Institut für Visualisierung und Interaktive Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Themen-spezifische komparative Visualisierung von Wortverwendungen zwischen Nachrichtenmedien

Oliver Ferch

Studiengang:	Softwaretechnik
Prüfer/in:	Prof. Dr. Thomas Ertl
Betreuer/in:	Franziska Huth, M.Sc., Johannes Knittel, M.Sc.
Beginn am:	17. September 2019
Beendet am:	17. März 2020

Inhaltsverzeichnis

1 Einleitung 9

2 Grundlagen Framing 11
2.1 Was ist Framing? 11
2.2 Wie funktioniert Framing? 11

3 Related Work 13
3.1 Berechnung und Auswertung von Framing 14
3.2 Visualisierung von Wortverwendungen 15
3.2.1 LingoScope 15
3.2.2 Compare Clouds 16
3.3 Abgrenzung zu dieser Arbeit 17

4 Konzept und Prototyp 19
4.1 Konzept des Prototypen 19
4.2 Auswahl der Werkzeuge zur Erstellung des Prototyps 20
4.3 Datenquelle und Verarbeitung der Daten 22
4.4 Prototyp Design 23
4.4.1 Hauptseite 27
4.4.2 Artikelseite 29
4.4.3 Gruppierungsseite 29
4.5 Arbeitsablauf im Prototyp 33
4.6 Limitationen des Prototypen 33

5 Anwendungsfälle 35
5.1 Auswertung für das Suchwort „immigrant“ 35
5.2 Auswertung für das Suchwort „medicare“ 37
5.3 Auswertung für das Suchwort „climate“ 39

6 Zusammenfassung und Ausblick 43
6.1 Zusammenfassung 43
6.2 Ausblick 44

Literaturverzeichnis 45

Abbildungsverzeichnis

3.1 Related Work LingoScope 16
3.2 Related Work Compare Clouds 17

4.1 Übersicht zu Seiten des Prototyps 23
4.2 Einstellungsmöglichkeiten des Prototyps 24
4.3 Hauptansicht im Prototypen 24
4.4 Anwendung von Ranking 25
4.5 Beispiel der Reduzierung auf den Wortstamm 26
4.6 Anzeige der Wörter vor oder nach dem Wort 26
4.7 Beispiel Datum Diagramm 27
4.8 Beispiel Balken Diagramm 28
4.9 Beispiel Artikelseite 29
4.10 Gruppierungen der Top TF-IDF Wörter 30
4.11 Beispiel Sentiment Diagramm 31
4.12 Beispiel Prozent Diagramm 32
4.13 Beispiel Worthäufigkeiten Diagramm 32
4.14 Beispiel Wortstamm bei Gruppierungen 33

5.1 Ergebnis Balkendiagramm Beispiel 1 35
5.2 Ergebnis Ranking Beispiel 1 36
5.3 Ergebnis Gruppierungsdiagramm Beispiel 1 36
5.4 Ergebnis Sentiment Diagramm Beispiel 1 37
5.5 Ergebnis Balkendiagramm Beispiel 2 37
5.6 Ergebnis Ranking Beispiel 2 38
5.7 Ergebnis Gruppierungsdiagramm Beispiel 2 39
5.8 Ergebnis Balkendiagramm Beispiel 3 40
5.9 Ergebnis Ranking Beispiel 3 40
5.10 Ergebnis Gruppierungsdiagramm Beispiel 3 41

Abkürzungsverzeichnis

CSS Cascading Style Sheets. 20

GB Gigabyte. 22

HTML Hypertext Markup Language. 20

IDF Inverse Document Frequency. 20

JSON JavaScript Object Notation. 22

LDA Latent Dirichlet Allocation. 27

SML Supervised Machine Learning. 14

TF Term Frequency. 20

TF-IDF Term Frequency-Inverse Document Frequency. 15

1 Einleitung

In Zeiten immer größer werdender Informationsflut aus aller Welt wird Nachrichtenmedien immer wieder fehlende Neutralität bei ihrer Berichterstattung vorgeworfen. Wie kann die Beeinflussung der Berichterstattung, falls diese vorliegt, erkannt und nachgewiesen werden, ohne komplizierte und aufwendige Analysen der Daten durchführen zu müssen? Um einen Framing Bias [Ent07] nachzuweisen ist schließlich die Sichtung riesiger Datenmengen nötig um Unterschiede in der Art und Weise der Berichterstattung zwischen den einzelnen Nachrichtenmedien zu erkennen.

Um diese Arbeit zu erleichtern, ist die maschinelle Verarbeitung und Datenaufbereitung, sowie die Visualisierung dieser Daten in einer graphischen Oberfläche zur besseren Identifikation möglicher Framings ein sinnvolles Ziel. In Bezug auf dieses Thema treten dabei verschiedene Fragen auf.

Wie kann eine automatische Auswertung für diese Datenmengen aussehen? Welche Werkzeuge, Parameter und Einstellungen müssen für ein gutes Ergebnis verwendet werden? Wie ist die Erkennung eines bestimmten Framings aus diesen Daten per Visualisierung kenntlich zu machen? Welche Optionen sollen Nutzer zur Identifikation von Framing erhalten? Können unterschiedliche Framings in der Visualisierung erkannt werden?

In dieser Arbeit geht es darum ein Konzept und einen Prototypen zu entwickeln, der für definierte Themen oder Suchwörter eine automatisierte Auswertung der Daten von verschiedenen Nachrichtenmedien vornimmt. Aus diesen verarbeiteten Daten sollen Ergebnisse entstehen, mit deren Hilfe anhand der Visualisierung deutliche Unterschiede bei den Nachrichtenmedien für Wortverwendungen sichtbar werden und durch die Nutzer daraufhin ein Framing Bias nachgewiesen werden kann.

Die Arbeit setzt sich aus fünf größeren Teilen zusammen. Zunächst werden die Grundlagen für Framing behandelt, um anschließend im Abschnitt Related Work auf Forschungsarbeiten zum Thema Framing und Visualisierung einzugehen. Dabei werden Arbeiten zur Berechnung und Auswertung von Framing betrachtet, wie auch Prototypen für die Visualisierung von Wortverwendungen vorgestellt. Im vierten Kapitel wird auf das Konzept des eigenen Prototyps eingegangen, welche Methoden und Algorithmen verwendet wurden, sowie der erstellte Prototyp mit seinen verschiedenen Seiten, Funktionen und Optionen vorgestellt. Danach diskutiere ich die Ergebnisse für mehrere Suchwörter und werte diese aus. Abschließend wird ein Fazit über den erstellten Prototyp gezogen und ein Überblick über weitere mögliche Erweiterungen und Verbesserungen gegeben.

2 Grundlagen Framing

2.1 Was ist Framing?

Je nach Blickwinkel oder Perspektive auf ein Thema kann sich die Meinung zu einem Thema ändern, genau an diesem Punkt setzt Framing an. Die Art und Weise wie an ein Thema heran gegangen wird, ist entscheidend dafür wie unterschiedlich auch die Wahrnehmung des Themas und die damit verbundene eigene Meinung ist. Bestimmte Teile des Themas werden betont oder verstärkt hervorgehoben, um ein Framing herzustellen. Die Beeinflussung des Themas durch Framing kann so subtil sein, dass sie vom Leser nicht erkannt wird [MWY+18]. Da Framing viele Facetten hat und daher auch vielfältig einsetzbar ist, wird es entsprechend verwendet, um die Meinung zu einem Thema in eine bestimmte Richtung zu lenken.

Dennoch sind die einzelnen Dimensionen, die in die Meinung über ein Thema einfließen, von Mensch zu Mensch unterschiedlich, was auch der Grund dafür ist das Menschen unterschiedlicher Meinung sein können. Jeder wird durch das Framing unterschiedlich beeinflusst, dennoch können bestimmte Stellschrauben genutzt werden, um die öffentliche Meinung zu lenken. Durch positives Berichten über ein Thema wird es sehr viel wahrscheinlicher eine positive Meinung zu diesem Thema zu haben, als wenn darüber eher negativ berichtet würde. Daher beschäftigen sich viele Wissenschaftler damit, Framing zu untersuchen, nachzuweisen und zu zeigen, wie genau es eingesetzt wird bei der Meinungsbildung der breiten Masse.

Eine große Rolle spielt auch das Übernehmen von anderen Meinungen, welche in der Diskussion mit einer anderen Person über ein Thema aufgekommen sind. Wenn der Gegenüber anschaulich und für einen selbst nachvollziehbar argumentiert, ist es möglich, dass die Meinung zum Thema vom Gegenüber übernommen wird. Dies machen sich auch die Eliten zu Nutze, die Art und Weise wie sie etwas kommunizieren führt zu einer Beeinflussung, welche auch Framing Effekt genannt wird [CD07]. Die Eliten werden als gebildeter angesehen und dadurch auch ihre Meinung zu einem Thema als glaubwürdiger aufgefasst, als wenn dies eine unbekannte Person sagen würde. Letztlich ist Framing der Versuch, die Meinung der Leser zu beeinflussen und gegebenenfalls auch zu ändern. Die Beeinflussung findet dabei unterbewusst statt und versucht die Leser zu überzeugen, eine bestimmte Meinung bei einem Thema anzunehmen.

2.2 Wie funktioniert Framing?

Damit ein Framing Effekt eintreten kann, müssen zusätzliche, beeinflussende Informationen vorhanden sein, die für das Framing sorgen. Ohne Informationen über ein Thema kann die Person sich keine fundierte Meinung bilden, somit können die zusätzlich eingefügten Informationen für eine veränderte Meinung sorgen. Ebenfalls kann das Weglassen von bestimmten Informationen auch zu einer veränderten Wahrnehmung des Themas führen [MWY+18]. Um ein Framing zuverlässig durchzuführen, müssen Informationen passiv gestreut werden, damit sie im Unterbewusstsein als normale Informationen aufgefasst werden. Da sich die Meinungsfindung bei Menschen auf die ihnen zugänglichen Informationen stützt, sind diese nicht erkennbar veränderten Informationen ein Mittel, um die Meinung indirekt zu beeinflussen.

Beim Berichten über eine bestimmte Nachricht werden gezielt Fakten und Informationen hervorgehoben, um durch dieses Framing die Berichterstattung in die gewünschte Richtung zu lenken. Es gibt außerdem verschiedene Ebenen, auf denen Informationen verzerrt dargestellt werden können. Eine dieser Ebenen ist die Auswahl an Nachrichten, die von einem Nachrichtenmedium berichtet werden, meist eine Teilmenge an allen verfügbaren Nachrichten und Informationen zu dem Thema. Die Auswahl der Nachrichten und wie viel Aufmerksamkeit einem Thema durch das Nachrichtenmedium geschenkt wird ist eine Möglichkeit, um dieses „Agenda Setting“ [MWY+18] zu erkennen. Die zweite größere Ebene ist das bereits angesprochene Wie, wie über Nachrichten berichtet wird. Je nachdem wie ein Thema betrachtet und behandelt wird, ändert sich die Wahrnehmung zu diesem Thema.

3 Related Work

Die Erkennung von Framing ist eine aufwendige Arbeit, bei der mehrere Faktoren ausgewertet werden müssen. In der Arbeit von Morstatter, Wu, Yavanoglu, Corman und Liu: „Identifying Framing Bias in Online News“ wird mit einem maschinellen Lernansatz versucht Framing zu identifizieren [MWY+18]. Da bei Framing Bias immer wieder versucht wird das Thema in eine bestimmte Richtung zu lenken, entsteht ein Überfluss dieser beeinflussenden Faktoren, den man klassifizieren und erkennen kann. Daher wird in diesem Paper versucht, Sätze in beeinflusster Satz und nicht beeinflusster Satz zu kategorisieren.

Die Ansätze zur Analyse von Framing kommen meist aus der sozialwissenschaftlichen Literatur und sind je nach Betrachtungsart oder Herangehensweise unterschiedlich definiert. Daher gibt es für Framing von verschiedenen Wissenschaftlern einige Framing Definitionen, die jeweils unterschiedliche Aspekte und Betrachtungsebenen abdecken. Definitionen können generell gehalten sein oder sich auf bestimmte Teile und somit themenspezifisches Framing beziehen. Der meist zitierte Wissenschaftler im Bezug auf Medien Frames [Mat09], Robert M. Entman definierte vier Ebenen von Frames im Kommunikationsprozess im Bezug auf politische Prozesse [EMP09]:

- in der Kultur
- in den Ansichten der Eliten und politischer Kommunikation
- in der Berichterstattung der Medien
- in der öffentlichen Meinung

Andere Definitionen, die sich mit der Art des Framings befassen sind unter anderem [Mat09]:

- Konflikt Framing
- Menschliche Interessen Framing
- Wirtschaftliche Konsequenzen Framing
- Moralisches Framing
- Verantwortung Framing

Der Handelskrieg der USA mit China, geführt durch die Administration unter Präsident Trump nutzt verschiedene dieser genannten Framings, um Zustimmung für die eigenen Handlungen zu gewinnen. Da China sich laut Trump nicht an Absprachen hält und einen zu hohen Handelsüberschuss gegenüber der USA habe, müssen importierte chinesische Waren mit Zöllen belegt werden, um die eigene Wirtschaft zu schützen. Verschiedene Framings kommen hier zum Einsatz, die zusätzlich mit dem Slogan „Make America Great Again“ in den Ebenen der Kultur, Ansichten der Eliten und der öffentlichen Meinung, wie auch in mehreren Arten an Framings, beispielsweise des Konflikts, Wirtschaftlicher Konsequenzen oder menschlicher Interessen bei den Amerikanern Anklang finden.

3.1 Berechnung und Auswertung von Framing

In der Forschungsarbeit von Burscher, Odijk, Vliegenthart, Rijke & Vreese: „Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis“ wird die Anwendung des überwachten maschinellen Lernens, Supervised Machine Learning (SML) auf Frame Coding untersucht [BOV+14]. Durch die Automatisierung der Findung von Frames in Nachrichtenmedien, unterstützt der Ansatz mit SML die Analyse von großen Datenmengen, auch wenn das Budget knapp ist.

Normalerweise wird das Nachrichten Framing per Hand von ausgebildeten Personen analysiert, dies ist aber sehr aufwändig und kostspielig. Die Analyse ist zuverlässig, aber durch die zuvor genannten Aspekte ist der Umfang der Inhaltsanalyse beschränkt. Die manuelle Analyse wird durch verschiedene Fragen als Indikatoren für News Frames unterstützt [BOV+14]. Fragen sind dabei so gestellt, dass die Bedeutung des Inhalts für die gegebenen Frames festgehalten wird. Mehrere Fragen werden kombiniert, um verschiedene Aspekte des Framings abzudecken.

Um überprüfen zu können wie gut der Ansatz mit SML ist, wird er mit dem Ansatz der manuellen Analyse verglichen. Beim SML Ansatz werden vom Menschen kodierte Trainingsdokumente verwendet um automatisiert Variablen der quantitativen Auswertung von den Texten vorherzusagen [BOV+14]. Dies wird auf vier generische Frames angewendet, um zu sehen wie das Automatisieren für zukünftige Framing Studien anzuwenden ist.

Die meisten computergestützten Frame Codings folgen dem Wörterbuch basierten Ansatz. Vorher definierte Zeichenketten und Regeln werden verwendet, um Inhaltskategorien zu erhalten. Ein Nachteil ist, man muss diese Regeln und Zeichenketten manuell entwerfen und testen, was viel Zeit in Anspruch nimmt und die semantische Gültigkeit beeinträchtigen kann. Durch eingegrenzte Annahmen, Überlegungen und Suchparameter kann das Ergebnis verzerrt werden, da nicht alles vollständig abgedeckt werden kann.

Der in der Arbeit vorgestellte Ansatz der Verwendung von SML versucht diese zu kurz kommenden Faktoren anzugehen, indem der Computer einen Trainingssatz von relevanten Dokumenten erhält, die für Inhaltskategorien des Framings interessant sind. Es sind dennoch 3 Verarbeitungsschritte nötig: Die Dokumente müssen konvertiert und verwendbar gemacht werden, jedes Dokument wird ein Vektor von quantifizierbaren Textelementen (zum Beispiel Worthäufigkeit), die auch Feature genannt werden [BOV+14]. Als zweites werden diese Feature Vektoren aller Dokumente des Trainingssatzes mit der Dokument Inhaltserkennung genutzt, um die Kategorien des Inhalts zu trainieren. Dadurch erlernt die Maschine die einzelnen Kategorien des Inhalts und generiert ein Vorhersagemodell für zusätzliche Dokumente, um diese dann ebenfalls kategorisieren zu können. Im letzten Schritt werden Dokumente außerhalb der Trainingsdokumente genutzt, um für diese die Kategorisierung vorzunehmen.

Die Auswertung zeigt, dass dies effizienter und effektiver als der vorher genannte Wörterbuch basierte Ansatz ist. Außerdem ist es für die Inhaltsanalyse einfacher durchführbar und leichter zu erweitern. Da große Datenmengen untersucht werden können und nicht nur kleinere Stichproben wie bei der manuellen Analyse, können Stichprobenfehler verringert und somit die Genauigkeit gesteigert werden.

Als Dokumentmerkmal wird Term Frequency-Inverse Document Frequency (TF-IDF) genutzt, um ein Framing durch trainierte Klassifizierung zu ermöglichen. Anhand der Ergebnisse kann festgehalten werden, dass der gewählte Ansatz mit SML gut für Frame Coding geeignet ist, aber die Qualität der Ergebnisse von der Art und Weise abhängt wie es implementiert wurde. Je nachdem welches Framing betrachtet wird, sind die Ergebnisse unterschiedlich gut, insgesamt ist es aber eine effizientere und kostengünstigere Lösung als die manuelle Analyse.

3.2 Visualisierung von Wortverwendungen

Im Folgenden werden Prototypen ausgewählter vorheriger Arbeiten vorgestellt, die eine Visualisierung von Unterschieden bei Wortverwendungen untersuchen. Anhand der Visualisierung sollen mögliche Frames besser ersichtlich werden, als dies durch Text Analyse allein möglich ist.

3.2.1 LingoScope

In der Arbeit von Diakopoulos, Zhang & Salway: „Visual Analytics of Media Frames in Online News and Blogs“ wird der Prototyp von LingoScope vorgestellt [DZS13]. Dieser Prototyp soll die Nutzer per Visualisierung dabei unterstützen, durch Vergleich von zwei Nachrichtenquellen Unterschiede im Framing zu erkennen.

Durch die Visualisierung soll eine bessere Identifikation und Reflexion von Themen Frames entstehen, auch in Bezug auf längere Zeiträume. Abbildung 3.1 zeigt die verschiedenen Einstellungsmöglichkeiten im Prototypen, zum einen die Auswahl der gewünschten Nachrichtenmedien, als auch die Angabe des zu untersuchenden Themas. Mit der Auswahl *vor dem Suchwort*, *nach dem Suchwort*, *Satz der das Suchwort enthält* oder *der gesamte Paragraph der das Suchwort enthält* wird der Kontext angegeben, indem das Suchwort untersucht werden soll. Als zusätzliche Filter kann noch nach *Adjektiven* oder *positiven* und *negativen* Wörtern die Suche weiter spezifiziert werden.

Das Ergebnis der Suche wird der Häufigkeit nach sortiert angezeigt, dabei geben die Balkengrößen und Prozentangaben wie auch der zeitliche Verlauf einen Einblick in die Nutzung der Wörter für das jeweilige Nachrichtenmedium. Mit Klick auf eines der errechneten Kontext Wörter werden dafür auf der rechten Seite Artikelausschnitte angezeigt, die nochmal im Detail zeigen in welchem Zusammenhang das Suchwort und das Kontext Wort in Artikeln erscheinen.

Mit LingoScope erhält man somit die Möglichkeit, Kontext Wörter, die in der Nähe des Suchwortes auftreten zwischen verschiedenen Nachrichtenmedien zu vergleichen und je nach Häufigkeit der Wortverwendungen mögliche Frames zu erkennen.

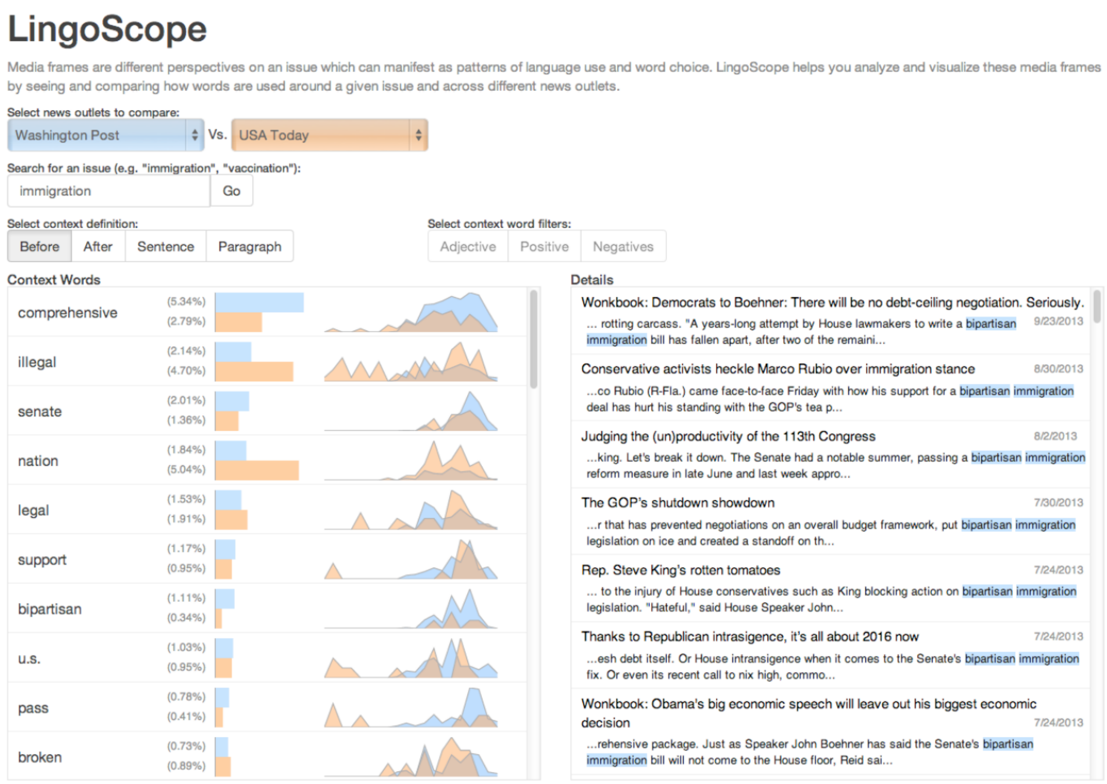

Abbildung 3.1: Ein Vergleich von Wörtern zwischen den Nachrichtenmedien für Wörter *vor* dem Suchwort „immigration“ [DZS13]

3.2.2 Compare Clouds

Die Ausarbeitung von Diakopoulos, Elgesem, Salway, Zhang & Hofland: „Compare Clouds: Visualizing Text Corpora to Compare Media Frames“ baut auf dem Ansatz von LingoScope auf und versucht einen breiteren Überblick wie auch eine Verbesserung des Vergleichs zwischen den Nachrichtenmedien zu erreichen [DES+15].

Der Prototyp von Compare Clouds soll die Analyse von Frames über die Worthäufigkeit und Kontextinformationen ermöglichen, dabei werden speziell drei textliche Ebenen behandelt, die von einer Visualisierung profitieren: Worthäufigkeiten, Wortkontexte und die Semantik. Medien Frames werden durch das Vorhandensein oder Fehlen bestimmter Schlüsselwörter oder Wortverwendungsmuster ersichtlich, welche im Prototyp anhand unterschiedlicher Schriftgröße, Schriftstärke und Farbe unterscheidbar gemacht werden sollen [DES+15]. In Abbildung 3.2 ist eine Auswertung im Prototyp zum Suchwort „surveillance“ dargestellt, dabei werden ausgewählte Mainstream-Medien mit Blogs verglichen. Untersucht werden in der Berechnung der Worthäufigkeiten hierbei Wörter, die im selben Satz wie das Suchwort vorkommen.

Wörter, die auf der linken Seite der Darstellung liegen und rötlich gefärbt sind, werden mehr von Mainstream-Medien verwendet. Dagegen sind Wörter auf der rechten Seite und in blau eher in Artikeln von Blogs zu finden. Wörter in der Mitte, im Übergang zwischen rot, blau und grau werden relativ gleichmäßig in den Vergleichsmedien genutzt. Mit Klick auf eines der erhaltenen Wörter werden rechts zusätzliche Informationen bezüglich der Kontext-Nutzungsrate sowie Vorkommen des Suchbegriffes in den jeweiligen Medien aufgezeigt. Außerdem werden Wörter, die im Zusammenhang mit dem ausgewählten Wort und Suchwort stehen in gelb hervorgehoben. Im unteren Teil werden Artikelausschnitte gezeigt, um einen genaueren Einblick auf den jeweiligen Kontext zu erhalten, in dem die Wörter genutzt werden.

Compare Clouds legt Wert darauf, interessante Begriffe anhand von Worthäufigkeiten hervorzuheben und durch räumliche Aufteilung den Nachrichtenmedien je nach Nutzung und passendem Kontext zuzuweisen.

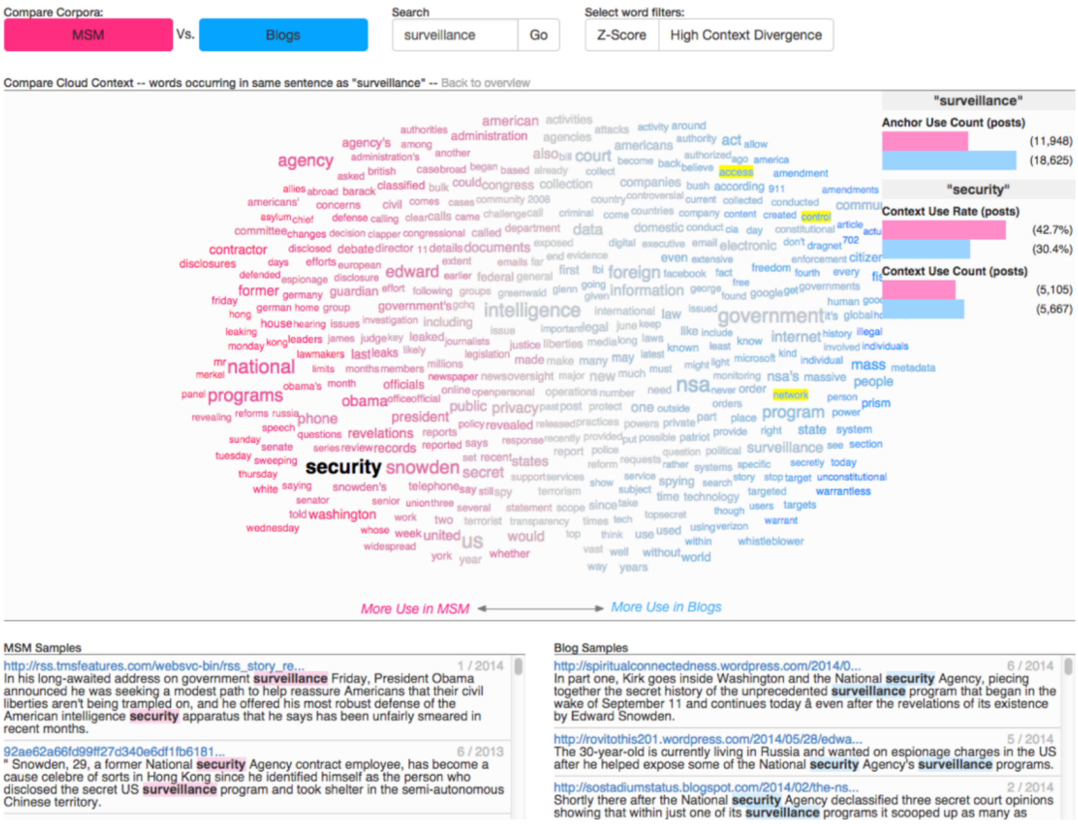

Abbildung 3.2: Visualisierung in Compare Clouds, die für das Wort „surveillance“ die Unterschiede in der Verwendung zwischen Mainstream-Medien und Blogs zeigt [DES+15]

3.3 Abgrenzung zu dieser Arbeit

Während in den genannten Arbeiten das Augenmerk hauptsächlich auf Sätze gelegt wird, die das Suchwort enthalten, wird in dieser Arbeit versucht eine Analyse der vollständigen Artikel zu erreichen die das Suchwort enthalten.

In Compare Clouds werden die einzelnen Wörter angezeigt, die im Bezug zum Suchwort stehen sollen, aber genauere Details über die einzelnen Wörter wie die Worthäufigkeit oder wieso ein Wort angezeigt wird, werden nicht dargestellt. In dieser Bachelor Arbeit wird versucht diese Werte mitzuliefern, um ein detaillierteres Verständnis zur Berechnung und Auswahl der Wörter zu erhalten. Es wird nicht ersichtlich wieso Wörter auf der linken Seite, in der Mitte oder auf der rechten Seite stehen. Man hat die vorgegebenen Definitionen des Prototyps, aber ein genaueres Nachvollziehen anhand der Werte ist nicht möglich. Durch die Gruppierung der Wörter in dieser Arbeit wird versucht die zusammenhängenden Wörter darzustellen und somit schlüssiger als in Compare Clouds zu zeigen welche Wörter zusammen auftreten und ein bestimmtes Thema beschreiben. In der gezeigten Abbildung 3.2 von Compare Clouds ist das Wort „security“ auf der linken Seite mit drei hervorgehobenen Wörtern der rechten Seite dargestellt. Eine genaue Aussage über die Wortverwendungen für beide Nachrichtenmedien erhält man dadurch nicht, da der nähere Zusammenhang dieser Wörter für beide Nachrichtenmedien nicht klar erkennbar ist. Wieso Wörter an einer bestimmten Stelle stehen, zum Beispiel aufgrund der Nutzungshäufigkeit, kann vom Nutzer nicht nachvollzogen werden. LingoScope enthält ebenfalls wenige zusätzliche Informationen, wie die Anzahl an Artikeln, die das Suchwort für die jeweilige Seite enthalten oder wieso Wörter die angezeigt werden ausgewählt wurden.

In beiden Arbeiten sind generell wenig Auswahlmöglichkeiten für die Nutzer gegeben, um spezielle Parameter wie „Suchbreite erhöhen“ einstellen zu können. Die Nutzer haben in den gezeigten Prototypen wenig Einsicht in Details, zum Beispiel wieso bestimmte Wörter vorkommen und auch insgesamt wenig Möglichkeiten vorhandene Daten in der Breite zu analysieren. Viele Einstellungsmöglichkeiten dieser Bachelor Arbeit sollen die Nutzer dabei unterstützen eigene Präferenzen zu definieren und so gewünschte Ergebnisse zu erhalten. Mit den Gruppierungen der Wörter soll ein Verständnis für häufig im selben Kontext auftretende Worte erlangt werden. Durch ein eigenes Ranking wird versucht, stark voneinander abweichende Werte beider Vergleichsseiten hervorzuheben und somit falls möglich ein bestimmtes Framing bei den Wortverwendungen sichtbar zu machen.

4 Konzept und Prototyp

4.1 Konzept des Prototypen

Basierend auf den Ergebnissen und Erkenntnissen der beiden vorgestellten Visualisierungsarbeiten, die aus einer Vielzahl weiterer Visualisierungsansätze ausgewählt wurden, wird in dieser Arbeit versucht Framing durch Wortverwendungen noch besser sichtbar zu machen. Zusätzlich sollen die Nutzer mehr Einstellungsmöglichkeiten und genauere Details geliefert bekommen um sich Ergebnisse der Visualisierung ihren Ansprüchen entsprechend anzeigen zu lassen. Durch die Nutzung stochastischer Prozesse soll die Verarbeitung von größeren Datenmengen für verschiedene Nachrichtenseiten ermöglicht werden. Außerdem bietet es sich an Textanalyse in Kombination mit der Visualisierung zu verwenden, um Unterschiede in den Wortverwendungen der verschiedenen Nachrichtenmedien heraus zu arbeiten. Die daraus folgenden Parameter werden weiterverarbeitet und in der visuellen Darstellung für die Nutzer aufbereitet. Das Interpretieren der Ergebnisse durch die Nutzer ermöglicht es, spezifische Themen oder Unterschiede zu erkennen, nach denen explizit gesucht wurde.

Wortverwendungen sind nützlich, um die Art und Weise zu erfassen, wie Nachrichtenmedien berichten, denn dadurch kann der sprachliche Wortschatz erfasst und daraus abgeleitet werden, welche Zielgruppen bestimmte Nachrichtenmedien ansprechen wollen. Je nachdem welche Wörter in Artikeln gewählt werden, können bestimmte Stimmungen über die Themen erzeugt werden. Benutzt man positive Wörter ist das Thema positiv besetzt, während bei negativen das Gegenteil der Fall ist. Das sieht man zum Beispiel an genutzten Wortkombinationen wie illegale Einwanderer (negative Betrachtung), hilfsbedürftige Flüchtlinge (positive Betrachtung) oder Klimawandel (neutral) und Klimahysterie (negativ). Durch diese gewählten Stimmungen im Text können Leser dieser Nachrichtenmedien beeinflusst wird.

Doch wie genau sollen Texte analysiert werden? Wie bereits in den vorgestellten Prototypen bietet es sich an zum einen die Häufigkeit der Wörter und Wortpaare/Wortgruppierungen zu untersuchen. Je häufiger ein Wort vorkommt, umso wichtiger ist es für ein bestimmtes Thema. Die Worthäufigkeit, also die Anzahl an Vorkommen eines Wortes, kann das Ergebnis aber verzerren, da Füllwörter oder für den Satzbau notwendige Wörter hier einen sehr hohen Wert erreichen. Um diese Wörter in den Ergebnissen zu reduzieren wird die Wichtigkeit der Wörter per TF-IDF Verarbeitung berechnet und die Worthäufigkeit als zusätzliche Information in der Visualisierung angezeigt. In einer sogenannten „Stopword“ Liste können sehr häufig auftretende Wörter oder unerwünschte Symbole angegeben und damit kurz vor der Verarbeitung herausgefiltert werden. Durch TF-IDF sollen als Ergebnis Wörter angezeigt werden, die für das Thema relevant sein können, da sie trotz ihres nicht häufigen Vorkommens ein wichtiger Bestandteil von Artikeln und eventuell essentiell für ein Thema sein können.

Um die Analyse zu erweitern sollen aber nicht nur einzelne Wörter betrachtet werden, sondern auch Wörter, die im direkten Bezug zu diesen stehen. Dadurch kann die Stimmung oder Richtung, in die ein Thema geht genauer analysiert werden und ein Einblick auf Themen entstehen, die in Verbindung mit dem Suchwort stehen. Der Kontext eines Themas soll erkennbar werden und die Gruppierungen der Wörter die Unterschiede zwischen den einzelnen Nachrichtenmedien herausarbeiten, indem die Unterschiede in der Verwendung der Wörter farblich festgehalten werden.

4.2 Auswahl der Werkzeuge zur Erstellung des Prototyps

Für die Umsetzung des Prototyps werden die Programmiersprachen Python [VD95] und JavaScript verwendet, zusätzlich wird die Dokumentbeschreibungssprache Hypertext Markup Language (HTML) mit Cascading Style Sheets (CSS) und der JavaScript Bibliothek D3.js [BOH11] genutzt. Python ist für die Hintergrundarbeit, die Datenverarbeitung durch stochastische Prozesse zuständig, während JavaScript im Zusammenspiel mit HTML, CSS und im speziellen D3 für die Darstellung der Visualisierungen anhand einer Website mit mehreren Seiten für die verschiedenen Darstellungen verwendet wird.

In Python werden im speziellen die Pakete scikit-learn [PVG+11] für die Berechnung der TF-IDF Werte verwendet, TextBlob [Tex] zur Analyse des Sentiments der Artikel (wie ist die allgemeine Stimmung in einem Artikel), sowie Gensim [ŘS10] mit Word2Vec für die Berechnung ähnlicher im Kontext vorkommender Wörter zur Bildung von Gruppen. Diese Pakete werden häufig für stochastische Prozesse eingesetzt, da ihr Aufgabenspektrum sehr weit gefasst ist. Aufbauend auf den errechneten Daten wird durch eigene Algorithmen eine Sortierung, beziehungsweise Einteilung vorgenommen, die dann in der Visualisierung dargestellt werden kann. In D3 wurde das Paket D3 Tip [Pal] verwendet, um ansprechendere Tooltips zu erstellen.

TF-IDF

Als Dokumentmerkmal wird TF-IDF genutzt, um ein Framing per trainierten Klassifizierungen zu ermöglichen. Jedem Wort wird zugeordnet wie oft es im Dokument vorkommt (Term Frequency (TF)) und die umgekehrte Häufigkeit von Artikeln in der Sammlung von Artikeln, die das Wort enthalten (Inverse Document Frequency (IDF)) [BOV+14]. Bei der TF-IDF Gewichtung wird darauf geachtet wie stark ein Wort zur Unterscheidung von Artikeln dient. Seltener in Artikeln benutzte Wörter werden als unterscheidender gesehen und haben dadurch einen höheren Wert. Daher wird TF-IDF verwendet, da es sich gut dazu eignet „spezielle“, „auffallende“ Worte hervorzuheben, die keine Füllwörter sind, sondern für die jeweiligen Artikel prägend sein können.

Als Trainingsdokument werden alle Artikel aus den Nachrichtenmedien genutzt, um einen möglichst großen Trainingssatz aufzubauen. Die für jeden Artikel spezifischen TF-IDF Werte werden am Schluss für das jeweilige Wort aufsummiert, damit Wörter mit den im Korpus insgesamt höchsten Werten im Prototyp absteigend sortiert in einer Visualisierung angezeigt werden können. Dadurch sollen die Nutzer einen Überblick darüber erhalten, welche Wörter im gesamten Korpus eine hohe Bedeutung in Bezug auf den Inhalt haben könnten.

Beim genutzten Algorithmus von TF-IDF aus dem Paket scikit-learn [Skl] sind die einstellbaren Parameter so gesetzt: `smooth_idf = False`, `use_idf = True`, `sublinear_tf = False`.

`smooth_idf` wird auf False gesetzt, da eine ungewollte Verzerrung dadurch auftreten könnte. Die Verzerrung besteht darin, dass für jedes Wort angenommen wird, dass in einem Artikel alle Wörter einmal enthalten sind. Da der Trainingssatz alle Artikel und somit alle Wörter mindestens einmal enthält, ist diese Einstellung nicht nötig und kann nur zu einer Verfälschung des Ergebnisses führen.

`use_idf` wird auf True gesetzt, um die Berechnung des TF-IDF mit IDF zu verwenden.

`sublinear_tf` bleibt auf False, da sonst der Wert von TF per Logarithmus aufgelöst wird.

Die verwendete TF-IDF Formel aus dem Scikit-learn Paket, mit Wort t eines Dokuments d in einer Dokumentensammlung ist [Skl]:

$$\text{tf-idf}(t, d) = \text{tf}(t, d) * \text{idf}(t)$$

Der IDF Wert wird mit den Parametern n = *Gesamtanzahl Dokumente* und df(t) = *Dokumenthäufigkeit von Wort t* (wieviele Dokumente in der Sammlung enthalten Wort t) berechnet:

$$\text{idf}(t) = \log\left[\frac{n}{\text{df}(t)}\right] + 1$$

Diese IDF Formel weicht von der bekannten IDF Formel ab, damit Wörter die in allen Dokumenten vorkommen dennoch einen IDF Wert größer 0 erhalten:

$$\text{idf}(t) = \log\left[\frac{n}{(\text{df}(t) + 1)}\right]$$

Word2Vec für spätere Gruppierungen

Durch Gruppierung sollen Wörter zueinander geführt werden, die oft im selben Kontext vorkommen und somit gemeinsam ein bestimmtes Thema beschreiben. Wenn man diese Gruppen eines Nachrichtenmediums mit den Gruppen eines anderen Nachrichtenmediums vergleicht können eventuell Unterschiede ersichtlich werden, die auf ein bestimmtes Framing hindeuten. Für Wörter mit den höchsten TF-IDF Werten wird eine Berechnung als Startwort mit Word2Vec durchgeführt, um ähnliche Worte zu diesem Ausgangswort zu finden.

Word2Vec analysiert Wortkontexte und kann Wörter, die eine ähnliche Bedeutung wie die angegebenen Startwörter dieser Berechnungen haben, über den gesamten Korpus hinweg in allen Texten finden. Dies ist ein Vorteil gegenüber den Vorkommen vor und nach einem Wort, da es Text übergreifend arbeitet und damit bessere Ergebnisse beim Finden ähnlicher Worte liefert. Nachdem diese ähnlichen Worte gefunden wurden, werden sie absteigend nach ihrem Ähnlichkeitswert sortiert, damit nur Wörter mit der höchsten Ähnlichkeit für die spätere Berechnung der Gruppen in Frage kommen. Werte liegen bei dieser Berechnung zwischen 0 und 1, mit einem Wert von 1 liegt die größte Übereinstimmung vor und passt somit am ehesten zum jeweiligen Startwort. Diese ähnlichen Worte werden später in einem eigenen Algorithmus verwendet, um Gruppierungen daraus zu erstellen. Die verwendeten Parameter für die Berechnung per Word2Vec sind [ŘS10]:

size = 100: Dimension der Wortvektoren, also die maximale Anzahl an Wörtern die bei der Berechnung im Bezug zum ausgewählten Wort stehen window: Distanz zwischen dem ausgewählten Wort und dem vorgeschlagenen Wort innerhalb eines Satzes, der Wert kann von den Nutzern durch „before/after words for groups“ auf der Hauptseite gesetzt werden

`min_count = 10`: minimale Anzahl an Vorkommen des Wortes im gesamten Korpus

workers: Anzahl an Prozessen die gleichzeitig zur Berechnung genutzt werden, wird durch die vorhandene Zahl an CPU Kernen ermittelt

4.3 Datenquelle und Verarbeitung der Daten

Der Datensatz im JavaScript Object Notation (JSON) Format wird von der Abteilung Grafisch-Interaktive Systeme (GIS) vom Institut für Visualisierung und Interaktive Systeme (VIS) an der Universität Stuttgart bereitgestellt und umfasst 170000 Online Nachrichten Artikel von sechs verschiedenen Nachrichtenseiten (1,3 Gigabyte (GB) an Daten). Die Nachrichtenseiten sind BBC News, AP News, Fox News, Al Jazeera English, The Guardian und RT News. Somit sind zwei amerikanische Seiten, zwei britische Seiten, eine russische Seite und eine Seite aus dem arabischen Raum vertreten. Die Daten umfassen den Zeitraum von Januar 2019 bis Juni 2019 und sind in englischer Sprache verfasst.

Nach Einlesen der einzelnen Artikel aus der JSON Datei werden diese zur leichteren Verarbeitung von Groß- und Kleinschreibung in Kleinschreibung umgewandelt und durch die Filter für Duplikate, das im Prototyp gesetzte Suchwort wie auch den gewählten Zeitraum für Nachrichtenartikel geführt. Alle verarbeiteten Artikel werden in einem großen Korpus hinzugefügt, der später den Trainingssatz darstellt. Gefilterte Artikel, die im Prototyp definierte Kriterien erfüllen werden ihrer Nachrichtenseite entsprechend in einem kleineren Korpus für diese Seite gespeichert. Auf Basis dieser Datensätze werden anschließend die Berechnungen für TF-IDF und die einzelnen Gruppen genutzt.

4.4 Prototyp Design

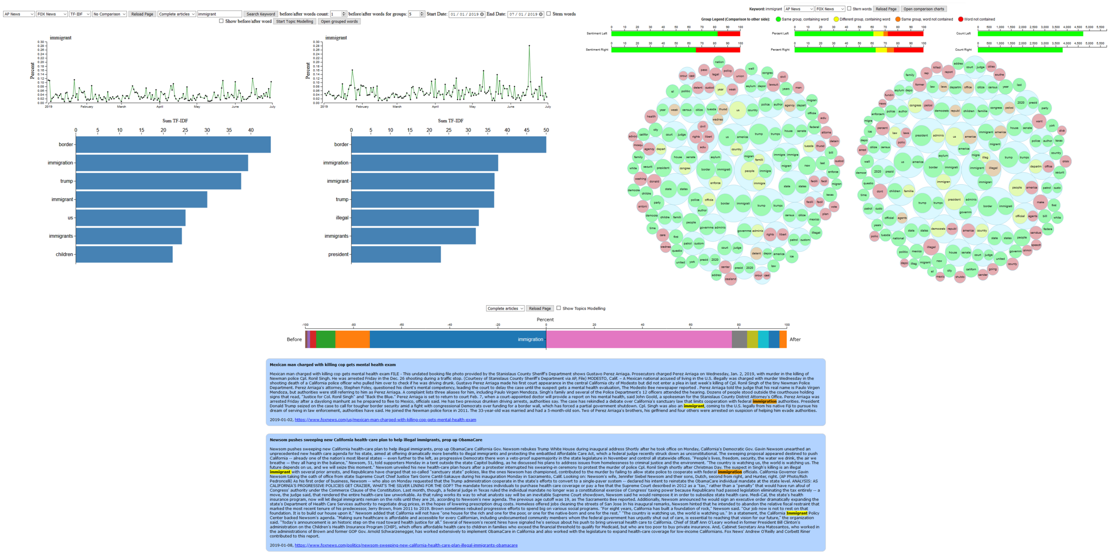

Abbildung 4.1: Die verschiedenen Seiten des Prototyps zur Analyse der Wortverwendungen

Die Visualisierung im Prototyp ist zum großen Teil auf zwei der drei in Abbildung 4.1 zu sehenden Seiten aufgebaut. Für den Vergleich von zwei Nachrichtenmedien werden die Seiten des Prototyps vertikal in zwei Hälften aufgeteilt, die Visualisierungen der linken Seite betreffen Ergebnisse des links ausgewählten Nachrichtenmediums, während die rechte Seite Ergebnisse des rechts ausgewählten Nachrichtenmediums darstellen.

Die Hauptseite des Prototyps stellt Balkendiagramme für die verschiedenen TF-IDF Werte der Wörter und Nachrichtenmedien dar, die zweite Seite die Gruppierungen der Wörter der jeweiligen Nachrichtenmedien. Per Klick auf eines der angezeigten Wörter erhält man Nachrichtenartikel für die Nachrichtenseite und das Suchwort, die dieses angeklickte Wort enthalten. Ausgangspunkt jeder Analyse für ein bestimmtes Suchwort oder Thema ist immer die Balkendiagrammseite, siehe Abbildung 4.3.

Die einzelnen Seiten des Prototyps (siehe Abbildung 4.1) haben unterschiedliche Einstellungsmöglichkeiten für die Nutzer. Auf der Balkendiagrammseite sind neben dem Auswählen der beiden Nachrichtenseiten, die verglichen werden sollen, die Angabe des gewünschten Suchworts wie auch eine Eingrenzung des Zeitraums möglich, wie in Abbildung 4.2 zu sehen ist. Die Auswahl der linken Nachrichtenseite betrifft die linke Seite der Ergebnisse und die rechte Auswahl die Ergebnisse der rechten Seite. Für die Suche nach einem bestimmten Suchwort sind zusätzlich die Angaben von Suchradien für Wörter vor oder nach dem Suchwort, wie auch der Suchradius für die Gruppierung der Wörter vorhanden.

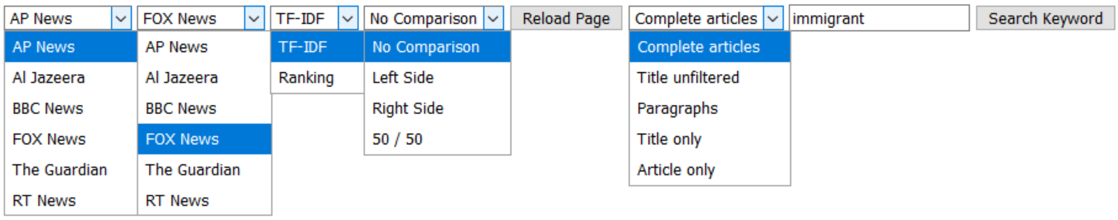

Abbildung 4.2: Verschiedene Einstellungsmöglichkeiten die es im Prototyp gibt

Über die standardmäßig auf „Complete articles“ gesetzte Auswahl kann zusätzlich angegeben werden ob die vollständigen Artikel verarbeitet werden oder eine spezifischere Auswahl stattfindet. Bei der standardmäßigen Auswahl wird jeder Artikel verwendet, der das gesuchte Wort entweder im Titel oder Artikel enthält. Bei der Auswahl von „Title unfiltered“ wird nach dem gesuchten Wort im Nachrichtentitel gesucht und falls es dort enthalten ist der gesamte Artikel verwendet. Falls das gesuchte Wort im Artikeltext enthalten ist wird nur wie in der Option „Paragraphs“ ein Teilausschnitt des Artikels verwendet. Über die Option „Paragraphs“ werden Sätze eines Artikels vor, nach sowie mit dem gesuchten Wort in die Auswertung übernommen. Bei der Wahl von „Title only“ muss das gesuchte Wort im Nachrichtentitel vorkommen, um den Artikel zu verwenden, mit der Option „Article only“ gilt genau der gegensätzliche Fall, dass das Wort im Artikel vorkommen muss. Die Optionen sind so gewählt, da bei vorkommen des Wortes im Nachrichtentitel eher davon ausgegangen werden kann, dass sich der Artikel auch zumindest zum Teil mit diesem Wort auseinandersetzt. Bei Vorkommen im Artikel kann das Wort „nur“ ein genutztes Wort sein, aber möglicherweise nichts über das Artikelthema aussagen.

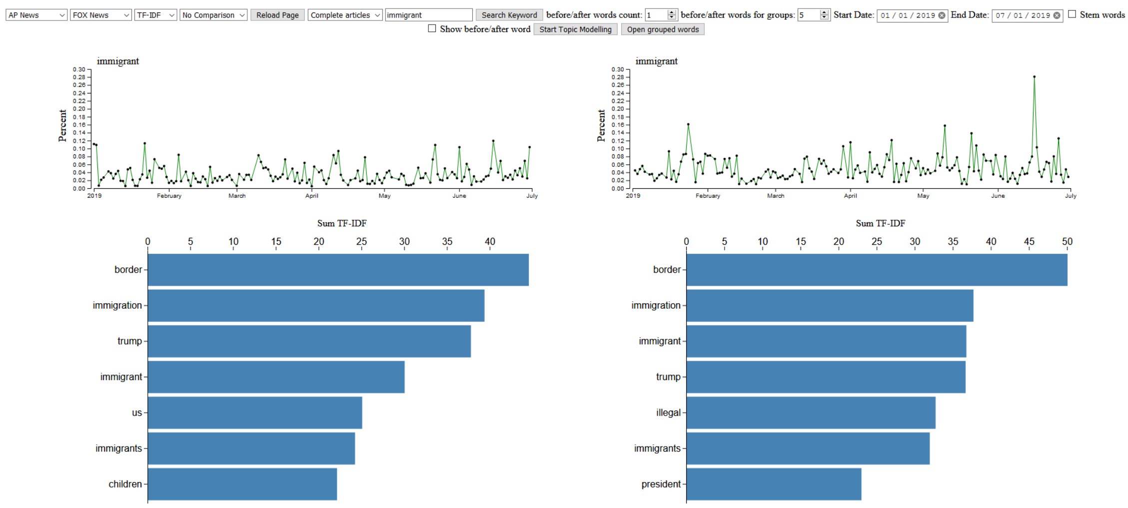

Abbildung 4.3: Die Hauptansicht des Prototyps mit den Einstellungsmöglichkeiten und Diagrammen

Mit Auswählen der Option „Stem words“ wird bei der Verarbeitung der Artikel eine Wortstammverkürzung vorgenommen, wodurch verschiedene Formen des gleichen Wortes in ihrem Stamm zusammengefasst werden. Durch Klick auf „Search Keyword“ wird die Berechnung der Ergebnisse für alle vorhandenen Nachrichtenseiten und Algorithmen ausgeführt und mögliche Gruppen berechnet. Nach Beenden dieser Berechnung wird die Seite neu geladen und die Ergebnisse in der Visualisierung aufbereitet dargestellt.

Weitere Optionen, die nach Berechnung des Ergebnisses gewählt werden können, sind die Auswahl „TF-IDF“ oder „Ranking“. Bei der Wahl von TF-IDF werden die einzelnen Worte nach Größe des TF-IDF Werts absteigend sortiert. Somit steht das Wort mit dem größten TF-IDF Wert ganz oben und danach die immer jeweils kleineren, bis 100 Wörter dargestellt sind. Durch die Wahl von „Ranking“ werden Wörter hervorgehoben, die eine große Differenz bei beiden Nachrichtenseiten für die Wörter aufweisen.

Durch das Ranking wird versucht, stark voneinander abweichende Werte im Vergleich zwischen den Nachrichtenmedien hervorzuheben. Ein häufig auf der einen Seite verwendetes Wort kann möglicherweise dafür genutzt werden, um ein Framing hervorzurufen. Daher werden diese Unterschiede durch das Ranking gesucht und anhand des TF-IDF Wertes wie in Abbildung 4.4 angezeigt. Die genutzte Formel für das jeweilige Wort ist:

| Worthäufigkeit links – Worthäufigkeit rechts | * IDF Wert des jeweiligen Wortes

Durch den Betrag auf das Ergebnis der Worthäufigkeiten ist dieser Wert für beide Seiten gleich, der Unterschied in den Ranking Werten entsteht durch den IDF Wert der jeweiligen Seite. Falls das Ergebnis der Worthäufigkeiten 0 ist wird nur der IDF Wert zurückgegeben. Der IDF Wert betont Wörter, die nicht in Masse im gesamten Korpus vorhanden sind, sondern eher Wörter die seltener auftreten und so theoretisch ein Thema beschreiben können. Nach der Sortierung anhand dieses Rankings werden in der Visualisierung die TF-IDF Werte angezeigt und diese sollten sich zwischen den beiden Nachrichtenmedien deutlich voneinander abgrenzen, was die Höhe der Werte angeht, siehe Abbildung 4.4.

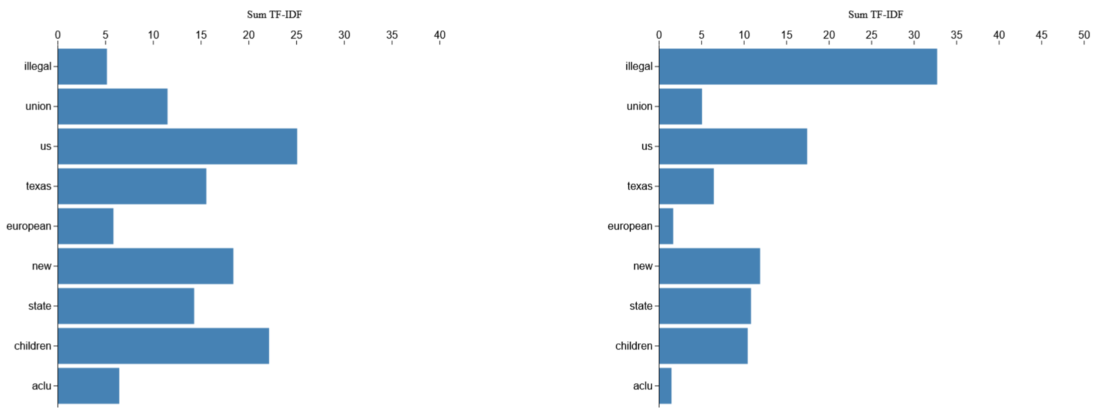

Abbildung 4.4: Anwendung des Rankings um größere Abweichungen bei den Wortverwendungen zwischen den Nachrichtenmedien hervorzuheben

Mit der Option „No Comparison“ werden beide Nachrichtenmedien *nicht* miteinander verglichen, die einzelnen TF-IDF Werte der Ergebniswörter sind also unabhängig von der anderen Seite. Mit der Auswahl von „Left Side“ oder „Right Side“ ändert sich dies, dabei werden die jeweils auf der bestimmten Seite angegebenen Wörter auch auf der anderen Seite gegenübergestellt, um eine leichtere Vergleichbarkeit zu ermöglichen. Durch die Wahl von „50 / 50“ werden die 50 größten Werte beider Seiten kombiniert, sortiert und angezeigt.

Das vorher bereits genannte Kontrollkästchen „Stem words“ kann auch nach der Berechnung eines Ergebnisses benutzt werden, dabei werden die 100 berechneten Wörter auf ihren Wortstamm reduziert und die entstandenen Lücken der wegfallenden Wörter bis zur Anzahl 100 wieder durch nachfolgende Wörter aufgefüllt, wie in Abbildung 4.5 zu sehen ist. Der Unterschied hier ist, dass die Berechnung auf die vorhandenen TF-IDF Werte vorgenommen wird, anstatt direkt bei der Berechnung aus den Artikeln für TF-IDF Werte einzufließen.

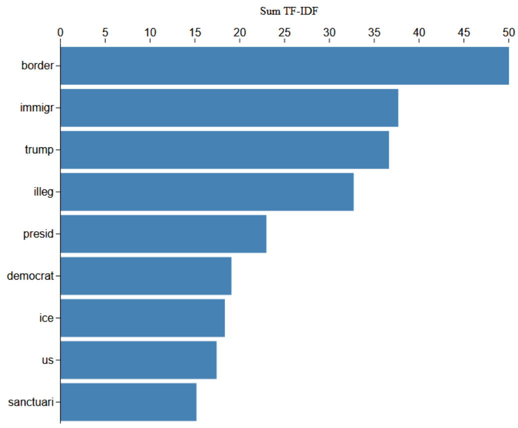

Abbildung 4.5: Ein Beispiel der auf ihren Wortstamm reduzierten Wörter

Durch das Kontrollkästchen „Show before/after word“ können im Balkendiagramm die am häufigsten vorkommenden Wörter vor und nach dem Wort in Prozent angezeigt werden (siehe Abbildung 4.6). Dadurch erhält man wie bei den beiden vorgestellten Prototypen eine Einsicht darauf, welche Wörter oft im direkten Zusammenspiel auftreten, was auf ein bestimmtes Framing hindeuten kann.

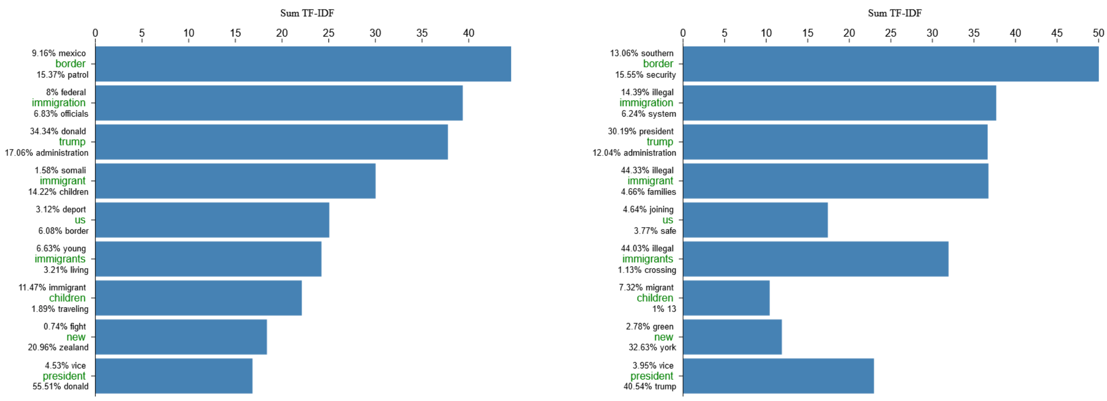

Abbildung 4.6: Prozentuale Anzeige der oft vor oder nach einem Wort stehenden Wörter

Über die Schaltfläche „Start Topic Modelling“ wird eine Latent Dirichlet Allocation (LDA) Berechnung durchgeführt, die auf den gesamten Korpus angewendet wird. Dadurch werden die verschiedenen Artikel der Nachrichtenmedien analysiert und anhand ihrer Ähnlichkeit zu bestimmten Themenbereichen zugeordnet, sodass eine Verteilung über die verschiedenen Themenbereiche dargestellt werden kann.

Per Klick auf „Open grouped words“ wird die zweite Visualisierung aufgerufen, um die Gruppierungen der Wörter, wie auch weitere Informationen wie das Sentiment der Artikel für die Nachrichtenseiten anzuzeigen, siehe Abschnitt 4.4.3.

4.4.1 Hauptseite

Im oberen Viertel der Hauptseite ist ein Datumsdiagramm enthalten, das den Verlauf zeigt, zu welchem Zeitpunkt Artikel über das angezeigte Wort veröffentlicht wurden und wie hoch der prozentuale Anteil des Wortes am gesamten Artikel ist, vergleiche Abbildung 4.7. Die Darstellung als prozentualer Wert ist so gewählt, um bei größeren Datenmengen erkennbar zu machen, dass an einem bestimmten Tag für die Nachrichtenseite vermehrt dieses Wort in Artikeln vorkam. Zusätzlich reduziert die prozentuale Angabe eine Verfälschung des Ergebnisses bei unterschiedlich langen Texten. Ist ein Artikel 2000 Wörter lang und enthält 20-mal das bestimmte Wort ist dies prozentual dasselbe wie ein Artikel mit 200 Wörtern und 10 Vorkommen des Wortes.

Im direkten Vergleich mit der anderen Nachrichtenseite kann man dann erkennen, zu welchem Zeitpunkt viel über das Wort/Thema berichtet wurde. Beim Mouseover über einem der Zeitpunkte im Diagramm werden weitere Informationen für diesen Tag angezeigt, zum einen das Datum, das durchschnittliche Vorkommen des Wortes an diesem Tag (inkludiert alle Artikel an diesem Tag, egal ob das Wort darin vorkommt oder nicht), die durchschnittliche Textlänge eines Artikels und die durchschnittliche Anzahl an Vorkommen des Wortes in Artikeln.

Zusätzlich wird die Anzahl an Artikeln mit oder ohne das Wort angegeben, dadurch kann zu einem gewissen Teil abgeschätzt werden wie präsent das Thema an einem bestimmten Tag war. Mit Klick auf einen der Datenpunkte im Diagramm erhält man eine Artikelübersicht mit dem gewählten Wort und dem ursprünglichen Suchwort an diesem bestimmten Tag.

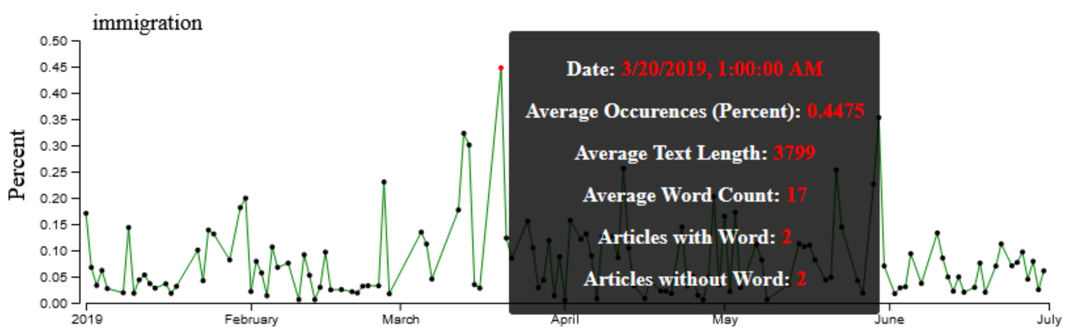

Abbildung 4.7: Genauere Informationen über das Wort „immigration“ am 20.03.2019

Im Balkendiagramm werden die größten TF-IDF Werte absteigend angezeigt, was eine gewisse Tendenz an Wortverwendungen bei den Nachrichtenseiten aufzeigen kann. Durch Mouseover über einem der dargestellten Wörter wird dieses Wort in beiden Balkendiagrammen hervorgehoben und Informationen über den TF-IDF Wert, die Anzahl an Vorkommen des Wortes, wie auch die prozentuale Größe in Relation zum größten TF-IDF Wert der Nachrichtenseite wie in Abbildung 4.8 zu sehen ist, angezeigt.

Durch das Hervorheben des ausgewählten Wortes auf beiden Seiten ist einfacher zu erkennen, wo sich das Wort auf der anderen Seite befindet, falls es enthalten ist. Nach längerem Mouseover über einem der Balken aktualisieren sich beide Datumsdiagramme auf das Wort des ausgewählten Balkens und zeigen entsprechend bei einem Mouseover über einem der Datenpunkte die aktualisierten verfügbaren Informationen für dieses Wort und den Datenpunkt an, wie in Abbildung 4.7 zu sehen ist.

Um die Vergleichbarkeit beider Seiten bei unterschiedlicher Artikelanzahl zu gewährleisten, wird die Nachrichtenseite mit der niedrigeren Artikelanzahl auf die der anderen Seite angepasst. Dabei wird die Artikelanzahl hoch skaliert, während sie auf der anderen Seite gleich bleibt. Die Formeln dafür sind:

$$\text{linke Seite} = \frac{\text{Anzahl Artikel rechts}}{\text{Anzahl Artikel links}} \quad \text{oder} \quad \text{rechte Seite} = \frac{\text{Anzahl Artikel links}}{\text{Anzahl Artikel rechts}}$$

Nur Werte über 1 werden zum Skalieren genutzt. Auf der skalierten Seite werden dann als zusätzliche Informationen beim Mouseover der skalierte wie auch der nicht skalierte Wert für TF-IDF und Worthäufigkeit angezeigt. Durch Klick auf den ausgewählten Balken werden Artikel für dieses Wort und das Suchwort angezeigt.

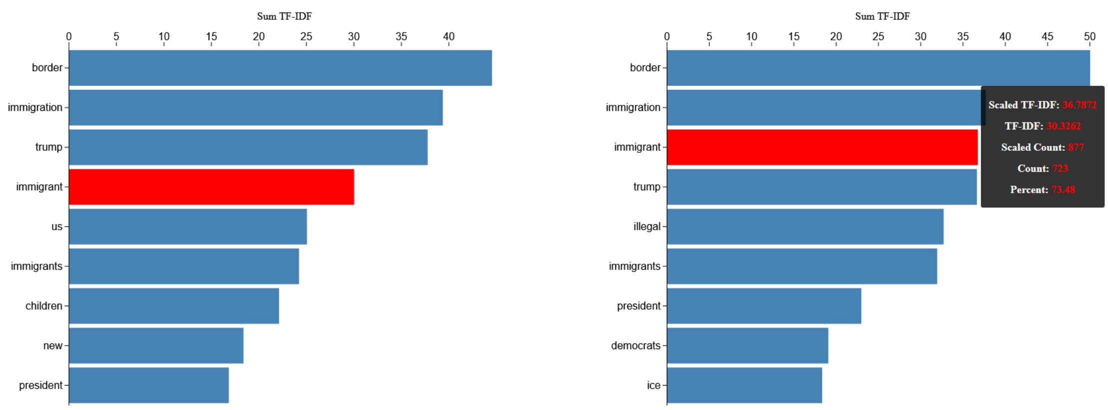

Abbildung 4.8: Hervorheben des ausgewählten Balkens und anzeigen der TF-IDF Werte oder Worthäufigkeiten zum Wort „immigrant“

4.4.2 Artikelseite

Auf der Artikelseite werden alle Artikel für das angeklickte Wort, das Suchwort und falls angegeben für ein bestimmtes Datum angezeigt. Die beiden Wörter werden in den Artikeln hervorgehoben dargestellt. Am oberen Ende der Seite wird außerdem wie in Abbildung 4.9 zu sehen ist, die Verteilung der Wörter vor und nach dem Suchwort angezeigt, ähnlich zum Balkendiagramm, aber diesmal mit bis zu 5 Wörtern und visueller Darstellung.

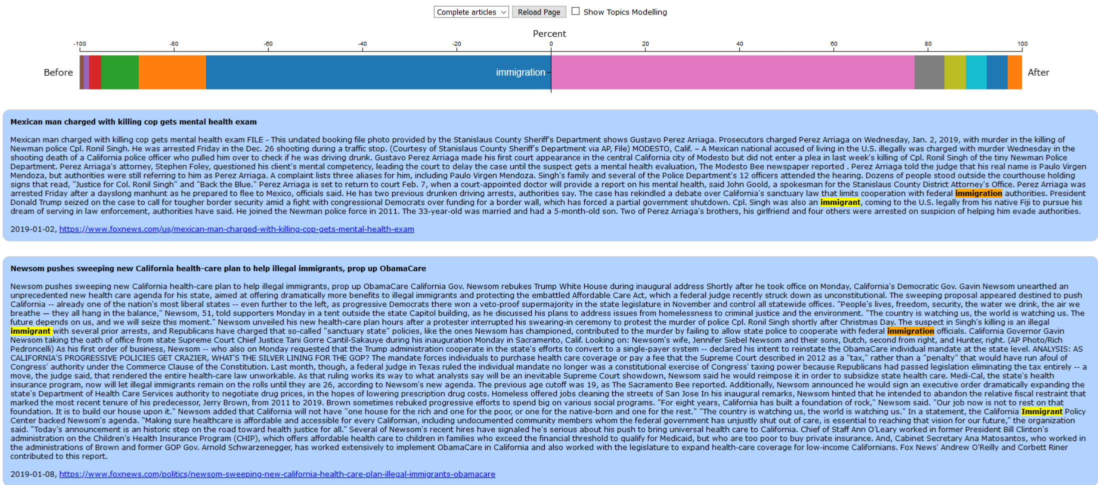

Abbildung 4.9: Neben den häufig vor oder nach dem Wort vorkommenden Wörtern, die Anzeige aller Artikel die die gesuchten Wörter enthalten

Die zusätzliche Option „Show Topics Modelling“ zeigt falls LDA Themen berechnet wurden 30 Themen mit beschreibenden Wörtern an, sowie die Verteilung der Artikel über diese Themen. Über die Auswahl „Complete articles“ wie auf der Hauptseite des Prototyps werden die kompletten Artikel angezeigt, durch die Option „Title unfiltered“ wird bei Vorkommen des angeklickten Wortes im Artikeltitel der gesamte Artikel angezeigt, ansonsten nur der Satz der das Wort enthält, wie auch der Satz vor und nach diesem Satz. Die Option „Paragraphs“ hat genau dieselbe Funktion, nur der Titel des Artikels wird nicht überprüft. Bei „Title only“ werden alle Artikel angezeigt, die das angeklickte Wort enthalten und mit „Article only“ werden die Textausschnitte angezeigt die im Artikel das ausgewählte Wort enthalten.

4.4.3 Gruppierungsseite

Beim Öffnen dieser Seite wird nach kurzer Berechnungszeit für die einstellbaren Seiten die Einteilung in die jeweiligen Gruppen angezeigt, siehe Abbildung 4.10. Die in einer Gruppe enthaltenen Wörter finden sich in den ähnlichen Worten zum Gruppennamen wieder, wie auch der Gruppenname sich in den ähnlichen Worten dieser Gruppenwörter befindet. Nur wenn Wörter sich gegenseitig enthalten werden sie in die jeweiligen Gruppen aufgenommen. Die Gruppen werden auf Grundlage der TF-IDF Top 100 Wörter der entsprechenden Nachrichtenseite berechnet. Für Wörter, die keine Übereinstimmung mit ihren ähnlichen Worten aufweisen können, werden einzelne Gruppen nur mit dem jeweiligen Wort erstellt. Es gibt zwei Einschränkungen bei diesem Algorithmus, jedes der

Wörter kann maximal einmal als Gruppenname wie auch in einer anderen Gruppe als Gruppenmitglied auftreten. Ist das Wort in einer anderen Gruppe enthalten, die eigene Gruppe aber leer oder identisch mit der anderen Gruppe wird die Gruppe des Wortes entfernt.

Zum besseren Verständnis ein Beispiel mit zwei Wörtern, für die ihre ähnlichen Worte berechnet wurden. Der Einfachheit wegen werden nicht alle enthaltenen Wörter mit ihren ähnlichen Worten hier dargestellt.

- Trump = [President, Donald, Administration, Organization, Campaign]
- President = [Republican, Vice, Donald, Trump, Barack]

Für diese zwei Wörter und ihre berechneten ähnlichen Worte wird nun versucht Gruppen anzulegen, begonnen wird mit dem Wort „Trump“. „Trump“ wird als Gruppenname angelegt und anschließend die ähnlichen Worte betrachtet. Das erste ähnliche Wort für „Trump“ ist „President“, somit wird beim Wort „President“ nachgesehen ob bei den ähnlichen Worten hier ebenfalls „Trump“ enthalten ist. Dies ist der Fall und somit wird zur Gruppe „Trump“ „President“ hinzugefügt. Nach diesem Muster wird für die weiteren Wörter verfahren und am Ende erhält man die folgenden Gruppen:

- Trump = [President, Donald, Administration, Organization]
- President = [Vice, Donald, Trump]

Daraus schließend kann gesagt werden, dass Trump nicht in den ähnlichen Worten für „Campaign“ enthalten war und für das Wort „President“ die Worte „Republican“ und „Barack“ ebenfalls nicht „President“ enthielten und somit sich nicht gegenseitig gefunden haben. Die Gruppennamen werden in der Visualisierung ebenfalls in den Gruppen dargestellt, da sie maßgeblich für die Gruppenzusammenstellung verantwortlich sind.

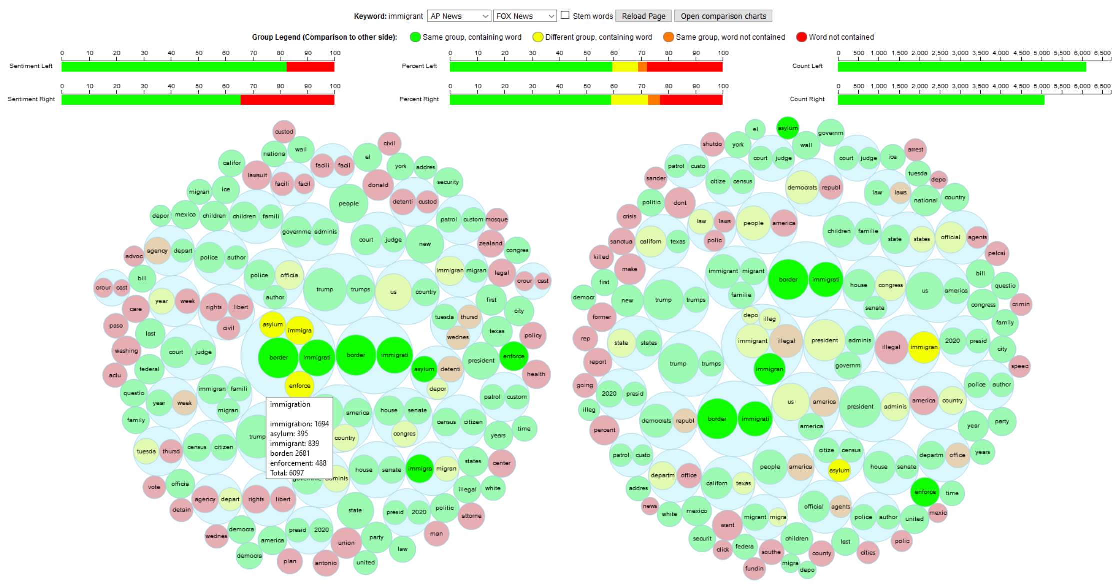

Abbildung 4.10: Für die Top TF-IDF Werte werden Gruppierungen angezeigt und mit den Gruppen des anderen Nachrichtenmediums verglichen

Die Farben der einzelnen Wörter geben an, wie dieses Wort auf der anderen Seite wiederzufinden ist. Grüne Farbe signalisiert, das Wort befindet sich in derselben Gruppe auf beiden Seiten. Bei gelb ist das Wort auf der anderen Nachrichtenseite enthalten, aber die Gruppe stimmt nicht überein. Orange bedeutet, das Wort ist zwar in einer gemeinsamen Gruppe enthalten, aber nicht in den Gruppen der anderen Nachrichtenseite wiederzufinden. Und wenn das Wort auf der einen Seite vorzufinden ist, aber nicht auf der anderen, ist die Farbe Rot.

Die einzelnen Größen der Kreise orientieren sich an der Worthäufigkeit des jeweiligen Wortes, je höher der Wert umso größer der Kreis, die Seiten werden wie auch schon beim Balkendiagramm per hochskalieren angeglichen. Durch Mouseover über einer Gruppe oder einem Wort werden die Gruppen oder Wörter auf beiden Seiten hervorgehoben, bei längerem Mouseover werden Informationen zur Gruppe beziehungsweise dem Wort für die Worthäufigkeit angezeigt. Mit Klick auf eine der Gruppen oder eines der Wörter werden Artikel für das Suchwort und den angeklickten Gruppennamen oder das angeklickte Wort angezeigt.

Die drei zusätzlichen Diagramme im oberen Teil der Darstellung zeigen die prozentuale Verteilung für das Sentiment der Nachrichtenartikel, die prozentuale Verteilung der Farben für Gruppierungen und die Worthäufigkeiten der hervorgehobenen Gruppen. Zur leichteren Vergleichbarkeit sind die jeweiligen Diagramme übereinander angeordnet, dabei ist das Diagramm der linken Seite oben und das der rechten Seite unten.

Mit dem Diagramm auf der linken Seite (Abbildung 4.11) wird die prozentuale Verteilung für positives, neutrales oder negatives Sentiment der Nachrichtenartikel des jeweiligen Nachrichtenmediums angezeigt. Für die Einordnung der Artikel in die entsprechende Kategorie werden die Artikel per Algorithmus auf Verwendung positiver, neutraler und negativer Worte analysiert und entsprechend einer dieser Kategorien zugewiesen. Die Anzahl und prozentuale Verteilung der Artikel in der jeweiligen Kategorie werden anschließend im Sentiment Diagramm dargestellt. Durch diese Verteilung kann ein Eindruck darüber entstehen, wie ein Nachrichtenmedium über ein Thema berichtet, welche Stimmung erzeugt werden soll und somit eine Beeinflussung des Lesers stattfinden kann. Da die Diagramme für beide Nachrichtenseiten untereinander angeordnet sind, kann hier sehr schnell eine Schlussfolgerung entstehen, welches der beiden Nachrichtenmedien positiver oder negativer über ein Thema berichtet.

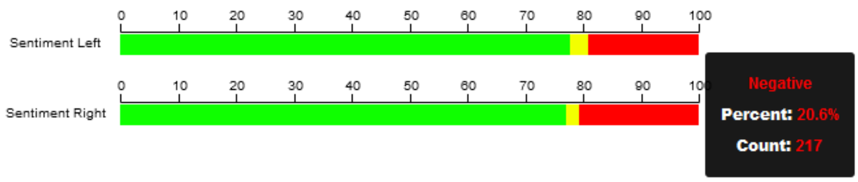

Abbildung 4.11: Prozentuale Verteilung des positiven, neutralen oder negativen Sentiments der Nachrichtenartikel

Das Diagramm in der Mitte zeigt die prozentuale Verteilung für die Farbe der Gruppen an. Dadurch erhält man einen schnellen Überblick darüber, welche Farben wie häufig auf beiden Seiten vertreten sind und ob die einzelnen Gruppen der beiden Seiten häufig übereinstimmen, oder es größere Abweichungen zwischen den Seiten gibt. Eine größere Anzahl an Abweichungen kann auf

Unterschiede in der Berichterstattung beziehungsweise vorhandene Framings hinweisen. Durch Mouseover über der Darstellung erhält man die Anzahl der Wörter, die eine bestimmte Farbe haben, siehe Abbildung 4.12.

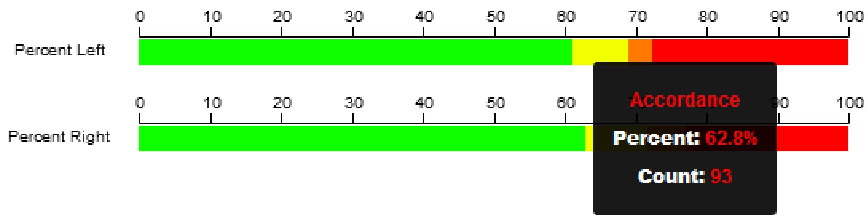

Abbildung 4.12: Prozentuale Verteilung der Farben für die Wörter in den Gruppen

Auf der oberen rechten Seite wird nach einem ersten Mouseover über einem Wort oder einer Gruppe das Worthäufigkeit-Diagramm für beide Seiten angezeigt, bei dem je nach Worthäufigkeit für das Wort oder die einzelnen Wörter der Gruppe die Diagramme dargestellt werden, wie in Abbildung 4.13 zu sehen ist. Die farblichen Unterschiede der Gruppen und Wörter werden dabei auch dargestellt. Durch Mouseover über diesem Diagramm erhält man noch einmal eine Anzeige über die einzelnen Wörter einer Gruppe inklusive ihrer Worthäufigkeit, oder für ein einzelnes Wort für dieses allein. Durch Skalierung wird hier eine gewisse Vergleichbarkeit ermöglicht, um besser sehen zu können wie häufig ein Wort auf der einen Seite genutzt wird im Vergleich zur anderen.

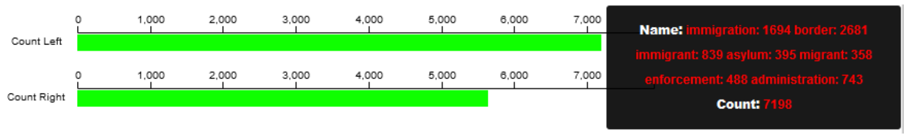

Abbildung 4.13: Aufaddierte Darstellung der einzelnen Worthäufigkeiten in den ausgewählten Gruppen

Als Kontrollkästchen „Stem words“ gibt es wieder die Möglichkeit, die Top 100 TF-IDF Worte vor dem Berechnen der Gruppen auf ihren Wortstamm zu reduzieren, um so verwandte Wörter und Plurale zu entfernen und wie in Abbildung 4.14 anzuzeigen. Mit Klick auf „Open comparison charts“ gelangt man wieder zurück zur Balkendiagrammseite.

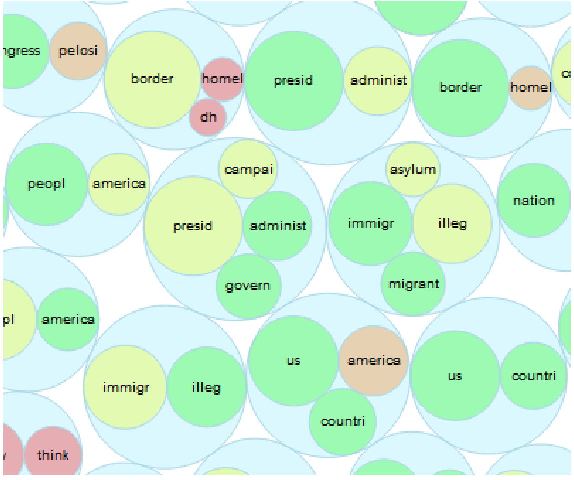

Abbildung 4.14: Auf Wortstamm reduzierte Wörter in den einzelnen Gruppen

4.5 Arbeitsablauf im Prototyp

Nach dem Aufrufen der Adresse des laufenden Prototyp Servers gelangt man auf die Hauptseite des Prototyps. Nach Eingabe eines gewünschten Themas als Suchwort und Klick auf „Search Keyword“ wird nach der Berechnung, die je nach System, Datensatz und gewählten Parametern unter einer Minute oder bis zu einer Stunde dauern kann, das Ergebnis zu diesem Suchwort angezeigt. Zum einen kann danach analysiert werden welche Wörter auf beiden Seiten häufig auftreten und welche Wörter in ihrem Kontext davor oder danach verwendet werden. Durch das Auswählen der Option „Ranking“ können stark voneinander in den Werten abweichende Worte angezeigt werden, die eventuell eine bestimmte Beeinflussung der jeweiligen Nachrichtenmedien zeigen. Außerdem kann im Datumsdiagramm nachvollzogen werden, wann sehr häufig in Artikeln über das Suchwort berichtet wurde.

Mit Klick auf „Open grouped words“ gelangt man zu den Gruppierungen der Wörter. Diese Gruppen und die Unterschiede in Größe, Farbe und enthaltenen Wörtern im Vergleich zum anderen Nachrichtenmedium können analysiert werden und daraus bestimmt werden welche Wörter für einen bestimmten Themenbereich genutzt wurden. Durch Anzeige der Worthäufigkeit-Diagramme kann nachvollzogen werden wie häufig diese Wörter in den jeweiligen Medien verwendet wurden. Mit dem Sentiment Diagramm kann eingeschätzt werden wie die Stimmung über alle Artikel hinweg für das Suchwort und die jeweilige Nachrichtenseite ist.

Anhand dieser vorhandenen Informationen auf beiden Seiten des Visualisierungs Prototyps sollen in Kombination Rückschlüsse auf bestimmte Framings der Nachrichtenmedien möglich sein.

4.6 Limitationen des Prototypen

Durch das gleichzeitige Verarbeiten der einzelnen Nachrichtenmedien aus dem 1,3 GB großen Datensatz wird mindestens 16 GB Arbeitsspeicher empfohlen. Mit Eingrenzen des Zeitraums können auch kleinere Arbeitsspeicher verwendet werden. Die Prototyp-Seiten sind für die Darstellung mit einer Auflösung von 1920 x 1080 Pixel optimiert, durch Nutzung der Skalierung des jeweiligen Internetbrowsers können auch andere Auflösungen genutzt werden, dies ist aber nur für Auflösungen über der Angegebenen empfohlen. Es ist keine Kombination aus nicht direkt hintereinander vorkommenden Worten als Suchwörter möglich, sondern nur die direkte Kombination wie zum Beispiel „climate change“. Man kann somit dem Prototyp nicht mitteilen für das Wort „Greta“ und/oder „climate“ die Berechnung zu kombinieren und somit einen spezifischeren Bereich zum gewünschten Thema zu erhalten. Je größer die Datenmenge zum Suchwort ist, umso länger dauert die Berechnung des Ergebnisses. Da die Kalkulation für LDA Topic Modelling recht lange dauert (über 30 Minuten), muss dies Falls benötigt per zusätzlichem Klick auf „Start Topic Modelling“ auf der Hauptseite gestartet werden. Für den Fall, dass einem Rechner weniger als 4 Threads zur Verfügung stehen, werden die Algorithmen ohne Parallelisierung ausgeführt, was zu längeren Berechnungszeiten führt, dies kann aber im Python Code angepasst werden. Mehr als zwei Nachrichtenmedien können im Prototyp nicht direkt miteinander verglichen werden und die einzelnen Seiten des Prototyps nicht nebeneinander angezeigt werden, außer man nutzt mehrere Bildschirme oder verkleinerte Internetbrowser Fenster mit der jeweiligen Prototypseite.

5 Anwendungsfälle

5.1 Auswertung für das Suchwort „immigrant“

Für das Suchwort „immigrant“ kann man zwischen den beiden Nachrichtenmedien AP News und Fox News Unterschiede bei den Wortwahlen vor und nach den Wörtern „immigrant“, „immigration“, „immigrants“ und „illegal“ erkennen, siehe Abbildung 5.1. Bei AP News werden neutrale Wörter genutzt, die über bestimmte aufkommende Sachverhalte berichten, wie „young“, „children“, „somali“, „federal“. Bei Fox News wird dagegen sehr häufig mit einem hohen Prozentsatz von 30 oder 40 Prozent von „illegal immigrants“ oder „illegal immigrant families“ gesprochen. Diese hohen Prozentzahlen für die Verwendung des Wortes „illegal“ deuten recht stark auf eine Beeinflussung auf die Berichterstattung zum Negativen hin, zusätzlich steht das Wort „illegal“ im TF-IDF Wert für Fox News an fünfter Stelle während es bei AP News nicht in den Top 25 vorkommt.

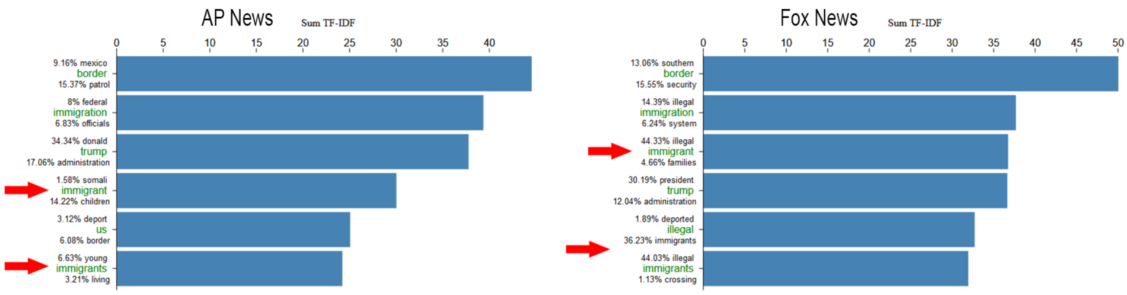

Abbildung 5.1: Linke Seite AP News im Vergleich zur Nachrichtenseite Fox News rechts. Unterschiede bei Verwendung von „illegal“

Verwendet man die Option „Ranking“ zur Analyse der TF-IDF Werte in Bezug auf die gerade erhaltenen Informationen, erkennt man einen deutlichen Unterschied in der Nutzung der Wörter „illegal“ und „children“, wie in Abbildung 5.2 zu sehen ist. Das Wort illegal hat einen sechs Mal höheren TF-IDF Wert bei Fox News und eine fünf Mal höhere Worthäufigkeit, was die deutlich höhere Verwendung des Wortes bei Fox News zeigt. Zusätzlich kann angemerkt werden, dass „illegal“ auf Seiten von AP News eher Sachbezogen ist da „illegal immigration“ die häufigste Kombination ist, während bei Fox News Personenbezogen mit „illegal immigrants“ berichtet wird. Das Wort „children“ wird bei AP News öfter verwendet, was wiederum auf eine neutralere beziehungsweise positivere Berichterstattung hindeuten kann.

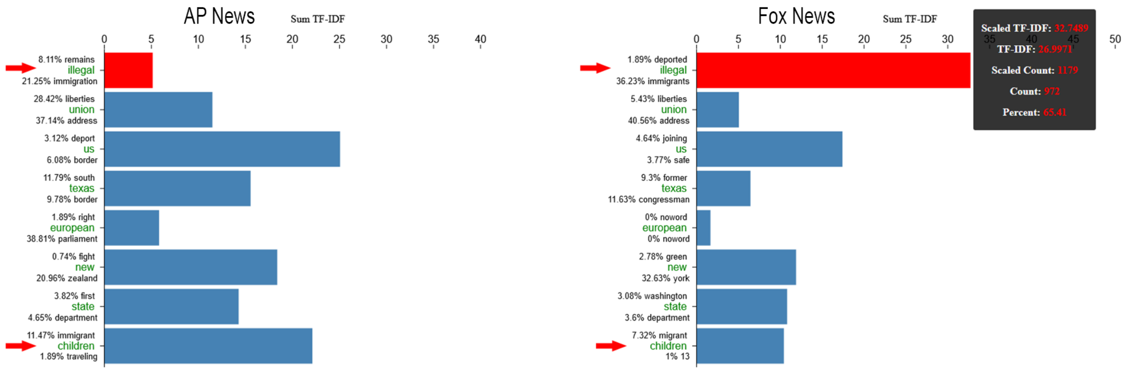

Abbildung 5.2: Das Ranking zeigt große Unterschiede bei der Verwendung der Wörter „illegal“ und „children“ zwischen AP News und Fox News.

Bei Ansicht der Gruppierungen in Abbildung 5.3 sieht man auf Seiten von Fox News auch eine Gruppe bestehend aus „immigrant“, „immigrants“, „migrant“, „illegal“ und „illegally“, während die vergleichbare Gruppe bei AP News die Wörter „illegal“ und „illegally“ nicht enthält. AP News enthält zusätzlich eine Gruppe mit den Wörtern „immigrant“, „immigrants“, „migrants“, „people“ und „families“, die wieder auf eine neutralere oder positive Berichterstattung hindeutet. Mit der Gruppe aus „civil“, „rights“, „liberties“ auf Seiten von AP News wird erneut deutlich, dass hier ein anderes Bild bei der Berichterstattung über das Thema „immigrant“ gewählt wurde, da bei Fox News diese Worte in den Gruppierungen nicht zu finden sind und übersetzt für Rechte und Freiheiten stehen.

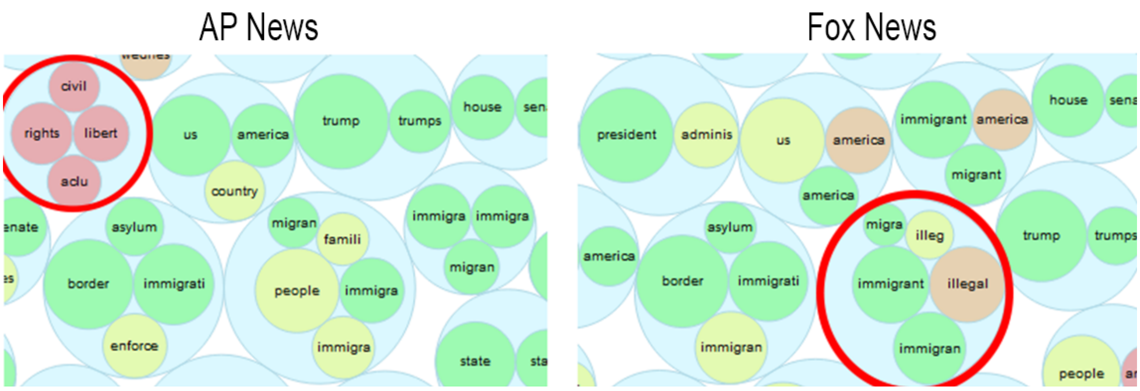

Abbildung 5.3: Auf der Seite von AP News zusätzliche Gruppe über Rechte und Freiheiten, auf der Seite von Fox News die Gruppe mit illegal

Auch im Sentiment Diagramm in Abbildung 5.4 erkennt man eine starke Tendenz zum Negativen bei Fox News (unterer Balken), da 34 Prozent der Artikel als negativ bewertet werden, während es bei AP News nur 17 Prozent sind. Die Nachrichten zum Thema „immigrant“ werden somit vom verwendeten Algorithmus für Fox News um einiges negativer bewertet, als bei AP News.

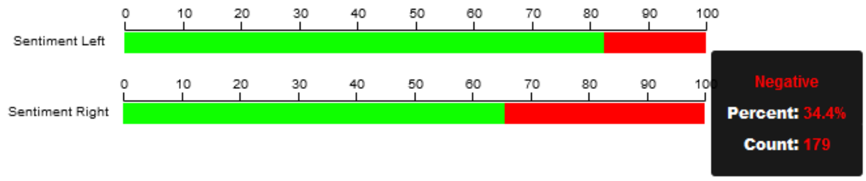

Abbildung 5.4: Vergleich der Sentiment Diagramme zeigt stärkere Tendenz zum Negativen bei Fox News (Sentiment Right Balken)

In der Gesamtbetrachtung kann festgehalten werden, dass Fox News eher negativ zum Thema „immigrant“ beziehungsweise „immigration“ berichtet, während AP News neutral oder positiv darüber berichtet.

5.2 Auswertung für das Suchwort „medicare“

Zum einen fällt der höhere TF-IDF Wert in Abbildung 5.5 für „medicare“ bei AP News im Vergleich zu Fox News auf und zum anderen das häufig danach auftretende Wort, welches bei AP News „beneficiaries“ ist. Hier geht es anscheinend um eine Begünstigung beim Gesundheitsdienst, während bei Fox News über den momentanen („current“) Zustand geschrieben wird. Ein Unterschied ist außerdem beim Wort „health“ zu sehen, da Fox News häufig über „private health care“ spricht, während bei AP News „universal health care“ am häufigsten angezeigt wird und somit etwas Positives für die Allgemeinheit ausstrahlt, im Vergleich zur privaten Vorsorge bei Fox News.

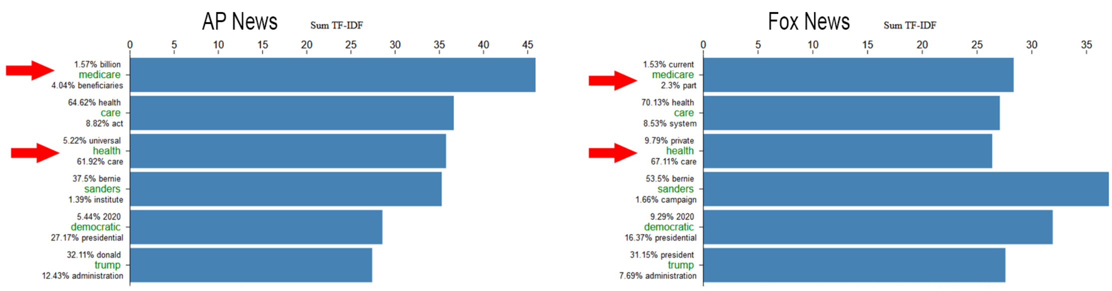

Abbildung 5.5: AP News im Vergleich zur Nachrichtenseite Fox News. Ergebnis für das Suchwort „medicare“

Bei Verwendung der Option „Ranking“ mit Sortierung auf die linke Seite, findet man die Wörter „medicare“ und „health“ im oberen Bereich wieder, was auf einen größeren Unterschied bei der Verwendung dieser Wörter zwischen den Nachrichtenmedien hindeutet, siehe Abbildung 5.6. Andere Wörter, die sehr unterschiedlich in den Nachrichtenmedien auftreten sind „medicare-for-all“, „socialism“ und „socialist“. Ein Grund für das unterschiedliche Vorkommen von „medicare-for-all“ kann der Schreibstil sein, durch die maschinelle Verarbeitung wird strikt von anderen Schreibweisen wie „medicare for all“ unterschieden. Mit Klick auf dieses Wort für die Seite Fox News erhält man alle Artikel für dieses Wort und bei genauerer Betrachtung wird hier oft auf die hohen Kosten für dieses Vorhaben der Gesundheitsvorsorge für alle eingegangen. In diesem Zusammenhang wird sehr oft auch die Politikerin Alexandria Ocasio-Cortez erwähnt, die für dieses Vorhaben einsteht. Ebenso fällt hier oft das Wort „socialism“ und „socialist“, was anscheinend in Amerika für etwas Negatives steht, aber von den Demokraten im Hinblick auf „medicare for all“ verteidigt wird.

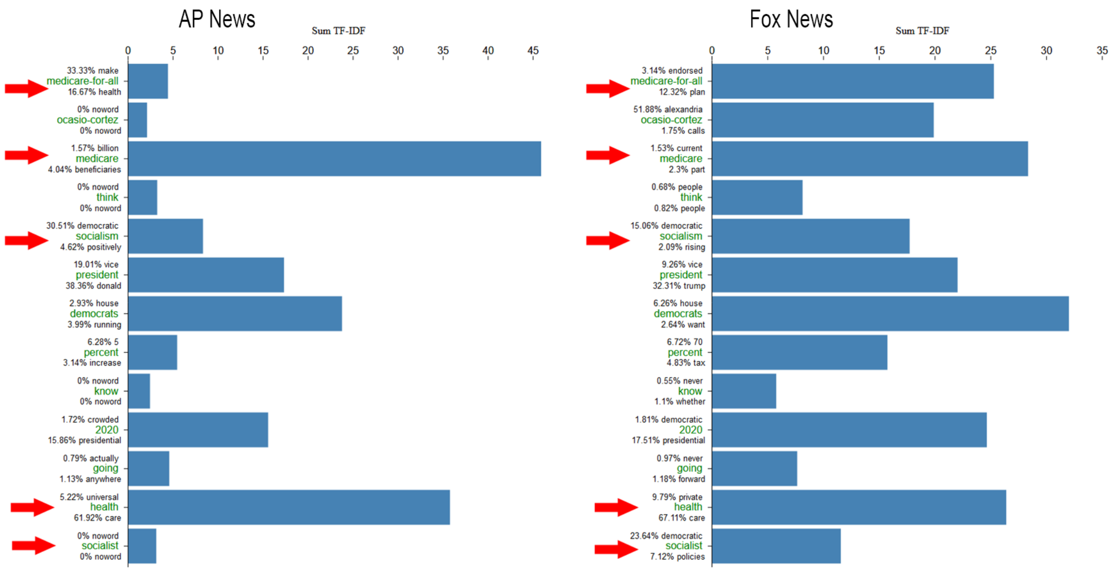

Abbildung 5.6: Große Unterschiede bei der Verwendung von Wörtern wie „socialism“ und „socialist“ zwischen AP News und Fox News.

Wie in Abbildung 5.7 zu sehen ist enthalten die Gruppierungen auf beiden Seiten jeweils interessante Gruppen, die für bestimmte Framings stehen können. Die Gruppe auf Seiten von AP News mit „prices“, „cost“, „costs“ und „premiums“ offenbart mit Klick die dazu passenden Artikel, in denen es darum geht, die Kosten der Gesundheitsvorsorge durch „medicare for all“ zu reduzieren. In der Gruppe von Fox News mit „socialist“, „socialism“ und „progressive“ wird per Klick erkennbar, dass die Demokraten sehr stark in die sozialistische Richtung geschrieben werden. Dies ist von den Demokraten auch zum Teil für die Durchsetzung der Gesundheitsvorsorge so beabsichtigt, nur wird dies beim Lesen der Texte als etwas Negatives dargestellt.

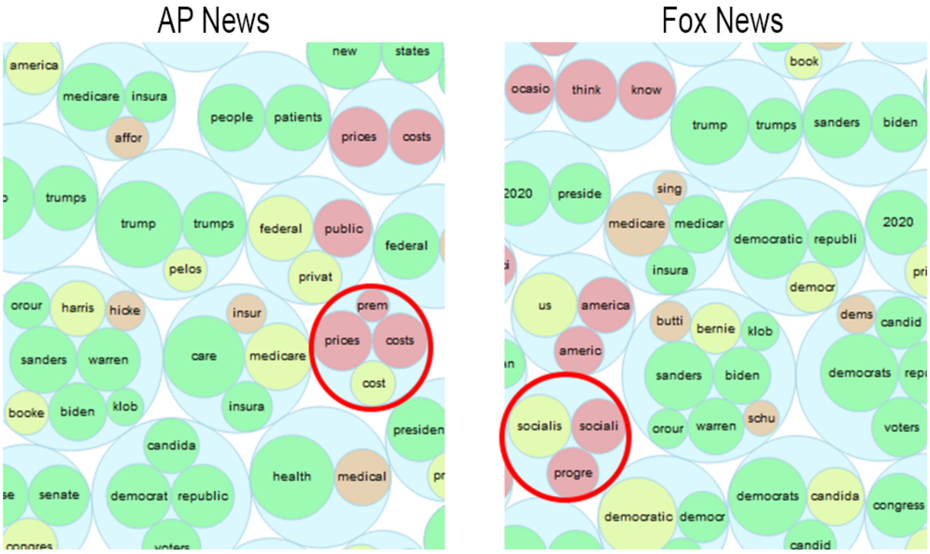

Abbildung 5.7: Auf der Seite von AP News zusätzliche Gruppe über Preise und Kosten, auf der Seite von Fox News die Gruppe mit Sozialismus

Im Sentiment Diagramm können keine größeren Unterschiede zwischen den beiden Nachrichtenmedien erkannt werden. Aus den Beobachtungen heraus kann man zum Schluss kommen, dass Fox News den Status Quo beim Gesundheitssystem bevorzugt und „medicare for all“ als etwas Negatives gesehen wird, AP News berichtet darüber eher neutral bis positiv.

5.3 Auswertung für das Suchwort „climate“

Da das Ergebnis für komplett ungefilterte Artikel nicht sehr aussagekräftig war, wurde hier der Filter „Title unfiltered“ gesetzt, wodurch Artikel mit dem Suchwort im Nachrichtentitel weiterhin ungefiltert in die Verarbeitung übernommen und bei Auffinden im Nachrichtentext Sätze vor, mit, und nach dem Suchwort übernommen wurden. Die größeren Unterschiede sieht man im Balkendiagramm bei weiter untenstehenden Worten und hierbei eher bei den Worten die vor diesen Worten stehen, siehe Abbildung 5.8. Das Wort „crisis“ hat bei The Guardian zu 44 Prozent das Wort „climate“ davorstehen, während es bei Fox News nur 14 Prozent sind. The Guardian spricht bei „people“ eher von jungen Leuten, während es bei Fox News amerikanische sind. Für das Wort „emergency“ ist bei The Guardian zu 56 Prozent das Wort „climate“ davorstehend, bei Fox News zu 50 Prozent „national“.

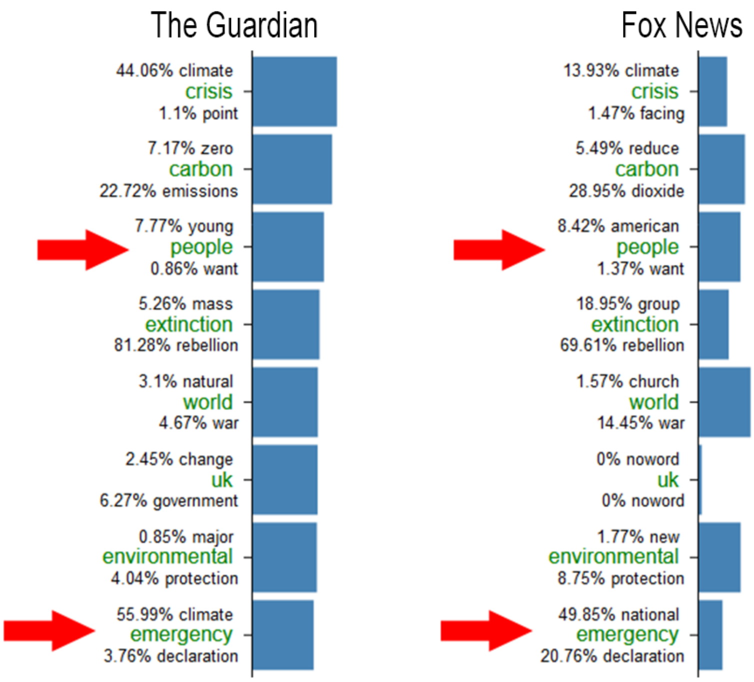

Abbildung 5.8: The Guardian im Vergleich zu Fox News zum Suchwort „climate“

Mit der Option „Ranking“ kommen wieder die angesprochenen Wörter „emergency“ und „people“ in Abbildung 5.9 zum Vorschein. Außerdem ist wie bei den TF-IDF Werten „carbon“ zu sehen, bei The Guardian geht es hier hauptsächlich um „zero carbon emissions“, die komplette Reduzierung des Kohlendioxids, bei Fox News um „reduce carbon dioxide“, also „nur“ die Reduzierung des Kohlendioxids. Da beides in dieselbe Richtung der Reduzierung geht, kann nicht genau gesagt werden ob hier ein Framing vorliegen könnte. Die Sentiment Diagramme zeigen ebenfalls keine auffälligen Unterschiede zwischen den beiden Nachrichtenmedien.

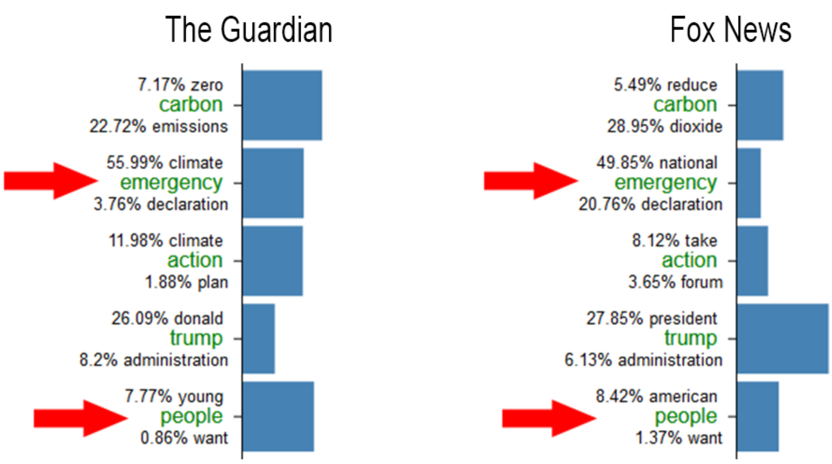

Abbildung 5.9: Das Ranking zeigt Unterschiede bei der Verwendung von „emergency“ und „people“ zwischen The Guardian und Fox News.

Gruppierungen auf beiden Seiten zeigen regionale Unterschiede auf, während auf der The Guardian Seite die Gruppen mit „strike“, „strikes“ und „protests“, „extinction“, „rebellion“, „thunberg“ vorhanden sind, ist die Gruppe „ocasio-cortez“, „biden“ und „sander“ bei Fox News zu finden, siehe Abbildung 5.10. Die Gruppe von Politikern bei Fox News steht im Zusammenhang mit dem Thema Klima, genauso wie die beiden Gruppen auf Seiten von The Guardian. Der Unterschied im Ranking für „carbon“ könnte daran liegen, dass bei The Guardian die Zahl 2050 mit Emissionen, „carbon“ und „greenhouse“ in einer Gruppe steht, da Extinction Rebellion die Auswirkungen des Klimawandels für 2050 thematisiert. Dieser Unterschied kann also daran liegen, dass die europäische The Guardian über diese Bewegung berichtet, während Fox News einen nicht so starken Fokus darauflegt. Ein regionales Framing kann hier unterstellt werden, ist aber nicht ganz untypisch für Berichterstattungen in den verschiedenen Regionen der Welt.

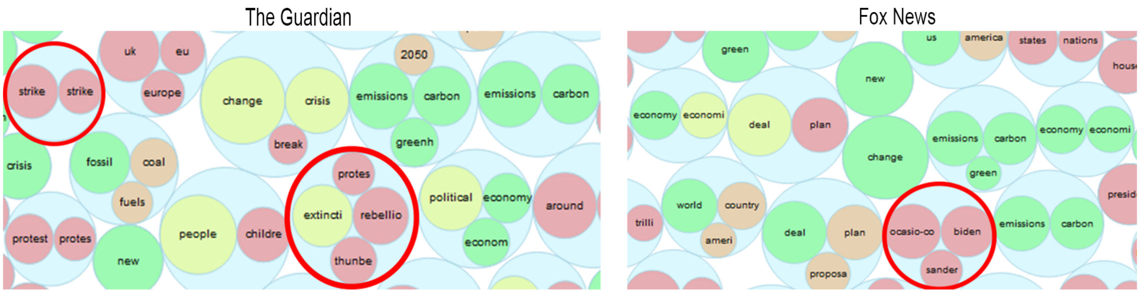

Abbildung 5.10: Bei The Guardian zusätzliche Gruppen über Proteste, Greta Thunberg und die Extinction Rebellion.

Einen wirklichen Rückschluss auf ein bestimmtes Framing lässt der Prototyp in diesem Fall nicht zu, da die Unterschiede zu gering und ein mögliches Framing auf regionale Unterschiede beziehungsweise fehlende Relevanz zurückzuführen sind. Das Thema Umweltschutz kam zu diesem Zeitpunkt in Europa erst ins Rollen, wobei dies auch eine persönliche Einschätzung der Datenlage darstellt.

6 Zusammenfassung und Ausblick

6.1 Zusammenfassung

Die Ergebnisse der Suchwörter, sowie Erkenntnisse, die man daraus ziehen kann, sind zum Teil stark von dem gewählten Thema und den einzelnen Parametern abhängig. Ist ein Thema sehr breit gefasst und somit die Datenmenge zu diesem Thema entsprechend groß, wird das Finden von Framings, falls sie existieren, erschwert. Mögliche Unterschiede zwischen den Nachrichtenmedien gehen im Durchschnitt der gesamten Datenmenge unter, daher ist es ratsam spezifischere Suchwörter zu verwenden und den Zeitraum, sowie die Art der Artikelverarbeitung (kompletter Artikel, nur Artikel mit dem Suchwort im Nachrichtentitel oder andere Möglichkeiten) zu verkleinern. Zusätzlich muss darauf hingewiesen werden, dass dargestellte Unterschiede in den Wortverwendungen nicht unbedingt auf ein bestimmtes Framing hindeuten müssen, da Unterschiede in den Datensätzen und eine andere Ausrichtung der Berichterstattung in den Medien bestehen können. Der Datensatz zu Fox News enthält im Vergleich zu AP News kaum Artikel zum Aktienmarkt, da dieser Teil der Berichterstattung in Fox Business stattfindet und nicht im Datensatz enthalten ist.

Diese Erkenntnisse können aber ebenfalls darauf hindeuten, dass der Prototyp sehr anfällig bei bestimmten Parametern ist und verfälschte Ergebnisse liefern kann. So ist zuallererst ein guter Datensatz nötig, der über alle gewünschten Nachrichtenmedien und Zeiträume genügend Artikel enthält, um Vergleiche anstellen zu können. Bei der Suche nach bestimmten Themen muss eine Eingrenzung des Themas vorgenommen werden, da sonst möglicherweise keine Unterschiede zwischen den zu vergleichenden Nachrichtenmedien erkennbar werden. Eine weitere Schwachstelle des Prototyps ist die lange Berechnungszeit (über 10 Minuten) bei größeren zu verarbeitenden Mengen, die das Suchwort enthalten.

Durch die beiden Visualisierungsseiten im Prototyp können Erkenntnisse der Visualisierungen kombiniert und so eine Analyse für mögliche Framings vereinfacht werden. Mit den verschiedenen Einstellungsmöglichkeiten kann das Ergebnis angepasst werden und somit durch diese genauere Definition ein Framing gefunden werden. Da der Programmcode über einen Python Server ausgeführt wird kann dieser ebenfalls auf einem Webserver gestartet werden, dadurch können größere Datenmengen verarbeitet und der Prototyp im Netzwerk leicht verfügbar gemacht werden.

Zusammenfassend kann gesagt werden, dass der Prototyp für die Erkennung von Frames verwendet werden kann, je nach Thema und Datensatz können die Ergebnisse aber besser oder schlechter ausfallen. Bei großen Datensätzen können Unterschiede in der Masse verloren gehen, durch eine spezifischere Definition des Themas und Zeitraums, sowie der anderen Parameter kann das Ergebnis verbessert werden. Die Ergebnisse sind aber ebenfalls davon abhängig ob überhaupt ein eindeutig zu beobachtendes Framing bestimmt werden kann, oder eben kein Framing für das Thema vorhanden ist. Dabei müssen auch regionale Unterschiede und andere Aspekte bei der Suche nach Framing beachtet werden, da zum Beispiel eine Suche nach „election“ für die amerikanischen und britischen

Nachrichtenmedien ein regional unterschiedliches Ergebnis ergeben und dadurch ein direkter Vergleich zwischen diesen Nachrichtenmedien nicht sinnvoll ist. Außerdem ist anzumerken, dass die gezeigten Ergebnisse und Erkenntnisse in dieser Arbeit nicht wie im Abschnitt Related Work mit definierten Fragestellungen und Kategorisierungen durchgeführt wurden und somit die persönlichen Einschätzungen zu möglichen Framings falsch sein können.

6.2 Ausblick

Mögliche Verbesserungen für den Prototyp können die Verwendung anderer Python Pakete oder Algorithmen für einzelne Berechnungen sein oder eine eigene Implementierung dieser, sowie die Umsetzung anderer Ideen und Anpassungen am Prototyp. Eine Möglichkeit der Erweiterung ist beispielsweise die Analyse kontroverser Wörter, dabei werden die Artikel der Nachrichtenmedien nach vorher definierten kontroversen, neutralen oder nicht kontroversen Worten analysiert [MZDC14] und anschließend dafür eine Visualisierung angeboten. Mit einer Erweiterung der Suchmaske durch eine Kombination mehrerer Wörter, die für ein bestimmtes Thema aussagekräftig sind und einer Kombinierung der daraus resultierenden Daten könnte eine bessere Unterscheidung zwischen den Nachrichtenmedien entstehen. Außerdem könnten Kombinationen von Worten oder Gruppen, die der Nutzer per Klick zusammenfasst, zur Bündelung von Informationen genutzt werden und somit das Ergebnis der Darstellung verbessern. Als Einstiegsseite in den Prototypen könnte eine Übersichtsseite erstellt werden, auf der die verschiedenen Themen des Korpus per Visualisierung angezeigt werden und durch Klick auf eines der Themen die Analyse von Wortverwendungen durchgeführt werden kann. Die verschiedenen Darstellungen im Prototyp sind ebenfalls ein möglicher Punkt der Änderungen unterzogen werden könnte, um besser erkennbar Unterschiede anzuzeigen.

Um die Nützlichkeit des Prototyps für die Erkennung von verschiedenen Frames nachzuweisen, können Nutzerstudien durchgeführt werden, in denen untersucht wird wie gut Erkenntnisse zu Framings aus dem Prototyp gezogen werden können. Darüber hinaus könnten andere Visualisierungsansätze mit denen des Prototyps verglichen werden, um herauszufinden, wie gut die Visualisierungen des Prototyps im Vergleich zu anderen Ansätzen für das finden von Framing funktioniert.

Im Allgemeinen sind weitere Forschungen im Gebiet der visuellen Darstellung von Wortverwendungen sehr sinnvoll, da es nach wie vor sehr viele Möglichkeiten der Umsetzung gibt. Die Darstellung der Visualisierung, verwendete Algorithmen, wie auch verwendete Werkzeuge sind in einer Vielzahl an Möglichkeiten unterschiedlich umsetzbar und somit viele weitere Ideen und Umsetzungen möglich.

Literaturverzeichnis

[BOH11] M. Bostock, V. Ogievetsky, J. Heer. „D^3 data-driven documents“. In: *IEEE transactions on visualization and computer graphics* 17.12 (2011), S. 2301–2309. URL: https://ieeexplore.ieee.org/abstract/document/6064996/ (zitiert auf S. 20).

[BOV+14] B. Burscher, D. Odijk, R. Vliegenthart, M. de Rijke, C. H. de Vreese. „Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis“. In: *Communication Methods and Measures* 8.3 (2014), S. 190–206. DOI: 10.1080/19312458.2014.937527. eprint: https://doi.org/10.1080/19312458.2014.937527. URL: https://doi.org/10.1080/19312458.2014.937527 (zitiert auf S. 14, 20).

[CD07] D. Chong, J. N. Druckman. „Framing Theory“. In: *Annual Review of Political Science* 10.1 (2007), S. 103–126. DOI: 10.1146/annurev.polisci.10.072805.103054. eprint: https://doi.org/10.1146/annurev.polisci.10.072805.103054. URL: https://doi.org/10.1146/annurev.polisci.10.072805.103054 (zitiert auf S. 11).

[DES+15] N. Diakopoulos, D. Elgesem, A. Salway, A. Zhang, K. Hofland. „Compare clouds: Visualizing text corpora to compare media frames“. In: *Proceedings of IUI Workshop on Visual Text Analytics*. 2015, S. 193–202. URL: http://vialab.science.uoit.ca/textvis2015/papers/Diakopoulos-textvis2015.pdf (zitiert auf S. 16, 17).

[DZS13] N. Diakopoulos, A. X. Zhang, A. Salway. „Visual analytics of media frames in online news and blogs“. In: *Proc. IEEE InfoVis Workshop on Text Visualization*. 2013. URL: http://vialab.science.uoit.ca/textvis2013/papers/Diakopoulos-TextVis2013.pdf (zitiert auf S. 15, 16).

[EMP09] R. M. Entman, J. Matthes, L. Pellicano. „Nature, sources, and effects of news framing“. In: *The handbook of journalism studies*. Routledge, 2009, S. 195–210 (zitiert auf S. 13).

[Ent07] R. M. Entman. „Framing bias: Media in the distribution of power“. In: *Journal of communication* 57.1 (2007), S. 163–173. URL: https://academic.oup.com/joc/article-abstract/57/1/163/4102665 (zitiert auf S. 9).

[Mat09] J. Matthes. „What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990-2005“. In: *Journalism & Mass Communication Quarterly* 86.2 (2009), S. 349–367 (zitiert auf S. 13).

[MWY+18] F. Morstatter, L. Wu, U. Yavanoglu, S. R. Corman, H. Liu. „Identifying Framing Bias in Online News“. In: *Trans. Soc. Comput.* 1.2 (Juni 2018). ISSN: 2469-7818. DOI: 10.1145/3204948. URL: https://doi.org/10.1145/3204948 (zitiert auf S. 11–13).

[MZDC14] Y. Mejova, A. X. Zhang, N. Diakopoulos, C. Castillo. „Controversy and sentiment in online news“. In: *arXiv preprint arXiv:1409.8152* (2014). URL: https://arxiv.org/abs/1409.8152 (zitiert auf S. 44).

[Pal] J. Palmer. *D3-tip*. URL: http://labratrevenge.com/d3-tip/ (zitiert auf S. 20).

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. „Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830. URL: https://scikit-learn.org/stable/ (zitiert auf S. 20).

[ŘS10] R. Řehůřek, P. Sojka. „Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, Mai 2010, S. 45–50. URL: https://radimrehurek.com/gensim/ (zitiert auf S. 20, 21).

[SGG10] A. C. Saguy, K. Gruys, S. Gong. „Social problem construction and national context: news reporting onöverweightändöbesityïn the United States and France". In: *Social problems* 57.4 (2010), S. 586–610. URL: https://academic.oup.com/socpro/article-abstract/57/4/586/1667272.

[Skl] Sklearn. *Sklearn Tf-Idf*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html#sklearn.feature_extraction.text.TfidfTransformer (zitiert auf S. 20, 21).

[Tex] TextBlob. *TextBlob Sentiment Analysis*. URL: https://textblob.readthedocs.io/en/dev/# (zitiert auf S. 20).

[VD95] G. Van Rossum, F. L. Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995 (zitiert auf S. 20).

Alle URLs wurden zuletzt am 12. 03. 2020 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift