Universidad Politécnica
de Madrid

Escuela Técnica Superior de

Ingenieros Informáticos

M. Sc. Data Science

Intelligent Systems:

NLP Practical Work

First Semester 2021/2022

Author:

Marcel Schweiker

Table of Contents

List of Figures

Project Overview

Background

Out of personal interest I follow the recent debate in the US-American society regarding the voting rights legislations proposed by the democrats [1]. These bills revolve mainly around national standards for early voting and voting by mail, in addition to abolishing partisan gerrymandering as well as the restoration of the Voting Rights Act that were struck down by the Supreme Court during the past decade. This reformation is a polarising, red-hot topic in the news right now - especially after President Bidens speech in Georgia in which he called on Senate to change filibuster rules to pass the voting right bills.

A filibuster is a political procedure in which Senate members prolong the debate on proposed legislation in order to delay or entirely prevent decision.

Material

As an example of the divergence within society regarding this topic, two news articles have been identified assessing the voting rights bills and Bidens Georgia speech entirely adversely.

The first source is an opinion piece from the *New York Post* which comes down hard on Biden [2][1]. On the one hand, the author criticizes the voting right bills as unnecessary for the majority of citizens. On the other hand, the author attacks Biden personally – portraying the announced suspension of the filibuster rules as a lie and denouncing Bidens rhetorical style as disgraceful and shameless.

An opposing opinion comes from a *The Hill* article by Donna Brazil [3].[2] She praises the efforts taken by the Biden administration to abrogate the filibuster rule in order to reform the voting rights as she assesses these as necessary actions to protect the nations democracy.

Objectives

The project which shall be conducted in R using RStudio [4] has several goals:

First, the data (both news articles) shall be scraped from the web followed by some NLP processing steps and checks in order to obtain cleaned data sets of the articles' natural language.

This data then serves as basis for a sentiment analysis. The objective is to try out different lexica, packages and methodologies to extract insights on the sentiments of both articles. Furthermore, main outcomes shall be visualized.

Based on the results of this sentiment analysis, interpretations shall be conducted e.g. in order to derive possible intentions, underlying attitude and biases of the publishers and authors as well as potential reasons for the fierce political polarization in US-American media.

---

[1] https://nypost.com/2022/01/11/bidens-disgraceful-lies-on-filibuster-dems-power-grab-over-us-voting-laws/

[2] https://thehill.com/opinion/campaign/589286-bidens-georgia-speech-was-a-call-to-save-democracy-as-we-know-it

Project Execution

Web scraping

In order to load the data into R, the articles have been scraped from the web. In order to identify the specific CSS tags, the Google Chrome Extension SelectorGadget has been used [5]. This returned a vector with the paragraphs of the articles.

```
##Webscraping the two relevant news articles
#create new variable for the link and get html document of this webpage
nypostlink = "https://nypost.com/2022/01/11/bidens-disgraceful-lies-on-filibuster-dems-power-grab-over-us-voting-laws/"
nypostpage = read_html(nypostlink)

#Scrape the article from the webpage using pipe operators with CSS tag from SelectorGadget
nypostarticle = nypostpage %>% html_nodes("p") %>% html_text()
```

Processing and basic checks

The web scraping returned more than just the relevant parts of the article, since the corresponding CSS tag also included Metadata. These paragraphs were consequently removed.

Furthermore, the text was checked for UTF-8 encoding and character normalization. Doubled or more spaces were replaced with a single space. Then, spacy tokenizer [6] was used to obtain single phrases from the article paragraphs whereupon empty sentences were removed. These processing steps were conducted based on the Hands-On 2 [7].

The sentence analysis resulted in a count of 20 non-empty sentences for the *New York Post* article and 31 for *The Hill* article.

Sentiment Analysis

Overall sentiment comparison by using NRC lexicon of syuzet package

Thus, these sentences have been assigned sentiment scores according to the NRC lexicon using the syuzet package [8]. Thereupon the scores were plotted (see Figure 1).

```
#obtaining sentiment scores for NYPost article and plotting them
sentimentscoresnypost <- get_nrc_sentiment(v_nypostarticlephrases)

cbind(v_nypostarticlephrases, sentimentscoresnypost)

barplot(sort(colSums(sentimentscoresnypost), decreasing = TRUE), las=2, col = rainbow(10), ylab = 'count', main ='Sentiment Scores for NYPost article')
```

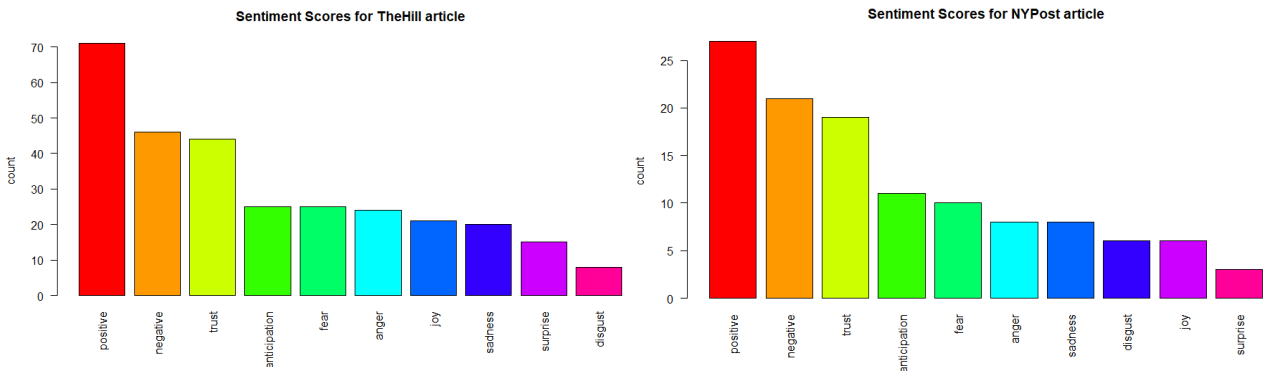

Figure 1: NRC Sentiment Scores

Moreover, a two-keyed bar plot has been created to compare both articles in one plot using ggplot2 [9]. For even better comparability, another depiction with relative sentiment score has been plotted since the number of sentiment scores differed greatly due to the varying length of text (see Figure 2).

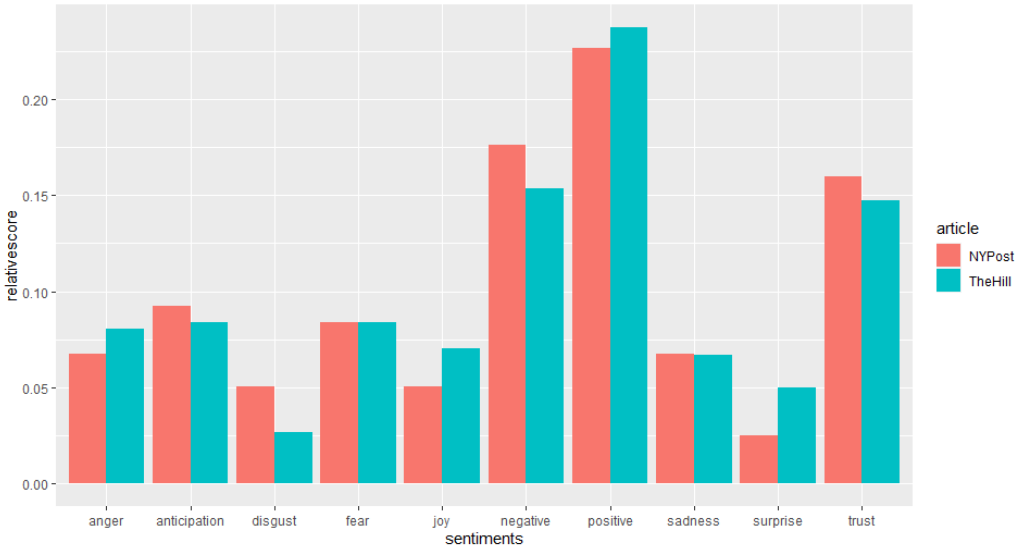

Figure 2: NRC Sentiments Comparison

Top words by sentiments using bing and loughran lexica

With the tidytext package [10], two other lexica have been used to retrieve sentiments – bing and loughran. The bing lexicon was used to retrieve words identified as positive respectively negative. Then, the frequency of these words was counted in order to display all of them in a word cloud (see Figure 3 – from left to right: *TheHill* negative, *TheHill* positive, *NYPost* negative, *NYPost* positive):

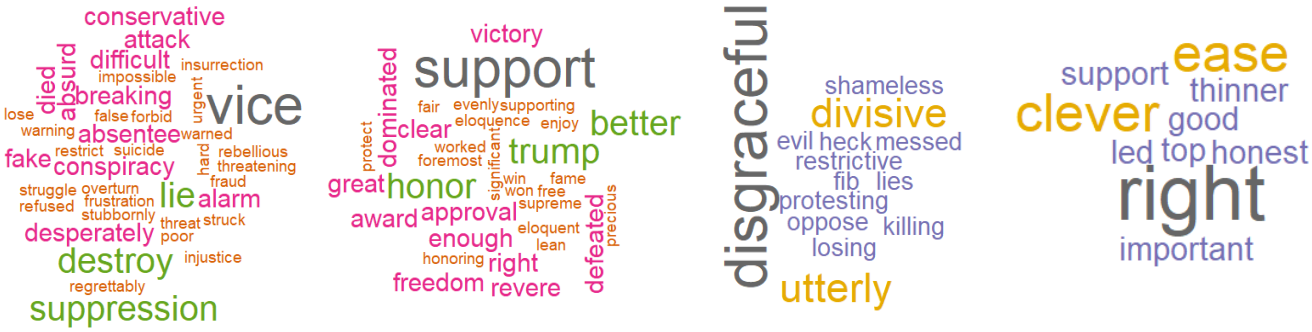

Figure 3: Bing Sentiments Wordclouds

After that, only the Top-10 words contributing to each bing sentiment by frequency were plotted in a bar chart.

The loughran lexicon classifies sentiments into 5 different categories: constraining, litigious, negative, positive and uncertainty. Again, for both articles, the Top-10 words contributing to each of those sentiment by frequency were plotted in multiple bar charts (see Figure 4):

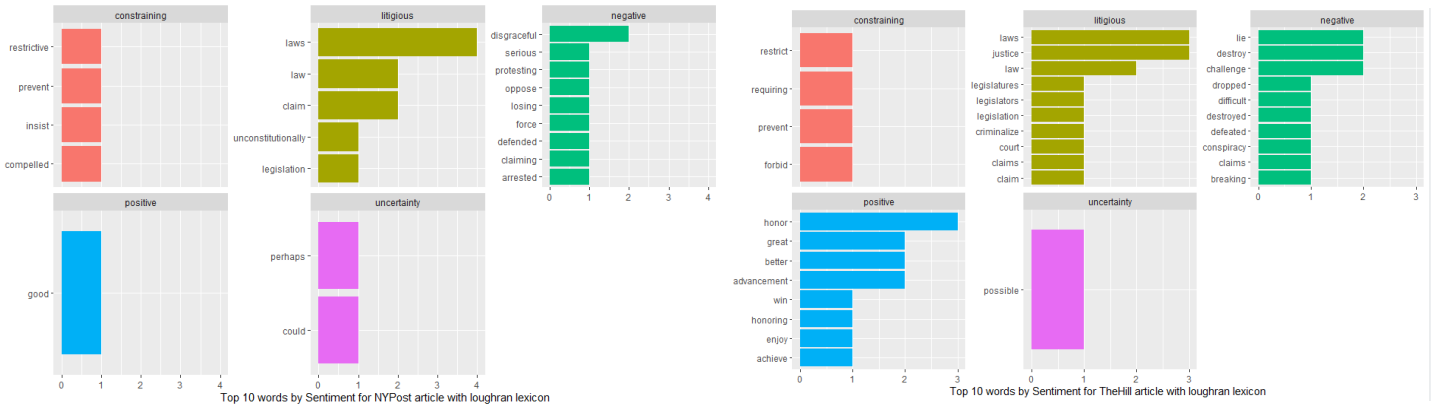

Figure 4: Laughran Top-10 Words by Sentiment

Project Evaluation

Conclusion and limitations

Looking back to the formulated objectives of this project, one can conclude that the first two objectives were purposefully fulfilled. Through implementing the web scraping of the articles, the R Script can run without any further input data. Furthermore, the basic checks and processing of the scraped articles ensured a clean data set for further analysis.

This sentiment analysis was then performed intensively, using three different lexica – nrc, bing and loughran – with different methods of extracting insights therefrom. On the one hand, a complete comparison of nrc sentiments between the articles was conducted, while on the other hand bing sentiments were used to plot positive and negative word clouds per article. Furthermore, the Top-10 words per bing and loughran sentiments were identified and plotted producing interesting insights in the articles.

However, the sentiment analysis did not allow clear interpretation in order to meet the third objective of this project sufficiently. While the presented articles are bipolar, the nrc sentiment comparison did not at all reflect that (see Figure 2). In all sentiments distinguished within the nrc lexicon, both articles score around the same relative amount without any outstanding differences. The usage of the laughran lexica resulted at least in a much greater number of identified positive sentiments in the *New York Post* article. But neither these results nor the results from the bing word clouds could be used for any interpretation approach for possible intentions, underlying attitude and biases of the publishers and authors. This is partly because positive and negative words are used in both articles, but their reference and annotation are differing. An example of these bilateral occurrences is revealed by a look into the bing word clouds.

In essence, this shows that the sentiment analysis performed is not particularly suitable for drawing conclusions, as it only remains at the word level and cannot evaluate and analyze the connections and context. The chosen methods are therefore more suitable for quantitative analyses with large data sets, but rather not useful for qualitative and contextual evaluation. In future projects, this might be improved by using coreferences or BERT models.

Github Link

https://github.com/marceey/NLPwithR

References

[1] Astead W. Herndon (2022): A Last-Gasp Push on Voting Rights; published by New York Times; https://www.nytimes.com/2022/01/19/podcasts/the-daily/voting-rights-senate-biden.html, last access: 19/01/2022.

[2] New York Post (2022): Joe Biden's disgraceful lies on the filibuster and Dems' power grab over US voting laws; https://nypost.com/2022/01/11/bidens-disgraceful-lies-on-filibuster-dems-power-grab-over-us-voting-laws/, last access: 29/01/2022.

[3] Donna Brazil (2022): Biden's Georgia speech was a call to save democracy as we know it, published by The Hill, https://thehill.com/opinion/campaign/589286-bidens-georgia-speech-was-a-call-to-save-democracy-as-we-know-it, last access: 29/01/2022.

[4] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

[5] Selector Gadget 1.1.1. (last update May 2020); https://selectorgadget.com/.

[6] Benoit, Kenneth, Akitaka Matsuo, and Maintainer Kenneth Benoit (2018): Package 'spacyr'; https://CRAN.R-project.org/package=spacyr.

[7] Mariano Rico (2021): Processing Don Quixote - NLP master course 2021-2022.

[8] Matthew Jockers (2017): Package 'syuzhet', https://cran. r-project. org/web/packages/syuzhet.

[9] Hadley Wickham, Winston Chang et al. (2016): Package 'ggplot2', https://cran.r-project.org/web/packages/ggplot2/index.html.

[10] David Robinson, Julia Silge et al. (2016): Package 'tidytext', https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html.