Tonlage von Überschriften, Unterüberschriften und Beiträgen in Zeitungsartikeln - ein Vergleich von Artikeln der Jungen Freiheit und der Tagesschau

Lea Wetzke
Department Linguistik
BSc Computerlinguistik
Automatische Textanalyse in der Politikwissenschaft WiSe 21/22
lwetzke@uni-potsdam.de

Abstract

Headlines are not only used to summarize the contents of a news article, but also to garner interest and incite emotional reactions - they can be polarizing. In the following project, an expandable German corpus consisting of headlines, subheadlines and paragraphs of 100 news articles published by Junge Freiheit and Tagesschau, is scored based on how polarizing the three article parts are and subsequently compared, followed by a comparison of how different the scores are between both publishers. The scores are calculated with a self designed formula that is mainly based on SentiWS occurrences and punctuation usage. Results show that while Junge Freiheit headlines have the highest polarity score when subheadlines and paragraph are combined, but overall Tagesschau subheadlines have the highest score when all 3 categories are viewed independently.

1 Einleitung

Die Überschriften von Zeitungsartikeln haben verschiedene Aufgaben: sie sollen zum einen Leser:innen eine kurze Zusammenfassung der Inhalte des Artikels geben, jedoch auch oft Leser:innen zur Interaktion motivieren, also entweder zum Kauf der physischen Ausgabe oder zum Klick auf den Link (clickbait). Dies kann durch bestimmte Tonlagen erzeugt werden. Doch wie groß ist der Unterschied der Tonlage zwischen Überschrift, Unterüberschrift und eigentlichem Beitrag eines Zeitungsartikels? Wie unterscheidet sich dies bei anderen Zeitschriften? Sind die Werte gleich oder verschieden? Im Rahmen dieser Fragen soll folgende Hypothese bewiesen oder falsifiziert werden: Eine Zeitschrift auf dem rechten Spektrum ("Junge Freiheit") weist eine stärkere Diskrepanz zwischen den Kategorien Überschrift, Unterüberschrift und Beitrag in Bezug auf Tonlage auf, als eine Zeitschrift, die eher neutral auf dem politischen Spektrum orientiert ist ("Tagesschau"). Hierfür wurde ein erweiterbares Korpus von jeweils 50 Junge Freiheit- und 50 Tagesschau-Artikeln mithilfe eines in Python programmierten Webscrapers erstellt. Zusätzlich wurde in Form eines R-Scripts ein Maß entworfen, dessen Ziel es ist, die Polarität von Überschriften, Unterüberschriften und Beitrag zu messen. Dieses Maß bedient sich hauptsächlich an der Häufigkeit von Wörtern, welche im SentiWS-Lexikon auftreten, und an der Häufigkeit von bestimmten Satzzeichen.

2 Korpus

Zum Zweck dieses Projektes wurde ein Grundkorpus bestehend aus insgesamt 100 Überschriften, Unterüberschriften, und Beiträgen erstellt, die im März und April 2022 veröffentlicht wurden. Es besteht aus insgesamt 623 Überschrift-Tokens, 1409 Unterüberschrift-Tokens und 63398 Beitrags-Tokens, verteilt über 100 .csv-Dateien. Das Korpus wurde automatisch mithilfe eines Webscrapers erstellt. Diese Daten sind in .csv-Form gespeichert und erweiterbar durch die Nutzung der Klassen JFWebscraper und TSWebscraper. Mithilfe des Moduls BeautifulSoup scrapen diese Klassen die relevanten Daten beim Aufrufen der Methode .scrape(): Eine URL-Liste im .txt-Format (jf-urls.txt für alle URLs der Jungen Freiheit, oder ts-urls.txt für alle URLs der Tagesschau) eingelesen, die Überschriften, Unterüberschriften und der Beitrag an sich werden extrahiert und dann in eine .csv-Datei exportiert, jeweils eine pro Artikel. Die einzelnen Kategorien sind in der .csv-Datei durch Semikolons getrennt. Mehr Informationen hierzu sind im github-repository zu finden[1].

3 Maß zur Bestimmung der Tonlage

Um die Tonlage zu bestimmen wurde ein spezifisches Maß entworfen, welches sich hauptsächlich an SentiWS orientiert, einer Sammlung von "positive and negative sentiment bearing words

[1] https://github.com/lewetzk/jf-ts-pol

weighted within the interval of [1; 1] plus their part of speech tag, and if applicable, their inflections" (Remus et al., 2011, S. 1). Mithilfe von SentiWS kann also bestimmt werden, welche tokens eine Polarität aufweisen, ob diese Polarität positiv oder negativ ist, ist für diese Untersuchung eher weniger relevant.

Das Maß ist folgend aufgebaut:

$$pol\_score = \frac{senti(c) + (excl(c) \cdot 2) + (q(c) \cdot 0.4)}{total\_tokens(c)} \quad (1)$$

Das Maß erschließt sich aus 4 Unterwerten, die sich durch die Anwendung von Funktionen auf das zu betrachtende Teilkorpus c ergeben: absolute Häufigkeit der SentiWS-Tokens (senti), absolute Häufigkeit der Ausrufezeichen (excl), absolute Häufigkeit der Fragezeichen (q) und Zahl der Gesamttokens (total_tokens). Diese Unterwerte wurden gewählt, da sie indikativ für einen Stil mit polarisierendem Ton sein können. Die absolute Häufigkeit der SentiWS-Wörter ist einer der Hauptfaktoren. Da SentiWS eine Sammlung von Wörtern mit "prior polarity [...] i.e. their polarity without any given context or discourse" (Remus et al., 2011, S. 1) ist, kann eine hohe Frequenz solcher Wörter ein Zeichen für eine polarisierende Tonlage sein. Laut Untersuchungen kommen Wörter mit hohem Sentimentscore, besonders positiv polarisierte, beispielsweise oft in Clickbait-Artikeln vor (Chakraborty et al., 2016, S.3).
Ausrufezeichen werden stärker gewichtet: sie kommmen seltener in Überschriften vor, wenn sie jedoch vorkommen, haben sie einen starken Effekt auf die Tonlage: "[s]ie haben oft eine kommentierende Funktion und verleihen der zugehörigen Aussage einen besonderen Nachdruck" (Gehr, 2016). Außerdem würden sie oft im Boulevardjournalismus und in Werbungen auftreten (Gehr, 2016), beides Medien, welche sich oft vor einem reißerischen Ton nicht scheuen.
Fragezeichen hingegen sind häufig im neutralen Journalismus fundiert, beispielsweise bei Unwissenheit in Berichterstattungen, treten aber auch selten als Mittel der bewussten Streuung von Falschinformationen oder -interpretationen auf (Gehr, 2016). Sie werden daraus folgend weniger gewichtet.
Letztendlich wird die Summe dieser teilweise modifizierten Unterwerte durch die Gesamtzahl der Tokens des Teilkorpus geteilt, um eine Art relative Häufigkeit zu erzeugen - der Wert, welcher sich zwischen 0 und 1 bewegen kann, ist somit ein Zeichen für die Tonlage eine Überschrift, einer Unterüberschrift oder eines Beitrages. Ist der Wert nah an 0, so ist das Teilkorpus extrem wenig polarisierend: es gibt wenige bis keine polarisierenden Wörter, wenige bis gar keine Ausrufezeichen oder wenige Fragezeichen. Ist der Wert nah an 1, kann man schließen, dass das Teilkorpus stark polarisierend ist. Es gibt wahrscheinlich eine hohe Zahl negativ oder positiv polarisierten Wörtern und eine hohe Anzahl and Ausrufe- oder Fragezeichen. Bei einem direkten Vergleich von zwei pol_score-Werten kann also leicht bestimmt werden, welches Teilkorpus polarisierender ist.

4 Kurze R-Code Zusammenfassung

Beim Ausführen von polscore.R werden die bereits vom Scraper erstellten .csv-Dateien, welche in 3 Spalten, je nach Artikelteil, geteilt sind, zuerst eingelesen und dann mithilfe von quanteda in Korpusform übersetzt. Zudem wird eine .Rdata-Form des SentiWS-Corpus eingelesen, welche von Cornelius Puschmann erstellt wurde.[2]Die Funktion make_sentiws_weights() wird darauffolgend genutzt, um die *senti*-Variable zu bestimmen, indem die Anzahl an positiv sowie negativ polarisierten tokens, die im SentiWS-Wortschatz vorkommen, summiert wird. Die Funktion search_punct() ermittlet mit die absolute Frequenz der Ausrufe- und Fragezeichen (und kann dadurch auch ggf. um andere Satzzeichen erweitert werden) in den Teilkorpora. Letztendlich wird die Polaritätsformel aus Punkt 3 angewendet und alle Ergebnisse für potentielle Weiterverarbeitung werden in results.csv gespeichert.

5 Ergebnisse

Figure 1 und 2 zeigen Ergebnisse der Score-Verteiling über die verschiedenen Teilkorpora

[2]https://github.com/cbpuschmann/inhaltsanalyse-mit-r.de.

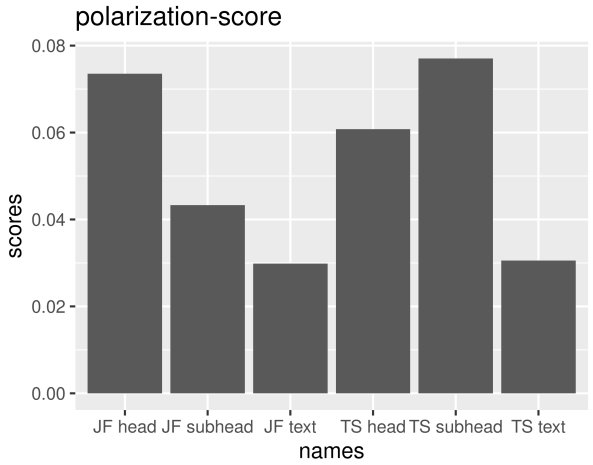

Figure 1: Ergebnisse der Maßanwendung

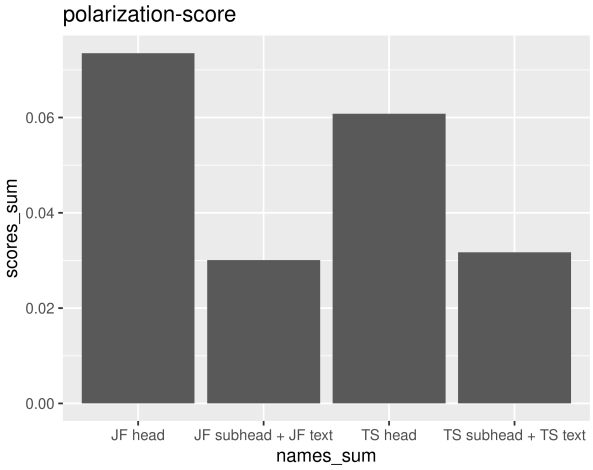

Figure 2: Ergebnisse mit Unterüberschrift- und Textwerten zusammengefasst

hinweg.

In Figure 1 sind die Ergebnisse des pol_scores angewendet auf alle Teilkorpara zu sehen. Wie man erkennen kann handelt es sich bei den ersten drei Balken um die Werte der Jungen Freiheit-Korpora, bei den restlichen drei um die der Tagesschau.

5.1 Score-Ergebnisvergleich der Junge Freiheit-Teilkorpora

Wie bereits etabliert befindet sich auf der y-Achse der Score-Wert, welcher zwischen 0 und 1 liegen kann. Angewendet auf das erste Teilkorpus, die Überschriften der Junge Freiheit (*JF head*), liegt der Wert bei ca. 0,0735, relativ hoch. Der Wert der Unterüberschriften (*JF subhead*) ist im Vergleich deutlich niedriger mit 0,0433. Noch niedriger ist der Wert des Beitrages selbst (*JF text*) mit ungefähr 0,0298. Man kann also erkennen, dass die Junge Freiheit polarisierendere Überschriften als Unterüberschriften hat, der Score ist fast doppelt so hoch. Besonders groß ist der Unterschied zwischen Überschriften und Beitragstext.

5.2 Score-Ergebnisvergleich der Tagesschau-Teilkorpora

Mit ca. 0,061 sind Tagesschau-Überschriften (*TS head*) weniger polarisierend als die Unterüberschriften (*TS subhead*) von Tagesschau-Artikeln mit einem Wert von 0,077. Beide liegen deutlich über dem Wert des Beitragtextes mit 0,031 (*TS text*). Es ist also zu erkennen, dass die Überschriften von Tagesschau-Artikeln weniger polarisierend als die Unterüberschriften sind. Der Beitragstext ist am wenigsten polarisierend, der Wert beträgt weniger als die Hälfte des Wertes der Unterüberschriften.

5.3 Vergleich zwischen den Zeitungen

Doch wie immens ist der Unterschied zwischen den beiden Zeitungen? Beide haben fast identische Werte in Bezug auf den Textbeitrag, mit einem aufgerundetem Wert von ca. 0,03. Die Überschriften-Polarität ist nicht extrem unterschiedlich, aber auch nicht sehr ähnlich mit einem Score-Unterschied von ca. 0,013. Der größte Unterschied liegt bei den Unterüberschriften. Der Wert ist der Tagesschau doppelt so hoch wie bei der Jungen Freiheit.

5.4 Score-Erebnis bei Kombination von Unterüberschriften und Beitragstext

Beitragstext und Unterüberschriften fügen sich zu einer Einheit zusammen: die Überschrift bezieht sich grundsätzlich auf die Summe beider. In Figure 2 ist ein Graph zu sehen, in dem das Unterüberschrift- und das Textteilkorpus zusammengefasst wurden. Anhand des Graphen kann man den Schluss ziehen, dass die Junge Freiheit durchschnittlich polarisierendere Überschriften im Vergleich zum Restartikel hat, fast dreifach so hoch wie der zusammengefasste Wert aus Unterüberschriften und Text (*JF subhead + JF text*). Der Score ist bei den Tagesschau-Überschriften etwas niedriger als der Junge Freiheit-Überschriften-Score, das kombinierte Korpus (*TS subhead + TS text*) hat einen Score, der halb so hoch ist wie der der Tagesschau-Überschriften. Kurzum: Überschriften der Jungen Freiheit sind generell polarisierender als die der Tagesschau und weisen eine größeren Unterschied zum Score des Beitragstextes selber auf.

5.5 Russischer Überfall auf Ukraine als möglicher Störfaktor

Der seit dem 24. Februar 2022 laufende Überfall der russischen Truppen auf die Ukraine dominiert die Nachrichten - und könnte somit einen Einfluss auf die Ergebnisse des Projektes haben. Das SentiWS-Lexikon besitzt eine Vielzahl von Wörtern, welche relevant für Kriegsreportage sind, wie *Krieg*, *Angriff*, *Invasion*, und vielen weiteren. Da das Beispielkorpus ausschließlich aus Artikeln von März und April 2022 besteht, kommen diese kriegsrelevanten Wörter häufiger vor als zu Zeiten des Friedens. Somit könnte die Verteilung mit einem Korpus aus einem anderen Zeitraum recht anders ausfallen - dieses Korpus ist aufgrund dieser Faktoren wahrscheinlich besonders polarisierend.

6 Zusammenfassung

Mit diesem Projekt sollte ergründet werden, dass Eine Zeitschrift auf dem rechten Spektrum ("Junge Freiheit") eine höhere Diskrepanz zwischen den Kategorien Überschrift, Unterüberschrift und Beitrag in Bezug auf Tonlage aufweist, als eine Zeitschrift, die eher neutral auf dem politischen Spektrum orientiert ist ("Tagesschau"). Mittels eines Python-Webscraper wurde ein Korpus bestehend aus 100 Zeitungsartikeln, unterteilt in Überschrift, Unterüberschriften und Beitragstext erstellt. Anhand eines in R implementierten eigens entwickelten Maßes, welches sich hauptsächlich an dem SentiWS-Lexikon und der Nutzung von verschiedenen Satzzeichen orientiert, konnte gezeigt werden, dass die Überschriften der Jungen Freiheit einen deutlich polarisierenden Tonfall als die zugehörigen Unterüberschriften oder den zugehörigen Beitragstext haben - was noch deutlicher zu erkennen ist, wenn die Teilkorpora der Unterüberschriften und des Beitragstextes zusammengefasst werden: es ist eine deutliche Diskrepanz feststellbar. Bei der Tagesschau sieht es etwas anders aus, hier sind die Unterüberschriften am polarisierendsten, was den Tonfall angeht, jedoch geht dieser Effekt bei dem zusammengefassten Korpus, bestehend aus Unterüberschriften und Beitragstext der Tagesschau-Artikel, unter. Die Diskrepanz zwischen Polaritäts-Score des Überschriftenkorpus und des zusammengefassten Korpus ist niedriger. In diesem Fall ist die Hypothese korrekt: die Zeitschrift auf dem rechten Spektrum zeigt in diesem Vergleich eine höhere Diskrepanz auf, und hat auch allgemein einen höheren Polaritäts-Score bei den Überschriften auf als bei der Tagesschau. Wenn jedoch alle Kategorien seperat untersucht werden, kann man erkennen, dass die Unterüberschriften der Tagesschau deutlich polarisierender sind als die Überschriften und die Beitragstexte dieses Nachrichtenportals, sogar noch polarisierender als die Überschriften der Jungen Freiheit. Aus diesem Blickwinkel ist die Hypothese entkräftet.
Weiterführend könnte das Projekt relativ einfach ausgebaut werden. Man könnte beispielsweise das Korpus mit neuen Links von beiden Portalen erweitern, indem der Webscraper benutzt wird. Zudem könnte man mit etwas mehr Aufwand mehr weitere Zeitungen zum Vergleich heranziehen oder die Formel so verändern, dass nur SentiWS-Tokens anerkannt werden, die über einem gewissen SentiWS-Wert liegen.

7 Quellenangabe

Chakraborty, Abhijnan Paranjape, Bhargavi Kakarla, Sourya Ganguly, Niloy. (2016). Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media.

Gehr, M. (2016, 16. Oktober). Überschrift. Journalistikon. Abgerufen am 5. April 2022, von https://journalistikon.de/ueberschrift/

Puschmann, C. (2021). Inhaltsanalyse mit R. Inhaltsanalyse mit R. Abgerufen am 5. April 2022, von http://inhaltsanalyse-mit-r.de/einleitung.html

R. Remus, U. Quasthoff G. Heyer. SentiWS - a Publicly Available German-language Resource for Sentiment Analysis.
In: Proceedings of the 7th International Language Resources and Evaluation (LREC'10), pp. 1168-1171, 2010