Sentiment and Subjectivity Analysis of Different News Sources

Trevor Moreci

2022-11-29

Introduction

The political leanings of different news sources is a highly talked about subject matter in today's day and age. These sources serve the purpose of updating they're readers on current events, and many claim to do so in a purely objective manner. Despite this claim, it is widely known that many of these sources have political leanings. For example, CNN and the New York Times are widely regarded as having a heavy liberal bias, while Fox News and the National Review are known for having a heavy conservative bias. Current American politics is very divided, and one of the main reasons for this is that people only consume media that aligns with and feeds into their political beliefs. These reasons were the inspiration for the research questions in this paper, which specifically look at the differences in language and sentiment regarding different political topics across different news sources.

In this experiment, I am using a dataset of around 150,000 news articles from 15 different sources and looking at how the language used and sentiment expressed in the articles differs between sources on different topics. Specifically, I looked at articles that had "Trump" in the title because of his polarizing nature. I believe that this is an interesting topic to explore, because it has been well documented that news sources often have biases either for or against specific political candidates or parties. I conducted this experiment by conducting a multidimensional analysis, or MDA, on the data. Following the results of this analysis I looked at the difference in mean sentiment and subjectivity scores for articles on the topics of Hillary Clinton and Donald Trump, specifically around election time. Elections are a very divisive time for a nation's politics, so looking at how different news sources were talking about the two main candidates will be beneficial for the analysis. Finally, I looked at the difference in sentiment on an subject that was political in a different manner, Colin Kaepernick. Kaepernick was a quarterback for the San Francisco 49ers in 2016, when he began to kneel during the national anthem to protest police brutality. This sparked a wide-ranging conversation where many people thought he was brave for speaking out and protesting in the manner that he did, while others thought he was being disrespectful towards the United States and its national anthem. Due to this topic's political nature and controversy, it was deemed a good topic to examine in this experiment.

Data and Methods

Data

The initial data that I gathered was found on Kaggle at the following link: https://www.kaggle.com/datasets/snapcrack/all-the-news. This dataset is composed of 143,000 news articles from 15 different sources between the years of 2016 and 2017. Some of the main sources include CNN, Fox News, The New York Times, The New York Post, Reuters, and Breitbart. A breakdown of how many articles were included from each source, along with their hypothesized or widely believed political leaning, is in Table 1. Since there was so many initial articles, I subset the data in various ways for the different analyses that I conducted. The details of each data transformation will be included in the methods section for each.

Table 1: Number of Articles by Source

Publication	Count	Political_Leaning
Breitbart	23781	Right
New York Post	17493	Right
NPR	11992	Left
CNN	11488	Left
Washington Post	11114	Left
Reuters	10709	Neutral
Guardian	8681	Left
New York Times	7803	Left
Atlantic	7179	Left
Business Insider	6757	Neutral
National Review	6203	Right
Talking Points Memo	5213	Left
Vox	4947	Left
Buzzfeed News	4854	Neutral
Fox News	4354	Right

MDA

In order to conduct MDA, I first decided to filter out the data set and include only articles that had "Trump" in the title and were written in 2017. This cut down the number of articles to 10,872. From there, I reduced the dataframe so that the only two columns were the publication and the content of the article.

After getting the data in this form, I started the analysis by pre-processing the text using the preprocess_text() function. Then I tokenized the text, removing punctuation, but keeping symbols and numbers. I decided to keep both symbols and numbers because I thought they would be valuable for determing the type of speech and analysis in each article. To conduct MDA, I decided to use the DocuScope tagger due to the ease of use and interpretability of the categories. In order to use DocuScope, I created a table with the normalized counts of each Docuscope category for each article. Then, I created a scree plot to decide how many factors would be in the MDA. Based on this scree plot, I created an MDA of the data with three factors.

Clinton vs Trump Polarity

In order to conduct sentiment analysis on I first created two subsets of the dataset. The first subset contains articles that have "Trump" in the title, but no mention of "Clinton" in the title. The second subset contains articles that have "Clinton" in the title but no mention of "Trump" in the title. I chose this method because it was the best way to filter the topics of the articles without doing full topic modelling on all of the articles. From there, I decided to look at articles from the following sources: Breitbart, Business Insider, CNN, National Review, Talking Points Memo, and New York Times. I chose these sources because most of them were identified as sources that contain more personal interaction through MDA. The National Review was the only source that wasn't on that side of the MDA that I included, but I chose to include it because it is known for having right-leaning proclivities. After subsetting the data in this way there were around 10,000 Trump articles and 5,000 Clinton articles. In order to even out the number of observations, I used a random sample of 5,000 of the Trump articles.

To actually conduct the sentiment analysis, I found the mean polarity scores for all of the articles of each source. In order to find the polarity score, I used spacytextblob, which is a spacy pipeline that enables sentiment analysis using the TextBlob library. This process uses a dictionary based approach where words are classified as either positive, negative, or neutral in a predefined dicitonary. With a score for positive being 1, and a score for negative being -1, these scores are mean-pooled from a word level to a document level. The entire sentiment analyses pipeline was conducted using Python. The mean polarity scores for each source were then compared in a bar graph.

Clinton vs Trump Subjectivity

In order to delve deeper into these findings, I also decided to look at the subjectivity scores of these same articles across the different sources. This was conducted in a very similar manner as the sentiment analysis, by using a spaCy TextBlob pipeline. In this case though, a subjectivity vs objectivity dictionary was used instead of a dictionary for polarity. In this case, an article with a score of 0 is deemed completely objective, while an article with a score of 1 is deemed completely subjective. The mean subjectivity scores for each source were then compared in a bar graph.

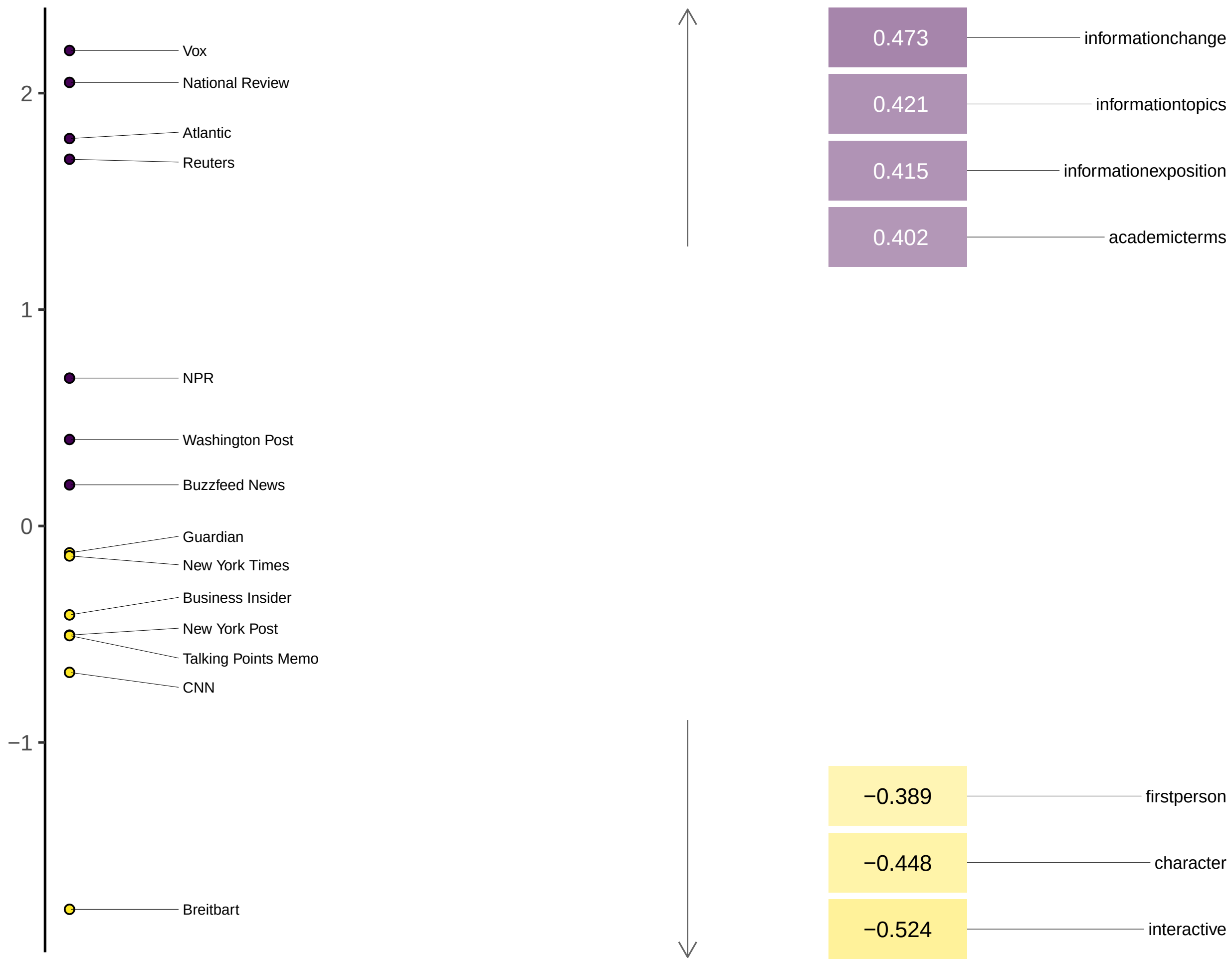
Kaepernick Articles Polarity

In order to conduct sentiment analysis on articles regarding Colin Kaepernick, the dataset was subsetted to only look at articles that included "Kaepernick" in the title. This created a dataset composed of only 172 articles. The sources that looked at in this analysis were Breitbart, CNN, Fox News, National Review, and the New York Times. These sources were chosen because all had over 5 articles on the topic, unlike many of the sources, and had some known political leanings. In order to get a better sense of the overall tone of the articles, the mean polarity of the titles of the articles was also examined. This was useful because of the nature of this controversy and the fact that the titles of the articles might reveal more about the intentions of the articles. One thing to note is that Breitbart makes up the majority of the articles in the dataset with 90, while the rest of the sources had somewhere between 5 and 15 articles.

Results

MDA

As mentioned in the methods section, I decided to use three factors for this MDA. After calculating the factor loadings and MDA scores for each factor, I ran linear regression models to test how explanatory each of the factors were. Factor 1 had an R-squared value of 0.15, factor 2 had an R-squared value of 0.10, and factor 3 had an R-squared value of 0.04. These effect sizes are all very low, especially factor 3's value, and lead to a lot of concern regarding the usefulness of MDA in this situation. With that being said, many of the p-values of the MDA scores were significant, and the dimensions created by the MDA provide insight that could be useful for further analysis. In this analysis I will only look at the second factor since it has the most relevance to the questions at hand in this report. .



Dimension 2 is displayed above in an MDA heatmap. I named this dimension "Formal Information vs. Personal Interaction". As you can see in the heat map, categories on the Personal end of the dimension include first person, interactive, and character. This means that sources that fall under this category use more first person pronouns and address the audience or other people more often. Sources on this end of the dimension include Breitbart, CNN, and the New York Post. One possible takeaway from this is that these sources may have presented more editorials or opinion pieces.

On the other hand, categories on the Formal Information end of the dimension include academic terms, information change, and information topics. This means that sources on this end of the spectrum may be more objective in their writing and might use more nuanced or specialized language when writing about the topic at hand. Sources on this end of the spectrum include Reuters, Vox, National Review, and the Atlantic.

In my opinion, the results from this factor are very informative. In general if I am looking for any biases or differences in sentiment between different sources, I will probably find that in editorials or opinion pieces. Based on the results from this factor, I can see that it would be beneficial to look into comparing the language and sentiment in sources like Breitbart, CNN, the New York Post, and Talking Points Memo. These results factored into my selection of which sources to look at when comparing the polarity of articles.

Trump vs Clinton Comparisons

Polarity

As mentioned in the methods section, the first sentiment analysis that conducted was looking at how sentiment differed between sources in regards to their articles concerning Hillary Clinton and Donald Trump.

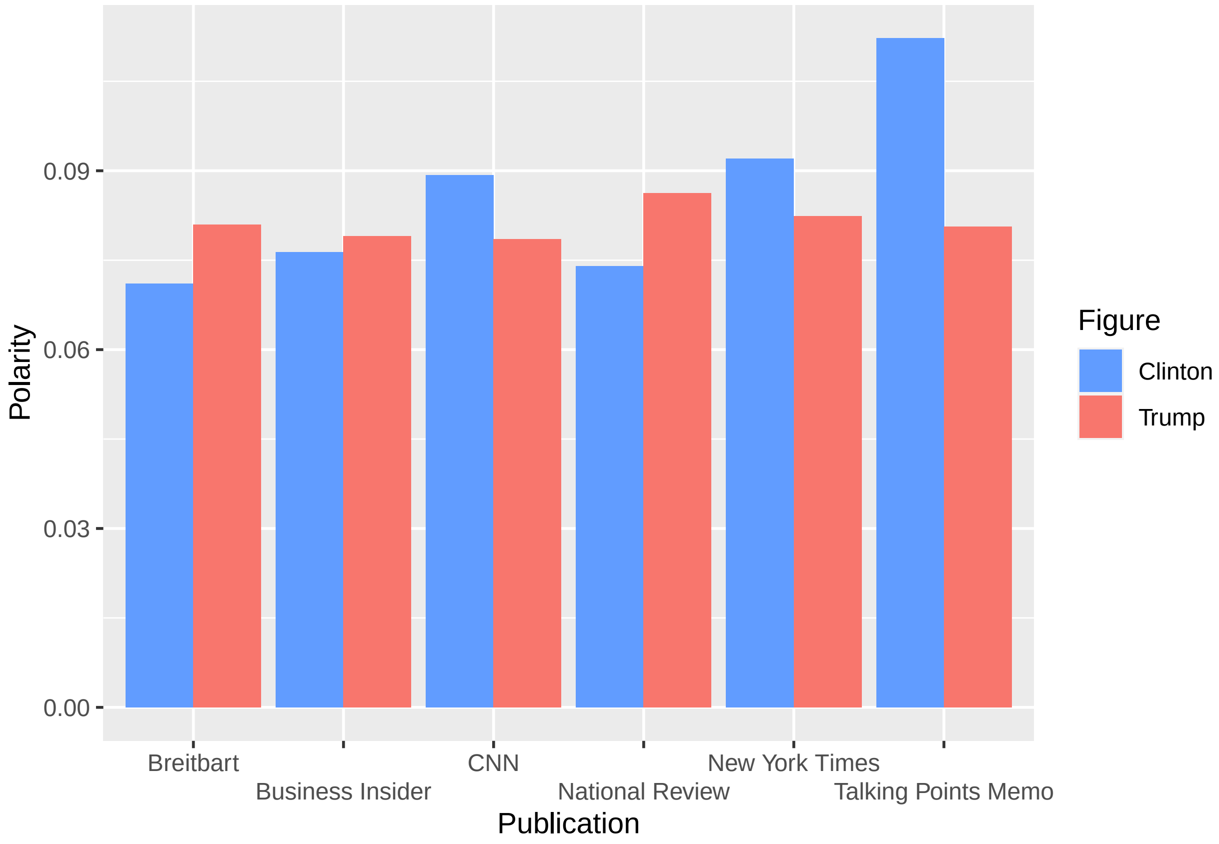


Figure 1: Polarity Scores by Source and Topic

The results from this analysis can be seen in Figure 1. While all of the polarity scores were greater than zero, or objectively more positive than negative, there are some meaningful differences in how positive the scores are. In this case, almost all of the sources that you would expect to be left-leaning in their audience and editorials have higher polarity scores for articles concerning Hillary Clinton, while all of the sources that you would expect to be right-leaning have higher polarity scores for articles concerning Donald Trump. For example, Talking Points Memo has a polarity score of 0.11 for Clinton articles, while it only has a polarity score of 0.08 for Trump articles. On the other hand, Breitbart and National Review both have over a 0.01 difference in polarities between the two, with the score for Trump articles being higher for both.

Subjectivity

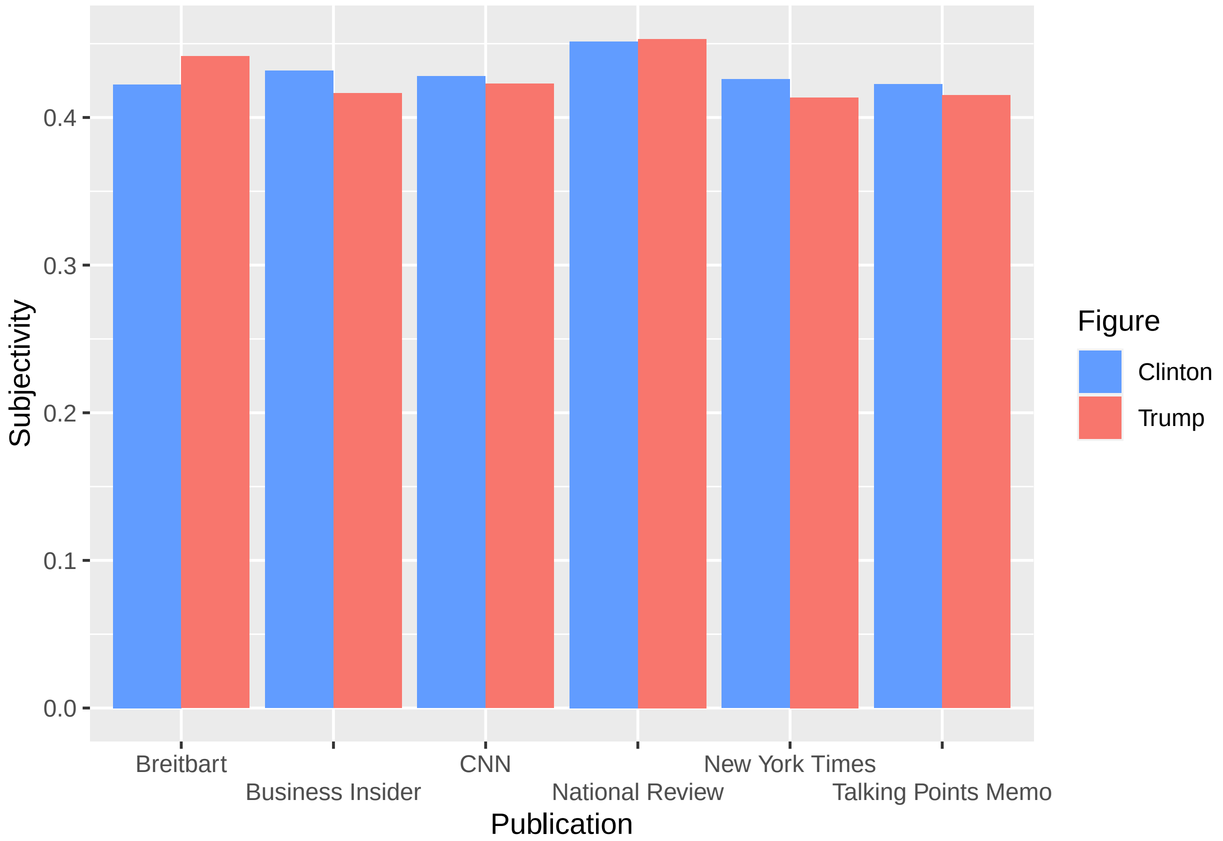


Figure 2: Subjectivity Scores by Source and Topic

Figure 2 shows the subjectivity scores for all of the sources included in the sentiment analysis. The main takeaway from this graph is that all of the sources mean subejctivity scores fall in the same range of between 0.41 and 0.45. This means that all of these source are slightly more objective in their language than subjective. One thing to note is that the sources deemed more liberal in their proclivites(Talking Points Memo, New York Times, CNN) are slightly more subjective when talking about Clinton, while the more conservative sources are slightly more subjective with articles regarding Trump.

Kaepernick Articles

Figure 3 shows the polarity scores for five different sources with articles that are about Colin Kaepernick. While all of the polarity scores are positive, much like with the results from the Clinton vs Trump analysis,

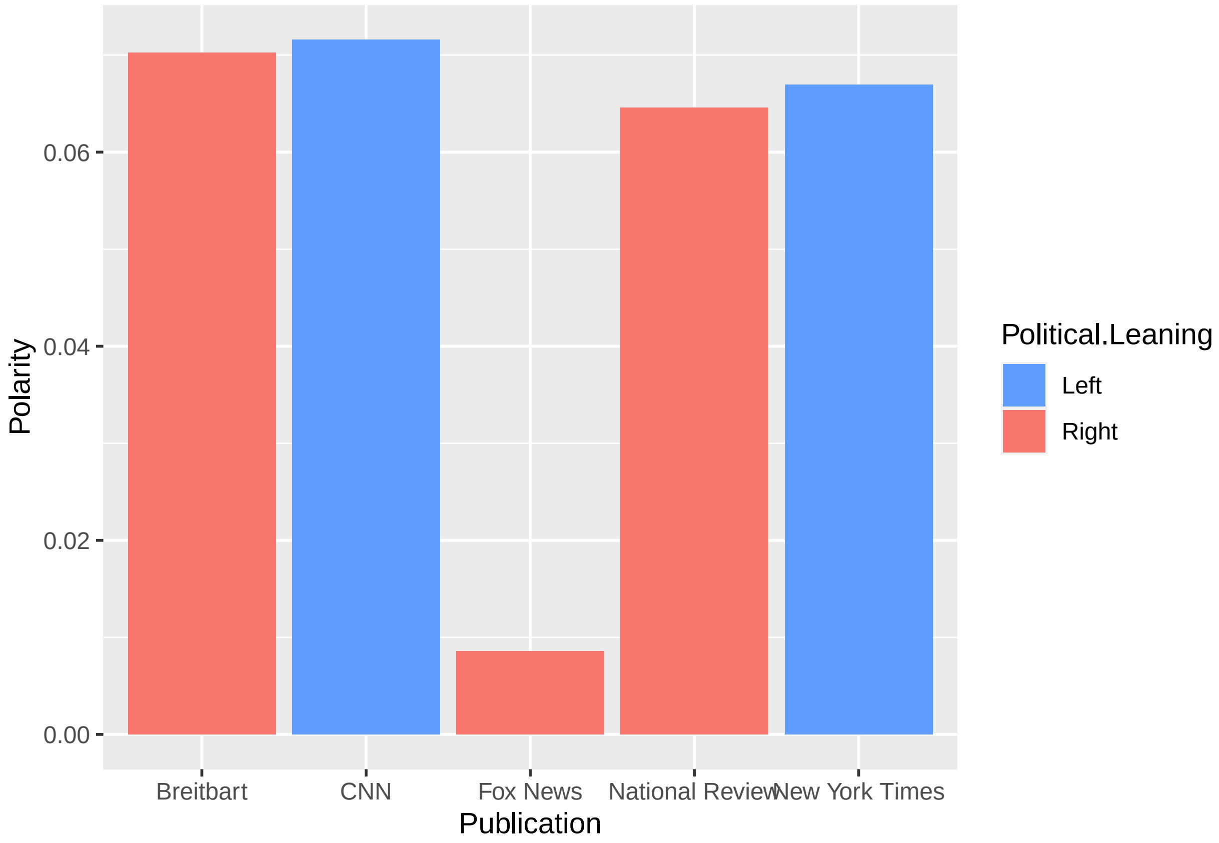


Figure 3: Polarity Scores of Kaepernick Articles by Source

there is again differences in the magnitude of the scores. The trend is very similar to the previous results as well, as CNN has the highest 0.075, while Fox News has the lowest polarity score of 0.009. This difference can really be seen in the titles of the articles from the two sources. For example, one of the CNN articles has the title "Obama defends Kaepernick's anthem protest", while one of the Fox articles has the title "NFL players blast Kaepernick's decision to sit during national anthem". When looking more closely at the mean polarity of the titles of the articles, CNN articles had a score of 0.022, while Fox News articles had a score of -0.04. Another interesting finding is that the National Review titles had a polarity score of -0.33. This really highlights the negative tone of the these articles, and makes sense with regards to the sources' political leanings. Another result that stands out is that Breitbart has the second highest polarity score, at 0.07. This is interesting since Breitbart is widely known for having a right-wing audience.

Discussion

Overall, this experiment produced some interesting results, even if they weren't as significant or interpretable as would have been expected. This trend starts with the results from the MDA. The effect size of the MDA ended up being very low, but its results shaped many decisions in the following methods. When looking at the results of the subjectivity analysis, I found it very interesting that the scores were all in the same range for the sources used.This points to a takeaway that while these sources may have varying levels of political biases, they all maintain a similar level of subjectivity in their writing.

One interesting observation from the sentiment analysis, is that all of the mean polarity scores were greater than zero, or slightly more positive than negative. One possible explanation for this is just the pure nature of the writing style of news articles. When writing on or reporting a story, an author would be more likely to use words that are more objective, or matter-of-fact. Furthermore, they might be more likely to show both sides of the story, using both positive and negative words, even if they have biases in their tones. This would lead to the polarity results of closer to zero that we saw in the results.

One of the main drawbacks of this experiment was the method that was used to find the articles that were about Clinton and Trump exclusively. While all of the articles contained the names in their title, there is no way of knowing for sure whether the article was purely about them. Many of the articles could have just included the names since they were involved, but been about completely different subject matters. Since reading through and filtering 5,000 articles one by one would have been very time consuming, using a topic modelling method, such as TF-IDF, might be recommended for future analyses.

Another drawback was the lack of data when it came to articles about Kaepernick articles. There were some very promising results, especially when it came to the difference in polarities between CNN and Fox News, but it would have been nice to have more data to really confirm these results. This is especially true concerning the sentiment analysis of the titles since there are so few words in a title.

References

Dataset: https://www.kaggle.com/datasets/snapcrack/all-the-news

spaCyTextBlob: https://spacy.io/universe/project/spacy-textblob

Source Political Leanings: https://www.allsides.com/media-bias

DocuScope: https://docuscospacy.readthedocs.io/en/latest/docuscope.html#categories