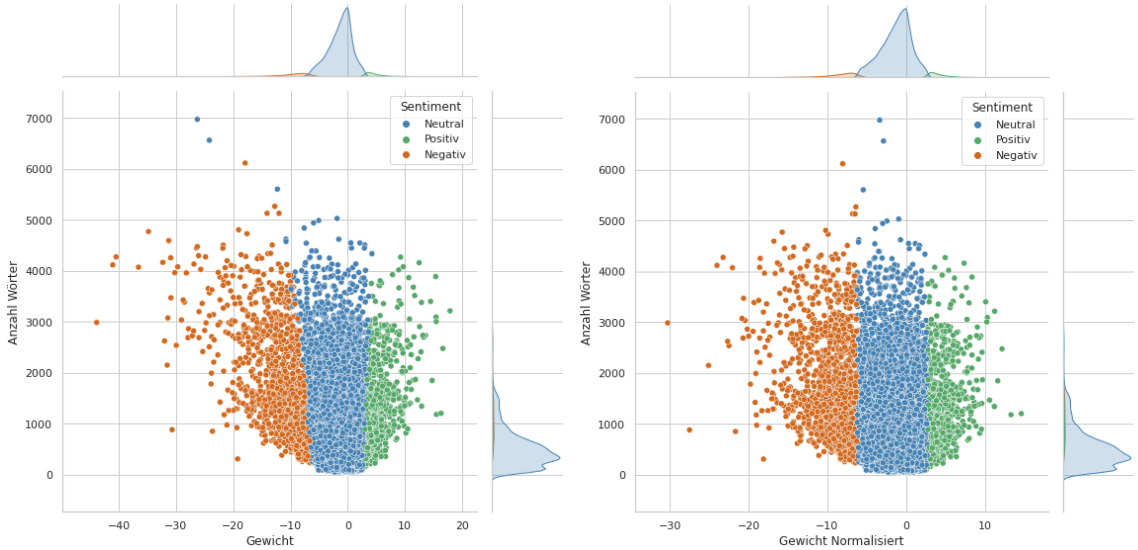

Abbildung 7: In beiden Bildern sehen wir die Verteilung der gelabelten Artikel. Links werden die Artikel mit der Anzahl an Wörter ins Verhältnis zu ihrem ermittelten Gewicht gestellt. Rechts steht die Anzahl der Wörter im Verhältnis zu dem angepassten Gewicht. Die Kurven, die sich oben und rechts an den Grafiken befinden, zeigen die Dichte der Daten an.

Der blaue Streifen in Abbildung 7 ist deutlich breiter, da die Toleranzgrenze größer ist. Die Toleranzgrenze muss größer werden, weil die Gewichte größer geworden sind. Natürlich beruht die verwendete Toleranzgrenze auf meiner theoretischen Herangehensweise an dieses Problem und die Parameter benötigen gegebenenfalls Feinjustierung. Solch ein Optimierungsprozess erfordert jedoch sehr viel Zeit und der Zeitrahmen meiner Arbeit lässt solch einen Prozess nicht zu. Somit könnte es sein, dass die Toleranzgrenze falsch justiert ist. Es könnte aber auch sein, dass die Ergebnisse tatsächlich valide sind. Im Bereich maschinelles Lernen hatte ich den Eindruck, dass viele Forscher dazu neigen, solange an den Parametern zu drehen, bis das Ergebnis herauskommt, das gewünscht war. So laufe auch ich Gefahr, mich bei der Justierung der Toleranzgrenze zu sehr an den Ergebnissen von GSBERT zu orientieren. Und da ich in dieser Arbeit nicht prüfen kann, wie nah SentiWS an GSBERT herankommt, werde ich darauf nicht weiter eingehen.

Sehen wir uns noch einmal Abbildung 5 auf Seite 20 an. Im Vergleich zu GSBERT in Abbildung 3 auf Seite 18 sehen wir hier einen entgegengesetzten Trend. Die Sätze sind seltener negativ. Ich könnte mir vorstellen, dass einzelne ziemlich stark gewichtete Wörter von der ‘taz’ seltener benutzt wurden und sich deshalb der negative Anteil verringert. Das könnten zum Beispiel Wörter sein wie ‘Gefahr’ mit einem Gewicht von -1, oder ‘bedenklich’ mit einem Gewicht von -0.77. Da der negative Anteil der Sätze weniger geworden ist, gilt es herauszufinden, ob dafür zumindest die Intensität der negativen Beiträge gestiegen ist.

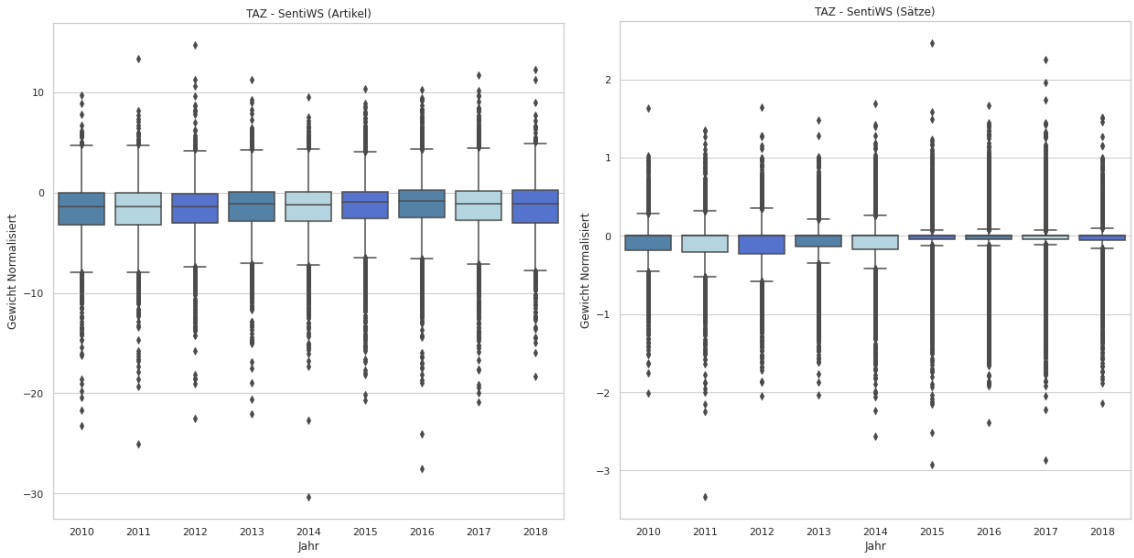

Abbildung 8: Hier sehen wir zwei Boxplots mit den ermittelten normalisierten Gewichten pro Jahr. Links stehen die Artikel, rechts stehen die Sätze. Die Verteilung der Werte bleibt über die acht Jahre relativ stabil.

In Abbildung 8 sehen wir zwei Boxplots mit den angepassten Gewichten pro Jahr. Auf der linken Seite, bei den angepassten Gewichten zu Artikeln, sehen wir minimale Veränderungen über den gesamten Zeitraum. Auf der rechten Seite, bei den angepassten Gewichten zu Sätzen, sehen wir hingegen, dass ab 2015 das Gewicht durchschnittlich näher bei Null liegt. Alles in allem sind die Veränderungen aber so klein, dass sie nicht auf eine Veränderung innerhalb des Overton-Fensters hindeuten können. Das Experiment zeigt damit auf, dass in 8 Jahren kaum Veränderungen in dem Sentiment zu den Texten der ‘taz’ in Bezug auf das Stichwort ‘Flüchtling’ erkennbar sind. Zudem sehen wir zwischen GSBERT und SentiWS widersprüchliche Trends. Da nicht evaluiert werden kann, welches Verfahren zuverlässigere Ergebnisse zu den ‘taz’ Daten liefert, können diese Trends nicht weiter interpretiert werden. Die gemessenen Veränderungen innerhalb eines Verfahrens sind klein und können damit nur schwer auf das Overton-Fenster übertragen werden. Es lässt sich zumindest sagen, dass dieses Verfahren mit diesen Daten und diesem Stichwort keine Möglichkeit bot, um Veränderungen innerhalb des Overton-Fensters zu beobachten.

4.2.3 Limitationen

Die Ergebnisse des ersten Versuchs sind nicht auf Veränderungen innerhalb des Overton-Fensters übertragbar. Mit dem Stichwort 'Flüchtling' hat sich sowohl das Sentiment von Sätzen als auch von kompletten Artikeln von 2010 bis 2018 bei der 'taz' kaum verändert. GSBERT zeigt eine Tendenz, dass mit steigendem Jahr häufiger negative Artikel und Sätze veröffentlicht wurden. Diese Tendenz ist jedoch sehr schwach. Das SentiWS Verfahren hingegen entdeckt keinerlei Veränderung bei den Artikeln und einen positiven Trend bei den Sätzen. Aber auch hier sind die Zahlen sehr klein. Zum einen ist kein Wandel in der Intensität des Sentiments zu sehen, zum anderen könnte die Toleranzgrenze inakkurat sein. Es wäre dementsprechend eine Möglichkeit, dass die verwendeten Methoden nicht geeignet sind, um Veränderungen aufzudecken. Die Ergebnisse können aber auch andeuten, dass es tatsächlich keine Veränderungen innerhalb des Overton-Fensters zu dem Thema 'Flüchtling' gegeben hat.
Für diese Möglichkeit spricht, dass die 'taz' über die Jahre hinweg zum Thema 'Flüchtling' eine relativ neutrale Berichterstattung praktiziert hat. Da jedoch unklar ist, welche Möglichkeit eingetreten ist, sollten Schlussfolgerungen zu dem Overton-Fenster, in Bezug auf dieses Thema, mit Vorsicht getroffen werden.

Ein weiteres Detail ist das Training von GSBERT. Wie wir in Tabelle 1 auf Seite 9 sehen können, stammen fast alle Trainingsbeispiele für neutrale Texte aus Wikipedia. Drei der Sammlungen aus Tabelle 1 haben alle ein ähnliches Problem wie SentiWS, es existieren kaum neutral gelabelte Texte. Es ist möglich, dass dementsprechend ein 'bias' vorhanden ist. Ich möchte nicht ausschließen, dass Zeitungsartikel eine gewisse Art von Satzbau nutzen, der dem aus Wikipedia sehr ähnelt. Wenn das der Fall sein sollte, dann würde das sicherlich die Klassifizierungsentscheidung von GSBERT in Richtung neutral lenken.
Einen ähnlichen 'bias' könnte auch das SentiWS Verfahren haben, da SentiWS mit 1.828 negativen und 1.645 positiven Wörtern scheinbar ein wenig mehr Möglichkeiten bietet, ein Wort mit negativem Gewicht zu finden. Ich will auch nicht ausschließen, dass es sicherlich noch elegantere Lösungen gibt, um mit SentiWS Neutralität zu definieren.
Zuletzt muss beachtet werden, dass Ironie und Sarkasmus auch weiterhin nur sehr schwer von Algorithmen erkannt werden können. Dies könnte auch einen Einfluss auf die Resultate gehabt haben.

4.3 Experiment 2 - Sentiment-Reaktion-Korrelation-Analyse

4.3.1 Ablauf

In dem zweiten Experiment richte ich den Fokus auf die Plattform Twitter. Twitter kann als ein soziales Netzwerk angesehen werden, das sich auf die Erstellung und Verbreitung von kurzen Nachrichten, sogenannte 'Tweets', spezialisiert hat. Twitter erfreut sich seit seiner Gründung im Jahr 2006 einer rasch wachsenden Popularität und 2021 gab es weltweit ungefähr 211 Millionen aktive Nutzer (Firsching, 2021). Das macht Twitter zu einer interessanten Quelle für die Studie des Overton-Fensters. In diesem Experiment möchte ich herausfinden, ob Tweets, die besonders viel 'Reaktionen' auslösen, besonders negativ oder positiv formuliert wurden. Als Reaktion zählt in diesem Experiment die Anzahl an Kommentaren, die ein Tweet erhalten hat. Die Idee dahinter ist, dass ein Tweet mit starkem Sentiment eine erhöhte Anzahl an Reaktionen hervorrufen könnte. Das bedeutet, dass der Tweet eine große Diskussionsbereitschaft ausgelöst hat. Eine erhöhte Diskussionsbereitschaft ist wiederum ein Indiz dafür, dass der Tweet den Grenzen des Overton-Fensters sehr nahe kommt, wenn nicht sogar diese überschreitet. Die Hypothese ist, dass ein solcher Tweet potenziell in der Lage wäre, eine Verschiebung der Grenzen des Overton-Fensters auszulösen. Der Tweet könnte aber auch die Folge einer Verschiebung innerhalb des Overton-Fensters sein. Das Sentiment eines jeweiligen Tweets wird exakt gleich wie im ersten Experiment ermittelt. Somit verwende ich auch hier sowohl GSBERT als auch SentiWS und unterteile das Sentiment wieder in die drei Kategorien 'positiv', 'negativ' und 'neutral'. Die verwendeten Daten sind im August 2021 mithilfe eines Pythoncodes direkt von Twitter gecrawlt worden. Insgesamt wurde über einen Zeitraum von 17 Tagen 8,8 Millionen deutschsprachige Tweets gesammelt. Da der Crawler hin und wieder von Google Colab getrennt wurde, ist der Datensatz mit Lücken behaftet. Ich war jedoch der Ansicht, dass mit 8,8 Millionen Tweets die Lücken nicht allzu ausschlaggebend sein sollten und die Idee der vorgestellten Hypothese dennoch geprüft werden kann. Neben dem Inhalt einer Nachricht wurden noch viele weitere Metadaten erfasst. Mit diesen Metadaten wäre es möglich, alle Relationen zwischen Tweets, Nutzern und Kommentaren aufzudecken. Das ist für dieses Experiment aber nicht nötig. Es reicht aus, dass nachvollzogen werden kann, welcher Tweet welchen Tweet kommentiert hat, wie oft ein Tweet favorisiert wurde und wie oft ein Tweet weitergeleitet wurde. Nachdem die Relationen aufgedeckt und die Anzahl an Reaktionen pro Tweet berechnet wurden, wende ich eine Korrelation-Analyse in Bezug zu Sentiment und Reaktionszahl an. Damit möchte ich überprüfen, ob besonders positive oder negative Tweets tatsächlich eine erhöhte Anzahl an Reaktionen aufweisen und dadurch möglicherweise die Grenzen des Overton-Fensters verschoben haben.

4.3.2 Ergebnisse

Von den 8,8 Millionen deutschen Tweets sind auf 336.623 reagiert worden. Das bedeutet, dass etwa 8,5 Millionen Tweets überhaupt keine Reaktion ausgelöst haben. Im Schnitt wurde in diesem Datensatz auf einen von 26 Tweets reagiert. Der meist kommentierte Tweet hat 852 Kommentare erhalten. Diese Zahl sinkt rapide mit größer werdendem Rang. Insgesamt haben im kompletten Datensatz nur 2.265 Tweets mehr als zehn Kommentare erhalten. Das entspricht etwa 0,025% aller gesammelten Tweets. Damit die Korrelation-Analyse dadurch nicht verfälscht wird, nehme ich die Top 500.000 Tweets mit Reaktionen. Das bedeutet, dass 163.377 Tweets im Datensatz keine Reaktion erhielten. Mit dieser Menge nähere ich mich an eine Verteilung nach Zipfschen Gesetz.

Sentiment	Tweets	∅ Reaktion	∅ Sentiment-Wert	Reaktionen
Neutral	245760	0.672	4.714	0.001
Negativ	200120	0.69	2.974	0.009
Positiv	54120	0.616	2.736	0.001
Sentiment	**Tweets**	**Verifiziert**	**Anzahl Follower**	**Totale Reaktionen**
Neutral	245760	0.058	0.038	-0.026
Negativ	200120	-0.014	-0.009	-0.01
Positiv	54120	-0.01	-0.017	0.001

Tabelle 3: Hier sehen wir die Ergebnisse der Korrelation-Analyse mit den Resultaten von BERT. '∅ Reaktion' und '∅ Sentiment-Wert' sind keine Korrelationswerte, sondern entsprechen den durchschnittlichen Werten vom gesamten Datensatz.

In Tabelle 3 sehen wir die Korrelationswerte zu den ermittelten Sentiment-Werten von GSBERT. Die Korrelationen wurden entsprechend ihres Sentiments des jeweiligen Labels berechnet. Wie wir oben rechts sehen können, sind die Korrelationen zwischen Sentiment und Reaktion sehr klein. Nichtsdestotrotz ist die Korrelation bei negativen Tweets im Vergleich 9 Mal höher. Alle Werte befinden sich jedoch im Promillebereich. Lediglich bei der durchschnittlichen Reaktion können wir größere Unterschiede erkennen. Auf einen positiven Tweet folgen durchschnittlich 0,616 Kommentare. Das ist die niedrigste Rate von allen drei Kategorien. Mit durchschnittlich 0,69 Reaktionen auf einen negativen Tweet ist diese Reaktionsrate am höchsten. Knapp dahinter befinden sich neutrale Tweets mit einer Rate von 0,672. Auch interessant ist die Korrelation zwischen verifizierten Twitteraccounts und neutralen Tweets. Mit 0,058 ist das der größte Korrelationswert im gesamten Datensatz. Es scheint so, dass verifizierte Nutzer eher darauf bedacht sind, möglichst neutral aufzutreten. Das ist plausibel. Der niedrigste Korrelationswert erscheint in Verbindung mit den 'totalen Reaktionen' und neutralen Tweets. Totale Reaktionen beinhalten die Anzahl der Kommentare, die Anzahl an 'Favoriten' – das kann mit einem aus vielen sozialen Netzwerken bekannten 'Like' gleichgesetzt werden – und die Anzahl an 'Retweets', das der Weiterleitung einer Nachricht gleicht. Es symbolisiert also die allgemeine Popularität eines Tweets. Neutrale Tweets scheinen also weniger populär als andere Tweets zu sein. Aber auch hier sind die

Werte sehr klein. Alles in allem, scheint es aber ein kleines Indiz zu geben, dass negative Tweets tatsächlich mehr Kommentare erhalten als neutrale oder positive. Betrachten wir mal die einzelnen Tweets genauer:

Reaktionen	Label GSBERT	Nachricht	Label SentiWS	Gewicht
852	neutral	Ich bitte alle die meine Frage beim #Sommerinterview als unangemessen oder gar sexistisch aufgefasst haben aufrichtig um Entschuldigung. @ABaerbock hat ihre Kinder selbst mehrfach thematisiert. Ich bin auch Mutter und bedauere deshalb sehr dass dieser Eindruck entstanden ist.	negative	-0.49
847	neutral	Noch nie sind so viele Plakate von mir zerstört oder verschmiert worden. Eine kleine Gruppe selbstgerechter Menschen, die nicht zwischen berechtigter Fürsorge und Angst unterscheiden können.	negative	-1.19
692	neutral	Als welchen Charakter bzw. Rolle seht ihr mich in GTA RP ?	neutral	0
439	neutral	Wie lautet dein Filmtitel für die aktuelle Lage in diesem Land?	neutral	0
427	negative	13:00 Uhr Verbotener #Querdenker Aufzug zieht lautstark durch Prenzlauer Berg, blockiert zeitweise die Danziger Straße mit (geschätzt) 700 Personen. Kollektiv ohne MNS @rbb24 #b2808 https://t.co/7S0kdyLoEs	negative	-0.66

Tabelle 4: In der Mitte sehen wir den Inhalt der Tweets mit den meisten Reaktionen. Links daneben sehen wir die Anzahl an Reaktionen und die Kategorisierung von GSBERT. Rechts sehen wir die Kategorisierung von SentiWS, samt dem ermittelten Gewichten.

In Tabelle 4 sehen wir die top Tweets mit Reaktionen zusammen mit den berechneten Labels beider Verfahren. GSBERT hat die ersten zwei Tweets als neutral eingestuft. SentiWS hat hingegen beide als negativ eingestuft. Der Inhalt des ersten Tweets scheint eine Entschuldigung über eine vorher getroffene Aussage darzustellen, der des zweiten eine Behauptung über Vandalismus. Bei letzterem ist die negative Einstufung von SentiWS aber durchaus nachvollziehbar, folgt der Behauptung doch sogleich eine herablassende Bemerkung. Hier könnte man die Einstufung von GSBERT hinterfragen. Doch es scheint auch so, dass keine der Aussagen an die 'Grenzen des Sagbaren' und damit an die Grenzen des Overton-Fensters gelangt. Dasselbe kann man zu Tabelle 5 auf der nächsten Seite sagen.

Reaktionen	Label GSBERT	Nachricht
427	negativ	13:00 Uhr Verbotener #Querdenker Aufzug zieht lautstark durch Prenzlauer Berg, blockiert zeitweise die Danziger Straße mit (geschätzt) 700 Personen. Kollektiv ohne MNS @rbb24 #b2808 https://t.co/7S0kdyLoEs
391	negativ	Kennt ihr das? Man erträgt die Berichte über Afghanistan nicht, erträgt Tweets dazu nicht, erträgt keine Nachrichten über den Horror. Gefühl von Ohnmacht und Scham, weil man selbst nichts tun kann. Und doch hat man die ganze Zeit das Handy in der Hand u. liest alles. Und leidet.
278	negativ	Bundesregierung: „Wir haben die Lage falsch eingeschätzt.” Welche? Corona, Wirecard, Flutkatastrophe oder Afghanistan?
146	negativ	"Wenn ihr eine Twitter-Floskel verbieten könntet, welche wäre das bei euch? Ich hasse Was macht das mit euch? und Weiß man da schon Genaueres? in etwa gleich.Ïch hasse Was macht das mit euch? und Weiß man da schon Genaueres? in etwa gleich."
143	negativ	Eure Meinung? https://t.co/wqnYO1Go6T

Tabelle 5: Hier sehen wir die ersten fünf negativen Tweets mit den meisten Reaktionen. Die Anzahl an Reaktionen sinkt rapide.

In Tabelle 5 sehen wir die top negativen Tweets kategorisiert von GSBERT. Auch hier scheint keine außergewöhnliche Aussage getroffen worden zu sein. Die Inhalte dieser Tweets sind sicherlich nicht schön und die Einstufung des negativen Sentiments ist nachvollziehbar, aber radikal oder extrem erscheinen mir diese Tweets nicht. Ich möchte aber betonen, dass ich kein Experte auf diesem Gebiet bin und für eine fundierte Einschätzung der Texte Experten herangezogen werden müssen.

Schauen wir uns daraufhin dieselbe Analyse mit dem SentiWS Verfahren an. Hier war es zusätzlich möglich eine Korrelation zu allen Kategorien zu berechnen, da alle Labels den gleichen Sentiment-Raum teilen.

Sentiment	Anzahl Tweets	∅ Reaktion	∅ Gewicht	Reaktionen
Neutral	275823	0.635	-0.003	-0.008
Negativ	98661	0.741	-0.532	-0.003
Positiv	125516	0.702	0.356	0.017
Gesamt	500000	0.673	-0.017	-0.002
Sentiment	Anzahl Wörter	Gewichtete Wörter	Negative Wörter	Positive Wörter
Neutral	-0.115	-0.13	-0.482	0.01
Negativ	-0.177	-0.295	-0.58	-0.036
Positiv	0.121	0.198	-0.06	0.218
Gesamt	-0.108	-0.067	-0.624	0.179
Sentiment	Verifiziert	Anzahl Follower	Anzahl Freunde	Totale Reaktionen
Neutral	0.008	0.004	0.002	-0.001
Negativ	0.016	0.007	0	0.005
Positiv	-0.004	0	0.008	0.039
Gesamt	0.001	0	0	0.028

Tabelle 6: Hier sehen wir die Ergebnisse der Korrelation-Analyse mit den Resultaten von SentiWS. '∅ Reaktion' und '∅ Gewicht' sind keine Korrelationswerte, sondern entsprechen den durchschnittlichen Werten vom gesamten Datensatz.

In Tabelle 6 erkennen wir zuerst, dass deutlich weniger Tweets negativ kategorisiert wurden und deutlich mehr Tweets positiv eingestuft wurden. Jedoch ist auch der neutrale Anteil gewachsen. Das ist tatsächlich bemerkenswert, wenn man die Ergebnisse mit denen der 'taz' vergleicht. Es scheint, dass Bert in Tabelle 3 auf Seite 26 die Tweets weitaus negativer betrachtet als die Sätze aus den Zeitungsartikeln (siehe Abbildung 3, Seite 18). SentiWS bleibt hingegen relativ stabil in seiner Verteilung. Im Vergleich zu Abbildung 5 auf Seite 20 sind die Verhältnisse ziemlich ähnlich, wobei Positiv- und Negativpositionen getauscht haben. Oben rechts in Tabelle 6 sind wieder die Korrelationswerte für Reaktionen zu sehen. Diese Werte sind ähnlich klein wie in Tabelle 3. Schauen wir uns die Werte für die Anzahl der Wörter, gewichteten Wörter und negativen Wörter an. Sie wirken zunächst kontraintuitiv. Wieso stehen negative Wörter in Antikorrelation zu negativen Tweets? Die Erklärung ist recht simpel. Ein negatives Sentiment ist kleiner als Null und hat damit einen negativen Wert. Die Anzahl an negativen Wörtern ist hingegen positiv. Dadurch, dass die Anzahl an negativen Wörtern besonders häufig mit negativen Sentiment-Werten auftritt, entsteht eine mittelstarke Antikorrelation. Die Werte entsprechen also mehr oder weniger den Erwartungen. Positive Wörter haben insgesamt eine schwächere Korrelation mit positiven Tweets. Ich gehe davon aus, dass es auf die unterschiedliche durchschnittliche Gewichtung von SentiWS zurückzuführen ist. Negative Wörter haben ein Gewicht im Durchschnitt von -0,16. Das ist mehr als das doppelte negative Gewicht gegenüber positiven Wörtern mit einem Durchschnitt von 0,07. Das passt auch in etwa auf das Verhältnis zwischen den zwei Korrelationen.
Tabelle 6 zeigt auch, dass die durchschnittliche Reaktionsrate bei negativen Tweets mit 0,741 wieder am höchsten ist. Dieses Mal besetzen die Positiven den zweiten Platz und die Neutralen sind

an letzter Stelle. Dementsprechend haben Tweets mit negativem Sentiment, auch bei dem SentiWS Verfahren, scheinbar auch eine höhere Wahrscheinlichkeit einen Kommentar zu bekommen. Das bestätigen nun beide Analyseverfahren. Die Idee, dass besonders negative oder besonders positive Tweets mehr Kommentare erhalten als andere, kann mit dieser Analyse jedoch nicht nachgewiesen werden.
Das größte ermittelte normalisierte Sentiment-Gewicht liegt im Datensatz bei 3,55. Das kleinste Gewicht liegt bei -8,29. Sehen wir uns als letztes die Tweets mit dem kleinsten Gewichten im Detail an.

Reaktionen	Label	Gewicht	Nachricht
0	negativ	-8.29	Die Corona Zwangsmaßnahmen sind Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht.
0	negativ	-3.21	Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht Unrecht. https://t.co/KB03DO2oKX
3	negativ	-3.68	@marykay05994602 @Doblerin @ebonyplusirony Beweise? Gut, aus eigenen Aussagen von Ihr: Schule schlecht, Lehrer und Rassismus schuld Karriere als StandUp gescheitert: Rassismus schuld Familiensituation: Die Männer schuld. Es sind immer andere schuld. Immer ne Ausrede Immer das Opfer. Und nun lies mal ihr Buch...
0	negativ	-3.46	Zweifach Geimpfter stirbt an Corona... die Ungeimpften sind Schuld Dreifach Geimpfter stirbt an Corona... die Ungeimpften sind Schuld Vierfach Geimpfter stirbt an Corona... die Ungeimpften sind Schuld ×-fach Geimpfter stirbt an Corona... die Ungeimpften sind Schuld
4	negativ	-3.42	@OnkelFester93 @DatMaXi @Bvborussia01 Das ist halt leider absolut kein Argument. Wenn eine Strecke ein absoluter Blindflug bei 300km/h wird, dann ist das ein unnötiges Risiko. Im Jahr 2021 sollte man so weit sensibilisiert sein, dass man einen Menschen keiner unnötigen Gefahr, die über die Norm hinausgeht, aussetzt.

Tabelle 7: Hier sehen wir die Tweets, die von SentiWS am negativsten eingestuft wurden.

Der erste Eintrag in Tabelle 7 hat ein deutlich kleineres Gewicht als die restlichen. Tatsächlich ist der negativste Tweet viermal von unterschiedlichen Nutzern kopiert worden. Diese vier Tweets wurden in der Tabelle zu einem zusammengefasst. Das Wort ‘Unrecht’ wurde 18 Mal in Folge verwendet. Mit einem Gewicht von -0,5086 gehört ‘Unrecht’ zu den stärker gewichteten Wörtern. Daraus ergibt sich jeweils das komplette Gewicht der ersten fünf Tweets, denn auch im zweiten

Eintrag in der Tabelle wird einfach nur ‘Unrecht’ wiederholt, aber ‘nur’ achtmal. Der angehängte Link bezieht sich auf eine Coronamaßnahme. Bei den nächsten drei Tweets ist jedoch weitaus mehr Inhalt zu finden. Es geht bei allen um heiß diskutierte Themen. Corona spielt dabei natürlich eine bedeutende Rolle, doch finden wir auch gesellschaftliche und politische Themen. Ein Sprachwissenschaftler könnte sicherlich einschätzen, ob hier extreme Aussagen zu finden sind.
Unabhängig davon sind es Themen, die von vielen verschiedenen Menschen unterschiedlich betrachtet und behandelt werden. Manche Menschen sind für die Coronamaßnahmen, manche dagegen. Die einen sind für ein Tempolimit auf Autobahnen und die anderen nicht. Manche verarbeiten Rassismus mit Humor, andere nicht. Es sind Themen, an denen sich die Gesellschaft stark reibt. Diese Themen sind im Endeffekt potenzielle Kandidaten, die eine etwaige Verschiebung der Grenzen des Overton-Fensters auslösen können.
Mir ist zudem aufgefallen, dass drei dieser Tweets selbst Kommentare auf andere Tweets sind. Schaut man sich die ersten 8000 aller negativen Tweets genauer an, sieht man, dass diese meist Kommentare sind, die je nach dem wieder kommentiert wurden. Es scheint also so, als sei das Ausschlaggebende nicht, wie oft auf einen Tweet kommentiert wurde, sondern ob der Tweet selbst etwas kommentiert. Daraufhin prüfte ich, ob es eine Korrelation zwischen dem Sentiment und dem Kommentar-Status gibt. Von den 500.000 Tweets sind 281.322 Kommentare. Es ist also gar nicht so selten, dass ein Tweet ein Kommentar auf einen anderen Tweet ist. Die Korrelation-Analyse hat jedoch einen Wert von 0,015 ergeben, also ist auch hier kein wirklicher Zusammenhang zu finden. Immerhin zeigen die Proben aus den Daten auf, dass es durchaus möglich sein könnte, dass die Inhalte von Tweets mit sehr negativen Sentiment eher starke Meinungen zu delikaten Themen haben. Damit befinden sich die Tweets potenziell an den Grenzen des Overton-Fensters.

4.3.3 Limitationen

Da ich von den Ergebnissen überrascht war, stellte ich Nachforschungen an. Wie sich herausstellt, ist über die Jahre die Nutzung von ‘hate speech’ - auf deutsch Hassrede - zu einem Problem für soziale Netzwerke geworden (de Gibert et al., 2018). Dies führte dazu, dass auf Twitter Kommentare in Ihrer Anzahl limitiert oder komplett ausgeschaltet werden können. Zudem kann ein Twitteraccount seine Tweets nur für ‘Follower’ sichtbar machen. Das heißt, dass man diesem Account zuerst eine Anfrage senden muss, ob man ihm ‘folgen’ darf, bevor man seine Tweets einsehen kann. Dadurch kann es sein, dass gewisse Aktivitäten auf Twitter überhaupt nicht einsehbar sind. Um das zu umgehen, bräuchte man einen Twitteraccount, der allen Accounts auf Twitter folgt. Das ist nicht nicht realisierbar und dadurch wird eine aussagekräftige Studie zu der Sentiment-Reaktion-Beziehung eines Tweets verhindert. Bauer (2021) hat außerdem eine interessante Abhandlung über die Mechaniken von Twitter geschrieben. Er zieht den Vergleich mit der Plattform und einem Videospiel. Er zeigt auf, dass viele Mechaniken aus Twitter tatsächlich sehr ähnlich zu Spielmechaniken aus Online-Multiplayer-Spielen sind. Soziale Netzwerke begünstigen die Entstehung von sogenannten ‘Bubbles’, Gruppierungen von Nutzern, welche ihre Ansichten energisch teilen und sich dabei gegenseitig unterstützen. Das ist an sich erst mal nichts Schlechtes, jedoch findet zwischen den Bubbles ein erbitterter subtiler Kampf um Nutzer statt. Er kommt letztlich zu dem Schluss, dass das momentane Design der Plattform Twitter und vieler anderer sozialen

Netzwerken, alle Gruppen von Nutzern dazu treibt, einander zu provozieren und auszuspielen. Das endet in einer endlosen, selbst nährenden Schleife des Konflikts, die den Gruppen scheinbar keine andere Wahl übrig lässt, als radikaler und extremer zu werden. Opuszko et al. (2018) kommen zu einem ähnlichen Schluss und auch Arruda et al. (2022) machen auf dieses Thema aufmerksam. Insgesamt wäre es auch schon bei einem deutlich kleineren Einfluss als bei Bauer geschildert notwendig, diesen in seinen Experimenten zu beachten. Dieser Einfluss ist sicherlich gegeben, rüsten sich doch immer mehr soziale Netzwerke mit selbst entworfenen Uploadfilter. Die Netzwerke sind sich vermutlich ihrer Auswirkungen in der Gesellschaft bewusst.
Mit dieser neuen Betrachtung auf die Plattform Twitter wirkt das Ergebnis nicht mehr so verwunderlich wie zuvor. Im Nachhinein scheint es also weniger vielversprechend, dass die Intensität eines Sentiments die Reaktionsbereitschaft erhöhen könnte. Es spielen viele andere Faktoren eine Rolle, zum Beispiel wer überhaupt einen Tweet sehen kann oder die Anzahl an 'Follower', die ein Nutzer hat. Dennoch wurde ein Indiz entdeckt, dass negative Tweets im Durchschnitt tatsächlich mehr Reaktionen erfahren als andere. Dass Tweets mit besonders intensivem Sentiment mehr Reaktionen erfahren, konnte jedoch nicht nachgewiesen werden. Dennoch gibt es zureichend Gründe anzunehmen, dass Twitter trotzdem geeignet ist, um Veränderungen im Overton-Fenster wahrzunehmen. Der verwendete Datensatz ist nur über drei Wochen gesammelt worden und ist lückenhaft. Mit moderner und stabil laufender Technik ist es sicherlich möglich, in einem weitaus größeren Zeitraum Tweets lückenlos abzufangen. Mit so einem Datensatz sollten durchaus mehr Tweets mit hoher Reaktionszahl auftauchen. Zudem scheint der Inhalt von besonders negativen Tweets mit wichtigen und viel diskutierten Themen verknüpft zu sein. Dieses Wissen könnte in zukünftigen Arbeiten angewendet werden.

Bisweilen sind zwei der drei Experimente mit eher ernüchternden Ergebnissen abgeschlossen worden. Beide Experimente bedienten sich eines Sentiment-Ansatzes und verwendeten dabei GSBERT und SenitWS. Das Overton-Fenster ist jedoch mit beiden Ansätzen weder mess- noch greifbar geworden. Es hat den Anschein, dass Sentiment-Analysen nicht zwingend die richtigen Verfahren darstellen, um Veränderungen innerhalb des Overton-Fensters zu entdecken. Es scheint somit sinnvoll, keine weiteren Sentiment-Analysen mehr vorzunehmen und sich einer anderen Methodik zu bedienen. Das führt uns zu dem nächsten Experiment.

4.4 Experiment 3 - Diachrone Satz-Ähnlichkeit-Analyse

4.4.1 Ablauf

Das dritte und letzte Experiment in dieser Arbeit wendet sich nun von der deutschen Sprache ab und richtet seinen Fokus auf den englischen Sprachraum. Um genau zu sein auf den US-amerikanischen Sprachraum. Ähnlich wie in Deutschland, mit einer vereinfachten Ansicht der politischen Haltung in Richtung Rechts und Links, können auch dort die politischen Haltungen in zwei Richtungen eingeteilt werden. So kann eine vereinfachte amerikanische Einteilung auch in Richtung 'republikanisch' und 'demokratisch' stattfinden. Im Allgemeinen wird eine Neigung zu einer beliebigen politischen Haltung auch als 'Hyperpartisan' bezeichnet. Der Hyperpartisan ist dabei nicht zwingend in die zwei erwähnten Richtungen geteilt, sondern umfasst viel mehr alle möglichen politischen Neigungen. Um in diesem Experiment dem Overton-Fenster näher zu kommen, möchte ich herausfinden, ob Sätze aus bereits veröffentlichten Zeitungsartikeln von einer anderen Zeitung, die einer anderen politischen Haltung zugeordnet wird, wiederverwendet wurden. Jeder Zeitungsartikel im Datensatz ist einer politischen Haltung zugeordnet, was wiederum bedeutet, dass ein wiederverwendeter oder kopierter Satz potenziell von einer anderen politischen Haltung übernommen wurde. Die Idee ist also, dass gewisse Sätze oder Aussagen im Laufe der Zeit von anderen Zeitungen mit anderen politischen Haltungen wiederverwendet wurden. Meine Hypothese dahinter ist, dass die Wiederverwendung selbst eine direkte Folge der Veränderung der Grenzen des Overton-Fensters darstellt. Als Datensatz verwende ich den von 'webis.de' bereitgestellten 'PAN-SemEval-Hyperpartisan-News-Detection-19' (Kiesel et al., 2019). Er ist ungefähr 1,2 GB groß und besteht aus 750.000 Zeitungsartikeln und ist ursprünglich als Datensatz für maschinelles Lernen gedacht, um ein Modell darauf zu trainieren, einen zugehörigen Hyperpartisan aus einem beliebigen Text zu ermitteln. Deshalb enthält jeder Artikel in dem Datensatz eine Kennzeichnung, welcher politischen Neigung er angehört. Das Hyperpartisan-Label teilt sich in diesem Datensatz jedoch nur in fünf Kategorien ein. Die Kategorien sind 'left', 'left-center', 'least', 'right-center' und 'right'. Least bedeutet hierbei 'am wenigsten einem Hyperpartisan angehörig' und kann mit einer politischen Mitte assoziiert werden. Die Daten der gesamten Zeitungsartikel gehen von 1997 bis 2018 und umfassen somit 21 Jahre.
Für meine Analyse werden zuerst alle Artikel nach Schlüsselwörtern durchsucht. Es wurden insgesamt 4 Sets an Schlüsselwörtern verwendet. Diese betreffen die Themen 'refugee', 'vaccine', 'gun-law' und 'healthcare'. Sobald ein Schlüsselwort gefunden wurde, wird der Artikel in Sätze zerlegt und nochmals werden alle Sätze nach den Schlüsselwörtern durchsucht. Am Ende bleiben alle Sätze, die mindestens ein Schlüsselwort enthalten, übrig. Diese Sätze werden bereinigt und SBERT übergeben. SBERT gibt daraufhin eine Liste mit Vektoren aus. Jeder Vektor hat 383 Einträge und repräsentiert einen Satz. Wie schon erwähnt, sorgt SBERT dafür, dass Vektoren aus semantisch ähnlichen Sätzen nah beieinander im Vektorraum liegen. Das ermöglicht das Vergleichen der Vektoren mithilfe der Kosinus-Ähnlichkeit. Als Ergebnis erhalte ich eine Matrix mit Ähnlichkeitswerten zu allen Sätzen. Je nach Anzahl an gefilterten Sätzen kann die Größe der Matrix schnell die Kapazität des verfügbaren RAMs übersteigen. So ist es unbedingt notwendig 'sparse' Matrizen zu verwenden. Sparse bedeutet so viel wie spärlich und meint in diesem Kontext eine

Matrix mit wenigen Einträgen. Eine Sparsematrix ist dementsprechend besonders nützlich, wenn die Matrix viele Einträge mit dem Wert 0 hat. Problem dahinter ist, dass die Kosinusähnlichkeit sehr selten den Wert 0 erreicht, sind doch zwei komplett entgegengesetzte Vektoren notwendig. Das macht den Sparse-Ansatz praktisch untauglich. Als Lösung musste ich einen Ähnlichkeitswert finden, der eine bedeutsame Ähnlichkeit repräsentiert, damit alle darunter liegenden Werte als eine Null gespeichert werden können.
Um ein Gefühl dafür zu bekommen, was der Ähnlichkeitswert ausdrückt, habe ich 7 Sets an Samples mit jeweils 50 Satzpaaren erstellt, die mit den Ähnlichkeitswerten zwischen 0,7 und 1,0 zufällig ausgesucht wurden. Die Sample-Sets steigern sich in 0,05 Schritten, damit ich eine insgesamt relativ ausgeglichene Anzahl an Ähnlichkeitswerten habe. Die 7 Samples-Sets wurden anschließend von mir von Hand analysiert und kategorisiert. Als ich bei den Ähnlichkeitswerten zwischen 0,90 und 0,95 angekommen war, fiel mir auf, dass die Daten praktisch unbrauchbar wurden. Das lag daran, dass gewisse Zeichen in den Datensätzen falsch in UTF-8 übersetzt wurden und dadurch die Satztrennung versagt hat. So verwarf ich die ersten von Hand gelabelten Samples und implementierte eine Lösung für die fehlerhafte UTF-8 Übersetzung. Doch die erste Iteration war nicht komplett nutzlos. Ich habe bis dato herausgefunden, dass Vergleichswerte kleiner als 0,75 nicht weiter beachtet werden müssen. Sätze mit einem kleineren Ähnlichkeitswert hatten überhaupt nichts mehr miteinander zu tun. Zudem ist mir aufgefallen, dass besonders kurze Sätze meist keinerlei Kontext beinhalten. Sie sind dementsprechend für diese Analyse nutzlos. Trotzdem sind die kurzen Sätze häufig mit anderen Sätzen semantisch ähnlich, weshalb sie tatsächlich ein Problem darstellen. Ich entschied mich deshalb dafür, Sätze mit weniger als 7 Wörtern aus dem Vergleich auszuschließen. Nach diesen Feinjustierungen erstellte ich ein weiteres Set an Samples. Diesmal entschied ich mich aber für 25 Satzpaare mit den Ähnlichkeitswerten 0,75 bis 1,0.

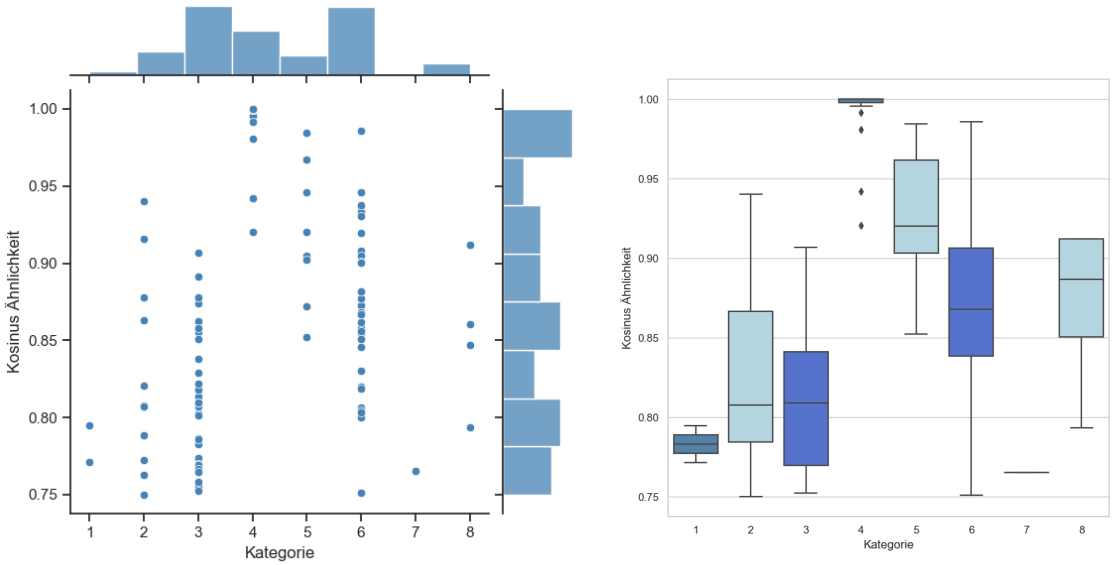

Abbildung 9: In dieser Grafik sehen wir die Ergebnisse der von Hand gelabelten Samples. Die Sätze wurden in acht Kategorien eingeteilt: (1) Fehlerhaft und nicht nutzbar, (2) Ähnliche Wörter aber ohne Kontext, (3) Gleiches Thema aber kein Kontext, (4) Kopie, (5) Kopie aber ein Satz hat mehr Informationen, (6) Gleiches Thema und neutraler Standpunkt, (7) Gleiches Thema, neutraler Standpunkt, aber ein Satz hat mehr Informationen, (8) Gleiches Thema, gleiche Meinung.

In Abbildung 9 sind die Ergebnisse zu den Samples zu sehen. Da die Kategorien 1 bis 3 für dieses Experiment nutzlos sind, erschien mir aufgrund dieser Grafiken ein Ähnlichkeitswert von 0,84 am geeignetsten. Alle Vergleichswerte darunter sind im Regelfall nicht hinreichend, um als wiederverwendeter Satz zu gelten.

Auf der nächsten Seite in Tabelle 8 sehen wir ein paar Beispiele, wie SBERT und ich die Sätze bewertet haben. Mit dem ermittelten Ähnlichkeitswert war es nun möglich eine Sparsematrix zu erstellen, das heißt, dass alle Werte kleiner als 0,84 als eine 0 eingetragen wurden. Die Einträge der Matrix werden daraufhin eingelesen und es wird eine Liste mit allen verglichenen Sätzen samt ihrer Metadaten erstellt. Diese Liste wird herangezogen, um herauszufinden, wann die Sätze veröffentlicht wurden und mit welcher politischen Haltung die Sätze assoziiert werden. Anschließend lässt sich auswerten, wer von wem einen Satz übernommen oder adaptiert hat. Das deutet wiederum darauf hin, dass sich eine Meinung in eine andere Richtung innerhalb des Overton-Fensters verschoben hat, was wiederum darauf hindeutet, dass sich die Grenzen verschoben haben.

Wert	Erster Satz	Zweiter Satz	Kategorie
0.764	[Left] There is another reason that this crisis is so severe: Politics within Europe are unusually hostile to refugees and migrants at the moment.	[Right] But the refugee crisis places an existential threat on European unity.	3
0.801	[Left] Here, 2million refugees outside the country, four and a half million inside the country.	[Left-Center]There are now more than 45.2 million displaced people — 15.4 million refugees, 937,000 asylum seekers and 28.8 million forced to flee within the borders of their own countries.	3
0.857	[Right-Center] Over 750,000 Palestinians were expelled from or fled the horrors of the militia-instigated war, and those who are still alive along with their descendants number over five million refugees.	[Left] In the course of the war, some 750,000 Palestinians became refugees.	6
0.912	[Right-Center] The U. S. and other nations refused to accept Jewish refugees, and the majority of the U. S. public supported that position.	[Left] The U. S. and other nations would not allow Jewish refugees in, and the majority of the U. S. public supported that position.	8
0.945	[Right-Center] Trump's order pauses America's entire refugee program for four months, indefinitely bans all those from war-ravaged Syria and temporarily freezes immigration from Iraq, Syria, Iran, Sudan, Libya, Somalia and Yemen.	[Right] Trump's order pauses America's entire refugee program for four months and indefinitely bans all those from war-ravaged Syria.	5
1	[Left-Center] This escalation is short-sighted and will lead to more dead civilians, more refugees, the strengthening of al-Qaeda and other terrorists, and a direct confrontation between the United States and Russia—which could lead to nuclear war.	[Right] This escalation is short-sighted and will lead to more dead civilians, more refugees, the strengthening of al-Qaeda and other terrorists, and a direct confrontation between the United States and Russia—which could lead to nuclear war.	4

Tabelle 8: Hier sehen wir Beispiele von den verglichenen Sätzen. 'Wert' entspricht der Kosinus-Ähnlichkeit.

4.4.2 Ergebnisse

Wie schon erwähnt, werden in diesem Versuch vier verschiedene Sets an Schlüsselwörter verwendet. Schauen wir uns als erstes die Ergebnisse mit dem Set ‘refugee’ an. Insgesamt wurden aus 750.000 Artikel 117.546 Artikel mit mindestens einem Schlüsselwort gefunden. Aus diesen Artikeln konnten 351.748 Sätze mit mindestens einem Schlüsselwort extrahiert werden. Jeder dieser Sätze wurde mit allen anderen Sätzen verglichen. Das ergibt eine Anzahl von über 123 Milliarden Vergleichen. Hieraus konnten 3.124.980 Satzpaare gefunden werden, die eine Kosinus-Ähnlichkeit von mindestens 0,84 haben. Manche dieser Satzpaare entstammen jedoch demselben Zeitungsartikel, weshalb diese nochmal gefiltert werden mussten. Das resultiert zu insgesamt 2.768.980 vergleichbaren Satzpaaren. Die meisten dieser Satzpaare besitzen jedoch dasselbe politische Label und nur 33.920 Satzpaare haben ein unterschiedliches Label. Das heißt, dass nur 1,22% aller vergleichbaren Sätze aus unterschiedlichen Zeitungen stammen. Es kommt noch hinzu, dass nicht jeder Zeitungsartikel einem Datum zugeordnet werden kann, was den Datensatz nochmals auf 28.238 Satzpaare verkleinert. Davon sind aber 4.368 Satzpaare am selben Tag erschienen, weshalb nicht herausgefunden werden kann, wer von wem kopiert hat. Letztendlich kommen wir damit auf insgesamt 23.870 vergleichbare Satzpaare mit dem Schlüsselwort ‘refugee’. Folgend verwende ich häufig die Begriffe ‘Links’, ‘Rechts’ und ‘Mitte’. Diese sollen sich jeweils auf die Labels innerhalb des Datensatzes beziehen. Sie drücken also die vom Datensatz eingestufte politische Haltung aus. Dementsprechend steht ‘Links’ für das Label ‘left’, ‘Mitte-Links’ für ‘left-center’, ‘Mitte’ für ‘least’, ‘Mitte-Rechts’ für ‘right-center’ und ‘Rechts’ für ‘right’.

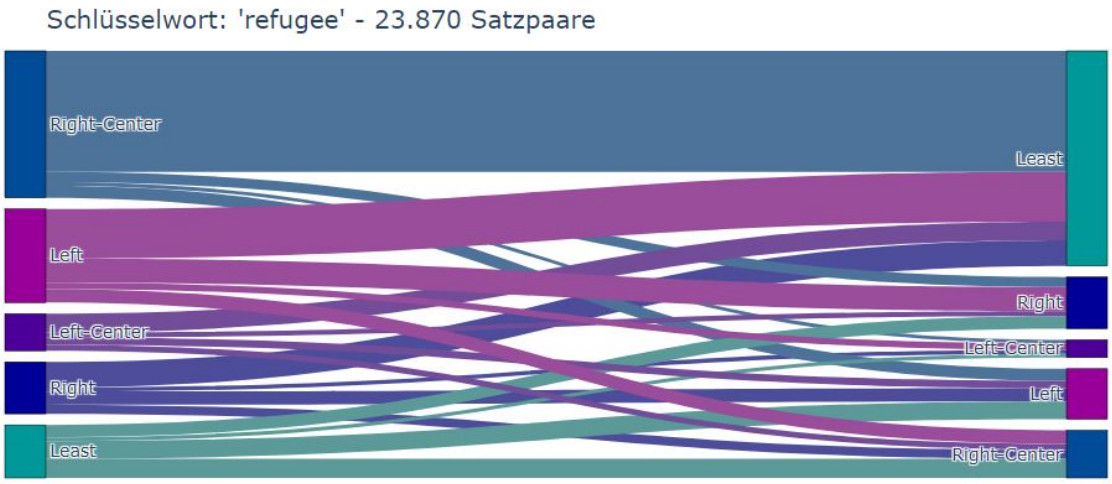

Abbildung 10: In diesem Sankey Diagramm zum Stichwort ’refugee’ sehen wir, welche Sätze von welchem politischen Label wieder verwendet wurden. Links sind die ursprünglichen Verfasser eines Satzes, rechts sind die Wiederverwender eines Satzes.

Abbildung 10 zeigt uns das Ergebnis zu den Satzpaaren und deren dazugehörigen, vom Datensatz eingestuften Hyperpartisan. Auf der linken Seite in der Grafik befinden sich die ursprünglichen Sätze und auf der rechten Seite befinden sich die Kopien. Im Durchschnitt wurde ein Satz nach 294 Tagen von einer anderen Zeitung wieder verwendet. Es wurden dabei 10.348 ‘linkere’ Sätze

von Rechts übernommen und 13.522 ‘rechtere’ Sätze von Links. Die mit Abstand größte Übernahme an Sätzen praktizierte die Mitte (Least). Ein Großteil dieser übernommenen Sätze entstammen aus Mitte-Rechts und Links. Interessant ist auch, dass Rechts mehr von Links übernommen hat als anders herum. Sätze, die der Mitte-Links entstammen, wurden am seltensten übernommen. Mitte-Links übernahm aber auch am wenigsten Sätze. Die Grafik zeigt eine deutliche Veränderung zu den Sätzen und ihrer assoziierten Zugehörigkeit einer politischen Neigung. Diese Veränderungen sind ein sehr schönes Indiz für den Wandel innerhalb des Overton-Fensters und der dynamischen Verschiebung seiner Grenzen. Natürlich gilt es zu beachten, dass diese Veränderungen im Verhältnis ein wenig mehr als nur ein Prozent des gesamten Datensatzes ausmachen. Das Overton-Fenster ist aber mutmaßlich auch kein Konstrukt, das sich schlagartig ändert und neu formt. Gesellschaftliche Veränderungen finden im Regelfall subtil und im kleinen Maßstab statt. Somit sind die Ergebnisse durchaus sehenswert und sehr spannend. Da mir das notwendige Wissen über die Vereinigten Staaten und die dort betriebene Politik fehlt, werde ich an dieser Stelle die Resultate nicht weiter interpretieren. Auch hierfür wären wieder Experten nötig, die ein fundamentales Wissen über die Politik und Gesellschaft haben. Ich möchte hier nur die Daten zeigen, auf Ihre Struktur und Muster aufmerksam machen und eine Verbindung zu dem Overton-Fenster herstellen.

Sehen wir uns also die Ergebnisse zu dem Schlüsselwort-Set ‘vaccine’ an. Hier wurden 119.160 Zeitungsartikel und damit 251.040 Sätze mit einem Schlüsselwort gefunden. Daraus ergaben sich 2.340.196 Satzpaare, von denen 2.066.086 aus verschiedenen Artikeln stammen. Hiervon sind wieder nur 17.030 mit einem unterschiedlichen Label klassifiziert worden. Davon haben 3.242 Paare keinen vollständigen Zeitstempel und 3.242 sind am selben Tag erschienen. Letztlich bleiben 10.178 Satzpaare zum Vergleich übrig.

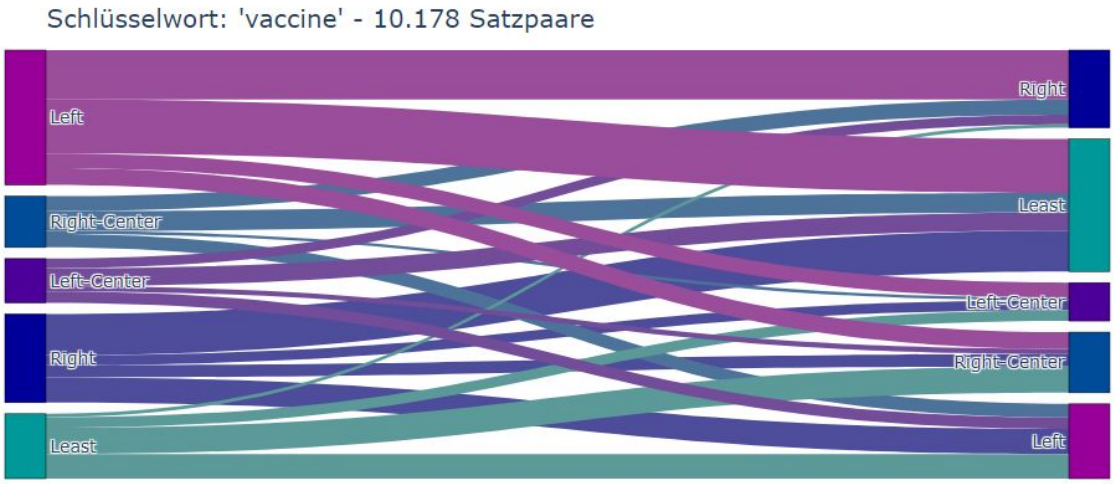

Abbildung 11: In diesem Sankey Diagramm zum Stichwort ’vaccine’ sehen wir, welche Sätze von welcher politischen Neigung wieder verwendet wurden. Links sind die ursprünglichen Verfasser eines Satzes, Rechts sind die Wiederverwender eines Satzes.

In Abbildung 11 sehen wir, dass die Verschiebungen weitaus ausgeglichener sind als zuvor. Mitte hat zu fast gleichen Anteilen von Links und Rechts adaptiert. Interessanterweise hat Rechts im

Verhältnis sehr viel von Links übernommen. Doch auch Links übernahm einen großen Anteil von Rechts. Mitte-Links scheint genau wie in Abbildung 10 auf Seite 37, am resistentesten gegenüber Veränderungen. Ein Satz wurde durchschnittlich im Abstand von 320 Tagen von einer anderen Zeitung übernommen. Es sind insgesamt 5.990 dieser Sätze von Links in Richtung Rechts gewandert und dementsprechend 4.188 in die andere Richtung. Hier hat also ein deutlicher Wandel in Richtung Links stattgefunden. Doch auch hier muss bedacht werden, dass die Veränderungen 0,82% der gesamten Daten ausmachen. Dass Mitte wieder einmal die meisten Sätze übernommen hat, ist sehr spannend. Es ist ein Indiz dafür, dass gewisse Aussagen, welche sich erstmals am Randbereich des Overton-Fensters befunden haben, weiter in das Zentrum gerückt sind. Sollten wir ein ähnliches Muster in den anderen Schlüsselwörter-Sets sehen, würde es das Indiz unterstreichen.

Das dritte Schlüsselwort-Set setzt sich mit dem Stichwort 'healthcare' auseinander. 302.366 Sätze aus 115.980 Artikel wurden gefunden. Von 2.943.752 Satzpaaren sind 2.618.590 aus unterschiedlichen Artikeln. Nur 0,77% davon haben ein unterschiedliches Label, wodurch 20.242 Satzpaare übrig bleiben. 5.858 dieser Satzpaare haben keinen Zeitstempel erhalten und 4.300 sind am selben Tag erschienen. Damit bleiben insgesamt 10.084 Satzpaare übrig.

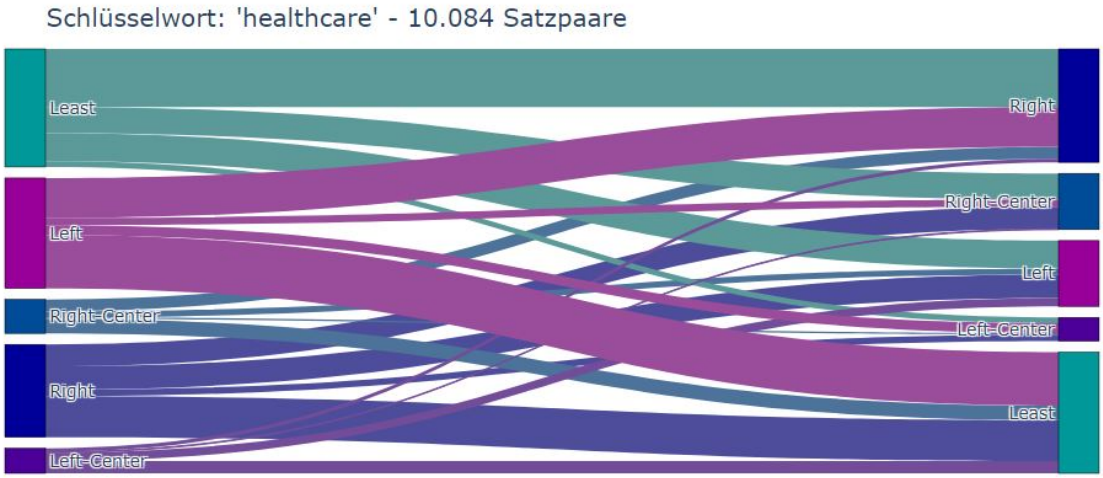

Abbildung 12: In diesem Sankey Diagramm zum Stichwort 'healthcare' sehen wir, welche Sätze von welcher politischen Neigung wieder verwendet wurden. Links sind die ursprünglichen Verfasser eines Satzes, Rechts sind die Wiederverwender eines Satzes.

Die Auswertung dieser Sätze können wir in Abbildung 12 betrachten. Wieder einmal hat die Mitte die meisten Sätze übernommen, wenn auch nur knapp. Bei genauerem Betrachten sieht es sogar so aus, als hätte es einen regen Austausch zwischen Rechts und Mitte gegeben. Links hat die meisten Sätze verbreitet, die praktisch fast nur von Rechts und Mitte übernommen wurden. Die durchschnittliche Zeit für eine Wiederverwendung beträgt in diesem Set mit 221 Tagen deutlich weniger als bei den anderen Sets.

Insgesamt wurden 5.918 Sätze von Links in Richtung Rechts und 4.166 von Rechts in Richtung Links verschoben. Auffallend ist auch hier, dass Mitte-Links kaum zu dem Austausch beiträgt.

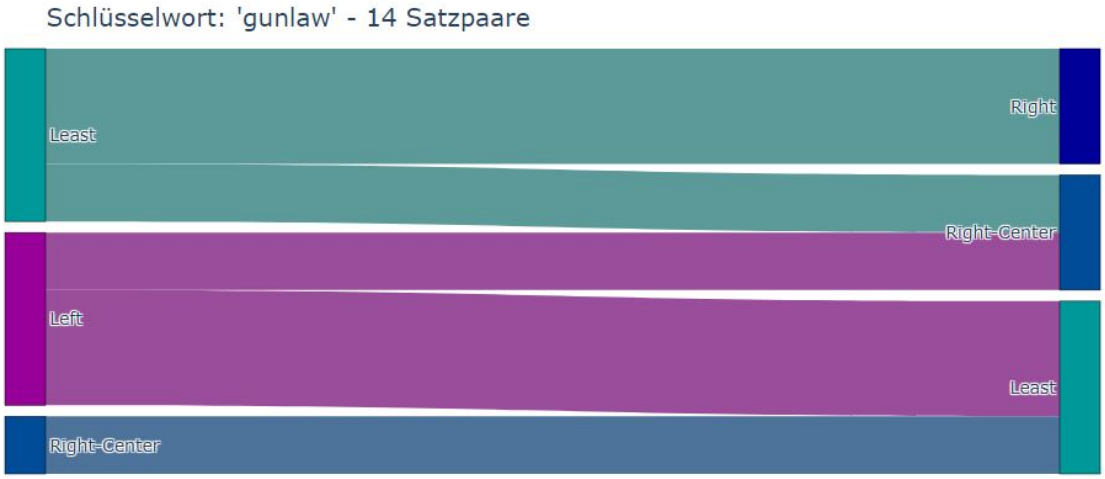

Abbildung 13: In diesem Sankey Diagramm zum Stichwort 'gunlaw' sehen wir, welche Sätze von welcher politischen Neigung wieder verwendet wurden. Links sind die ursprünglichen Verfasser eines Satzes, Rechts sind die Wiederverwender eines Satzes.

Schauen wir uns mit Abbildung 13 an, ob sich im letzten Set Mitte-Links auch genauso verhält. Mehr oder weniger überraschend ist die Mitte-Links diesmal überhaupt nicht vertreten. Das letzte Set umfasst das Stichwort 'gunlaw' und ist das mit Abstand kleinste Set von allen. Mit 841 Artikeln und 1.033 gefunden Sätzen scheint in diesem Datensatz das Thema Waffenkontrolle nicht besonders häufig vorzukommen. Nach allen Filterungen blieben gerade einmal 14 Satzpaare übrig und damit wurde diese Grafik erstellt. 12 Sätze wurden dabei von Links in Richtung Rechts verschoben und zwei von Rechts in Richtung Links. Niemand hat einen Satz von Rechts übernommen und Links hat von niemandem einen Satz übernommen. Auch wenn die Daten bei Weitem weniger Quantität haben, so lassen sich doch auch hier dieselben Beobachtungen wie bei den anderen Sets machen. Die Mitte hat die meisten Sätze übernommen, Mitte-Links hält sich so gut es geht aus der Affäre und es werden insgesamt mehr 'linke Sätze' von Rechts übernommen als anders herum.

Die Resultate des letzten Experiments sehen vielversprechend aus. Es lässt sich in allen Sets die Tendenz erkennen, dass die Mitte die meisten Sätze übernommen hat. Wenn wir das auf das Overton-Fenster übertragen, könnte das bedeuten, dass gewisse Meinungen und Aussagen, welche zuvor am Randbereich des Fensters waren, in die Mitte verschoben wurden, was wiederum bedeuten könnte, dass sich die Grundsätze des Fensters verändert haben. In drei von vier Sets erkennen wir eine eindeutige Mehrheit an Sätzen, welche von Rechts übernommen wurden. Das entspräche einer allgemeinen Verschiebung der Grundsätze in Richtung der linken politischen Haltung. Lediglich das Thema 'Flüchtling' scheint eine Verschiebung in Richtung Rechts zu erfahren.

4.4.3 Limitationen

Das letzte Experiment hat gute Ergebnisse geliefert. Jedoch existieren auch hier Limitationen. Als erstes komme ich auf die beschränkte Menge an Schlüsselwörtern zu sprechen. Wie zuvor erwähnt, wurden im Schnitt – mit Ausnahme des letzten Sets – deutlich über 100 Milliarden Satzvergleiche angestellt. Es benötigt selbstverständlich eine immense Zeit, diese Vergleiche abzuarbeiten. Ich hätte gerne noch größere Satzvergleiche ausgeführt, das war jedoch nicht im Rahmen dieser Arbeit und bietet also Stoff für zukünftige Arbeiten. Des Weiteren sind in dem Datensatz Werbungen und Web-Register aufgezeichnet worden. Da sie aber aus den gleichen Quellen entstammen und deshalb dem gleichen politischem Label zugeordnet sind, wurden sie aus meinem Versuch automatisch herausgefiltert. Damit konnten sie meine Ergebnisse nicht verfälschen. Es sollte aber auf diese Sätze geachtet werden, wenn man den Datensatz für sein Maschinelles-Lern-Modell benutzen möchte, denn dafür war der Datensatz ursprünglich gedacht.
Nun komme ich noch einmal darauf zu sprechen, dass die verglichenen Sätze meistens nur etwa 1% der gesamten Daten ausmachen. 99% aller Daten waren somit keiner Art von Verschiebung ausgesetzt. Wie sehr dieses extreme Verhältnis die Signifikanz der gewonnenen Resultate beeinflusst, lässt sich nur schwer einschätzen. Ich kam schon einmal darauf zu sprechen, dass das Overton-Fenster als ein Rahmen betrachtet werden kann. Der letzte Versuch hat zum größten Teil Aussagen betrachtet, die innerhalb dieses Rahmens liegen. Es ist vorstellbar, dass Bewegungen innerhalb des Rahmens nur mühsam und in kleinen Schritten stattfinden. Das würde natürlich dann auch für die Grenzen gelten. Dementsprechend könnte das Verhältnis von 1 zu 99 durchaus plausibel sein.
Zuletzt wurde in diesem Experiment nicht die allgemeine Verteilung der Daten beachtet. In dem Datensatz sind jeweils etwa 25% der Artikel Rechts und Links gelabelt. Ungefähr 30% gehören der Mitte an. Mitte-Links und Mitte-Rechts machen nur jeweils etwa 10% aus, wobei mehr Artikel mit Mitte-Links gekennzeichnet wurden. Überraschenderweise haben die Diagramme aus Experiment 3 vermuten lassen, dass die wenigsten Artikel mit Mitte-Links gelabelt wurden. Doch tatsächlich sind am wenigsten Artikel mit ‘right-center’ gelabelt. Das lässt vermuten, dass Mitte-Links tatsächlich am seltensten kopiert wird und auch am seltensten von anderen kopiert hat.

5 Zukünftige Arbeiten

Das erste Experiment in dieser Arbeit hat gezeigt, dass ein diachroner Sentiment-Ansatz zu den 'taz' Daten nicht ausreicht, um Veränderungen innerhalb des Overton-Fensters zu entdecken. Es wäre aber möglich, dass dieser Ansatz mit einem größeren Datensatz weitaus bessere Ergebnisse erzielt. Ich dachte dabei zum Beispiel an Plenarprotokolle des Bundestags. Diese wurden seit 1949 aufgezeichnet und sind auch öffentlich einsehbar. Ich kann mir gut vorstellen, dass mit einer längeren Zeitspanne auch deutlich mehr Unterschiede im Sentiment zu sehen sind.

Im zweiten Experiment zeigte sich SentiWS als geeignete Methode, Themen, welche in der Gesellschaft breit diskutiert werden, durch Sentiment aufzudecken. Mithilfe der Sentiment-Einschätzung von SentiWS könnte es möglich sein, Profile für Themen zu erstellen, welche potenziell die Struktur des Overton-Fensters beeinflussen. SentiWS ist eine gute und simple Methode, um Sentiment-Klassifizierungen vorzunehmen. Es wäre schön zu sehen, wenn SentiWS weiter verbessert und eine elegante Lösung für die neutrale Klassifizierung implementiert werden würde.

Da jedoch das letzte Experiment die aussagekräftigsten Resultate – in Bezug auf Veränderungen innerhalb des Overton-Fensters – lieferte, halte ich diesen Ansatz für den besten der drei vorgestellten. Mir sind zwei Ideen gekommen, wie man den letzten Versuch verbessern könnte. Zuerst wäre es spannend, die Sätze nicht nach Schlüsselwörtern zu filtern, sondern ohne Filterung alle Sätze direkt miteinander zu vergleichen. Anschließend würden wieder nur Sätze mit den gewünschten Ähnlichkeitswerten behalten. Diese Sätze könnten letztlich mithilfe einer 'clustering' Methode zu gewissen Themen zusammengeführt werden. Dadurch würde man einen kompletten Überblick zu allen Themen und ihren Erwähnungen bekommen. Mit dieser Methode wäre es gegebenenfalls tatsächlich möglich, ein echtes Bild von dem Overton-Fenster zu kreieren. Die zweite Idee ist, dass die Sätze zusätzlich einem Sentiment-Verfahren unterzogen werden können. Ähnlich wie bei SentiWS sollte dabei auch die Intensität des Sentiments ermittelbar sein. Diese Intensität könnte als Gewicht für die Wiederverwendung eines Satzes genommen werden. Dadurch ließe sich nachvollziehen, welches Sentiment übernommen wurde und man könnte daraus ermitteln, in welcher Gefühlslage sich eine politische Haltung gegenüber gewissen Themen befindet. Besonders spannend wäre selbstverständlich ein Versuch, der beide Ideen verwendet.

Da der letzte Datensatz für ein Maschinelles-Lern-Modell gedacht ist, liegt es auch nahe, die Einbindung einer KI zu erwägen. Mit der Möglichkeit, ungesehene Sätze einer politischen Haltung zuzuordnen und anschließend mithilfe von SBERT auswerten zu lassen, wäre ein überaus mächtiges Werkzeug geschaffen, das das Overton-Fenster vielleicht tatsächlich sichtbar werden ließe.

6 Diskussion

Um ein konkretes Bild von dem Overton-Fenster zu bekommen, ist es dringend notwendig weitere Experten in die Experimente einzubeziehen. Zum einen sind sie in der Lage, den Inhalt der gefundenen Ideen und Aussagen zu interpretieren, zum anderen müsste das Overton-Fenster an sich erst einmal genauer definiert werden. Viele Fragen dazu sind noch nicht ausreichend geklärt. Ich habe während dieser Arbeit immer Bezug auf *das* Overton-Fenster genommen. Bei genauerer Betrachtung existiert jedoch nicht nur ein einzelnes Overton-Fenster, sondern viele verschiedene Overton-Fenster. Dabei ist unklar, wann ein Fenster aufhört und ein anderes anfängt. Sprachen stellen sicherlich ein sehr einfaches Kriterium dar, um die Fenster auseinanderhalten zu können. Jedoch findet das Zusammenleben der Menschen und damit der menschliche Austausch auch über Sprache hinaus statt. Dadurch ist es schwer zu definieren, wie ein Fenster für sich alleine stehen könnte, scheinen sich doch alle Overton-Fenster gegenseitig zu beeinflussen.
Die nächste Schwierigkeit stellt die Festlegung eines Status quo dar. In meinen ersten beiden Versuchen gaben die Algorithmen und deren Sentiment Einschätzung den Status quo vor. Kann aber ein Konstrukt wie das Overton-Fenster in einem Moment genau eine Form haben, die wir einfangen können? Ich befürchte, dass dies nur sehr unwahrscheinlich möglich sein wird. Ich halte es für nicht exakt feststellbar, wann eine Idee entstanden ist und auch nicht wie sie sich ausgebreitet hat. Genauso wenig halte ich es für feststellbar, wo genau sich eine Idee innerhalb des Fensters befindet. Ich möchte hier einen Vergleich mit der Heisenbergschen Unschärferelation wagen (Heisenberg, 1930). Nach dem Unschärfeprinzip können nicht beliebig viele Eigenschaften eines Teilchens genau bestimmt werden. Das könnte sich bei dem Fenster ähnlich verhalten. Wenn sich eine Aussage in dem Fenster bewegt, können wir nicht mehr die Position der Aussage bestimmen, und wenn wir die Position einer Aussage gefunden haben, werden wir nicht ermitteln können, in welche Richtung sie sich bewegt. Das muss sicherlich nicht zwingend so sein und vor allem muss es nicht alle elementaren Teilchen des Fensters betreffen, jedoch schließt die Komplexität des Fensters diese Möglichkeit auch nicht aus.
Das Overton-Fenster kann auch als ein Instrument einer Instanz verstanden werden (Pigott, 2020). Als Beispiel müssen wir uns nur Diktaturen anschauen. Es ist üblich, dass diese ihre Ansichten und ihre Ideologie in das Zentrum des Fensters forcieren. Genauso werden Meinungen und Ideen, die diesen widersprechen, aus dem Rahmen geworfen. Das impliziert auch, dass es mehrere Fenster geben muss, denn die Regeln der Diktatoren gelten nicht für alle Menschen auf der Welt. Das bringt mich auch zu meinem letzten Punkt, denn die Grenzen des Overton-Fensters müssen sich damit auch nicht ausschließlich über verbale und schriftliche Kommunikation verändern, sondern gegebenenfalls auch durch Gewalt, Unterdrückung und Manipulation. In der Nachbetrachtung halte ich es dadurch für sinnvoller, das Overton-Fenster nicht als einen Rahmen zu betrachten, sondern viel mehr als ein Netzwerk aus Netzwerken.

7 Fazit

Diese Arbeit hatte sich als Ziel gesetzt, Indizien für den Wandel innerhalb des Overton-Fensters zu entdecken. Die abstrakte Natur und die theoretische Struktur des Fensters ließen dieses Modell bisweilen relativ unerforscht. Aus diesem Grund teilte sich die Arbeit in drei Experimente, die alle darauf abzielten, das Overton-Fenster greifbar zu bekommen. Zuerst wurde eine diachrone Sentiment-Analyse zu Zeitungsartikeln aus der Tageszeitung 'taz' vollzogen. Als Sentiment-Analyse-Methode wurden SentiWS und GSBERT verwendet. Die Ergebnisse zeigten jedoch auf, dass innerhalb der 8 Jahre keine großen Veränderungen zum Thema 'Flüchtling' stattgefunden haben. Die kleinen Veränderungen, die entdeckt wurden, sind jedoch je nach Sentiment-Verfahren widersprüchlich, wodurch keine Schlüsse auf Veränderungen innerhalb des Overton-Fensters gemacht werden sollten. Es ist auch nicht auszuschließen, dass sich das Fenster zu diesem Thema tatsächlich nicht verändert hat.

Das zweite Experiment bediente sich daraufhin einer Sentiment-Reaktion-Korrelation-Analyse zu deutschen Tweets. Auch hier kamen SentiWS und GSBERT zum Einsatz. Das Resultat hat ergeben, dass kein Zusammenhang zwischen der Intensität eines Sentiments und der Reaktionszahl eines Tweets nachgewiesen werden konnte. Es konnte aber gezeigt werden, dass Tweets mit besonders negativem Sentiment auf umstrittene Themen aufmerksam machen. Diese Themen sind wiederum gute Schlüsselwörter für weitere Analysen, da sie die vielversprechendsten Kandidaten sind, die auf eine Verschiebung innerhalb des Overton-Fensters hinweisen.

Das letzte Experiment verwendete eine diachrone Satz-Ähnlichkeit-Analyse auf englische Zeitungsartikel. Als Satz-Vektorisierung-Algorithmus kam SBERT zum Einsatz. Insgesamt wurden vier verschiedene Sets an Schlüsselwörtern verwendet. Die Ergebnisse zeigen auf, dass einige Aussagen von unterschiedlichen politischen Richtungen wiederverwendet wurden. In allen gezeigten Grafiken der vier vorgestellten Sets haben wir jeweils die gleichen Trends beobachten können. Daraus lässt sich vermuten, dass ein Wandel innerhalb des Fensters stattgefunden hat und dass sich dieser Wandel durch das Verfahren messen ließ.
Zuletzt stellte sich heraus, dass die Definition des Overton-Fensters nicht eindeutig ist, und es gegebenenfalls einer Überarbeitung der Definition bedarf. Alles in allem kann diese Arbeit jedoch bestätigen, dass ein Wandel oder eine Verschiebung innerhalb des Overton-Fensters durchaus gemessen werden kann. Damit wird das zuerst unfassbare Overton-Fenster letztlich doch noch greifbar.

8 Anhang

Link zur Google Colab Pipeline:

`https://colab.research.google.com/drive/1unqoLEab3Z9K1tU_BV1td2OlE_G_DZZV?usp=sharing`

Literatur

Arruda, H., Cardoso, F., Ferraz de Arruda, G., Hernández, A., da F. Costa, L., & Moreno, Y. (2022). Modelling how social network algorithms can influence opinion polarization. *Information Sciences*, *588*, 265–278. https://doi.org/10.1016/j.ins.2021.12.069

Bauer, R. (2021). The Twitter Game. In *Narrative Mechanics* (S. 337–350). transcript Verlag, Bielefeld, Germany. https://doi.org/10.1515/9783839453452-022

Church, K., & Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography. In *ACL '89: Proceedings of the 27th annual meeting on Association for Computational Linguistics* (S. 76–83). Association for Computational Linguistics, Vancouver, British Columbia, Canada. https://doi.org/10.3115/981623.981633

Cieliebak, M., Deriu, J. M., Egger, D., & Uzdilli, F. (2017). A Twitter Corpus and Benchmark Resources for German Sentiment Analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (S. 45–51). Association for Computational Linguistics, Valencia, Spain. https://doi.org/10.18653/v1/W17-1106

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate Speech Dataset from a White Supremacy Forum. In *2nd Workshop on Abusive Language Online*. HSLT Group at Vicomtech. Verfügbar 5. März 2022 unter https://arxiv.org/abs/1809.04444

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Verfügbar 5. März 2022 unter https://arxiv.org/abs/1810.04805

Firsching, J. (2021). *Twitter Statistiken 2021: Aktuelle Nutzerzahlen, Nutzerwachstum und Umsatz*. Verfügbar 5. März 2022 unter https://www.futurebiz.de/artikel/twitter-statistiken-nutzerzahlen/

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (S. 759–765).

Google. (2022). *Machine Learning Glossary*. Verfügbar 5. März 2022 unter https://developers.google.com/machine-learning/glossary/#logits

Guhr, O., Schumann, A.-K., Bahrmann, F., & Böhme, H.-J. (2020). Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In *Proceedings of the 12th Conference on Language Resources and Evaluation* (S. 1627–1632). European Language Resources Association (ELRA), Marseille, France. https://aclanthology.org/2020.lrec-1.202

Heisenberg, W. (1930). *Die Physikalische Prinzipien der Quantentheorie*. S. Hirzel Verlag, Stuttgart, Germany.

Jacobsen, L. (2018, 26. Juli). *Krasse Meinungen wehen uns mit voller Wucht ins Gesicht.* Verfügbar 5. März 2022 unter https://www.zeit.de/politik/deutschland/2018-07/overtone-fenster-diskussionen-debatten-diskurse-radikal

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). *Bag of Tricks for Efficient Text Classification*. Verfügbar 5. März 2022 unter https://arxiv.org/abs/1607.01759

Kahmann, C., & Heyer, G. Measuring Context Change to Detect Statements Violating the Overton Window. In: In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR,* SciTePress, 2019, 392–396. ISBN: 978-989-758-382-7. https://doi.org/10.5220/0008191803920396.

Kahmann, C., Niekler, A., & Heyer, G. Detecting and Assessing Contextual Change in Diachronic Text Documents using Context Volatility. In: In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR,* INSTICC. SciTePress, 2017, 135–143. ISBN: 978-989-758-271-4. https://doi.org/10.5220/0006574001350143.

Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., & Potthast, M. (2019). SemEval-2019 Task 4: Hyperpartisan News Detection. *13th International Workshop on Semantic Evaluation (SemEval 2019)*, 829–839. https://doi.org/10.18653/v1/S19-2145

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. Verfügbar 5. März 2022 unter https://arxiv.org/abs/1907.11692

Opuszko, M., Ulbricht, S., & Bode, L. (2018). Analysis of Twitter Communication During the 2017 German Federal Election. In *Proceedings of the 7th International Conference on Data Analytics 2018* (S. 34–38). IARIA, Athens, Greece.

Pigott, J. (2020). Terraforming the Internet: The media's plan for the Overton Window 2.0. The Iconoclast. Verfügbar 5. März 2022 unter https://www.researchgate.net/publication/344486882_Terraforming_the_Internet_The_media%5C%27s_plan_for_the_Overton_Window_20

Quasthoff, U. (2010). *Deutsches Kollokationswörterbuch*. deGruyter, Berlin, New York.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. Verfügbar 5. März 2022 unter https://arxiv.org/abs/1908.10084

Remus, R., Quasthoff, U., & Heyer, G. (2010). SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. *Proceedings of the International Conference on Language Resources and Evaluation*. http://www.lrec-conf.org/proceedings/lrec2010/summaries/490.html

Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2019). HateMonitors: Language Agnostic Abuse Detection in Social Media. *Working notes of FIRE 2019 - forum for information retrieval evaluation*. Verfügbar 5. März 2022 unter https://arxiv.org/abs/1909.12642

Sänger, M., Leser, U., Kemmerer, S., Adolphs, P., & Klinger, R. (2016). SCARE - The Sentiment Corpus of App Reviews with Finegrained Annotations in German. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1114–1121. https://www.aclanthology.org/L16-1178/

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., & Boeker, M. (2020). *GottBERT: a pure German Language Model*. Verfügbar 5. März 2022 unter https://arxiv.org/abs/2012.02110

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA*. https://doi.org/10.1109/cvpr.2015.7298682

Sidarenka, U. (2016). PotTS: The Potsdam Twitter Sentiment Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (S. 1133–1141). European Language Resources Association (ELRA), Portorož, Slovenia. https://aclanthology.org/L16-1181

Stone, P., Dunphy, D., Smith, M., & Ogilvie, D. (1967). *The General Inquirer: A Computer Approach to Content Analysis* (Bd. 4). American Educational Research Journal, Washington, USA. https://doi.org/10.2307/1161774

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 6000–6100. Verfügbar 5. März 2022 unter https://arxiv.org/abs/1706.03762

Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., & Biemann, C. (2017). Proceedings of the GermEval 2017-Shared Task on Aspect-based Sentiment in Social Media Customer Feedback. In *Proceedings of GermEval - Shared Task on Aspect-based Sentiment in Social Media Customer Feedback* (S. 1–12). GermanEval 2017, Berlin, Germany.

Zipf, G. K. (1972). *Human Behaviour and the Principle of Least Effort*. Hafner Pub. Co., New York, USA.

Selbständigkeitserklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann. Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Ort: Datum: Unterschrift: