QTM 340, "Practical Approaches to Data Science with Text"
Final Project: All The News
Research Question: How did the sentiment of the major candidates of the 2016 election vary based on publication?
Jong Min Park, Angela Guevarra, Zone Li

Introduction

There has been a large, on-going debate about how the press & many other media outlets may be politically biased in their channels. Since the 2016 presidential election cycle, there has been increased discussion of partisan bias in major media. President Trump has called out "mainstream media" on numerous occasions for portraying him in a negative light, while being more lenient with his Democrat opponents. Through this project, we will see if this bias is reflected in sentiment scores of articles in these major American publications as they describe the presidential candidates in the 2016 election. First, we will describe any relevant studies our project could engage in conversation in, then describe our data & methodology used to analyze the sentiment of those articles. We will then lay out our results and conclude with possible limitations and areas of improvement.

Relevant Studies

While there are multiple relevant studies performed by many intellectuals and researchers, we want to intervene in three particular conversations made by studies that also engaged in a similar research question as ours. A study performed by Julien Phalip proposes a similar question in identifying bias in media with sentiment analysis. However, his work differs because he pulls data from Youtube uploaded by prominent American TV news channels & downloaded the metadata using Youtube API. This differs from our project in that we use content from articles to analyze, not data from videos. Julien looked at the metadata & analyzed the video descriptions and video titles by performing sentiment analysis on them using Google Natural Language API. He filtered out specific topics to further analyze specific sentiment on specific topics/political candidates. We would love to engage in conversation with his study to compare how the sentiment of political candidates from the same publications were to differ from their article content compared to their video content. Would we see the same biases in both forms of media or would they drastically differ?

Another conversation we would love to engage in is a study done by Lucas Kohorst that ranked news sites based on their subjectivity and polarity using sentiment analysis on Donald Trump. While his work is similar to our project in that we will both be performing sentiment analysis of Donald Trump from different publications, the author only used the 30 most recent articles from these 12 publications. We would love to engage in conversation with this work to compare our sentiment scores of Donald Trump from the same publications. Given that our dataset pulls articles primarily from 2016 while Lucas pulls only the most recent articles from 2018, it would be interesting to see how his analysis could add on to the trend over time graph that we created and see if the trend continues in a similar, or different, direction. However, the only drawback would be that we could only use this work for Donald Trump's trends and be unable to compare it to Hillary Clinton's as Lucas only focused on political candidate Donald Trump.

The third and last conversation we would engage in using our project is a work by Erik Bleich that poses a different research question that asks how are Muslims portrayed in the media. While our areas of interest differ, we use the same methods behind determining sentiment scores for this certain topic. This conversation would differ from the first two because rather than talking about the sentiment scores and the results of our respective analyses, it will be more focused on our methodologies and how we used it in relation to our differing research questions.

Data

The dataset is called All the News. This data set has 143,000 articles from 15 American publications. Most of the articles were dated between 2016-2017 with quite a few from 2015 and some from earlier. The majority of the articles are from Brietbart followed by the New York Post and CNN. The articles in this dataset were prominently placed on the publications' website and whole sites were not scraped. The curator used archive.org and RSS feeds to find articles to scrape. The subject matter varies but is mostly political. As seen from the publications, the political alignment is also varied.

The instances that comprise the dataset are articles from publications like New York Times, Breitbart, CNN, Business Insider, the Atlantic, Fox News, Talking Points Memo, Buzzfeed News, National Review, New York Post, The Guardian, NPR, Reuters, Vox, and the Washington Post. These publications are not representative of all publications that exist, both print and digital. They were chosen based on familiarity of the domain and tried to be inclusive of all political alignments, but still include a good mix of print and digital publications. This data collected is raw data; it includes Database ID, Article Title, Publication Name, Author Name, Date, Year, Month of Publication, and all of the article content. All articles in the dataset have all of this information except for 13% missing author names. However, since we will mainly be focusing on the content of the articles, this missing data will not impact our analysis. The text within the articles have not been cleaned. This data is not representative of all articles present on every publication's domain, as is explained in the following paragraph.

The home-page headlines or RSS feeds were collected first by using a website called archive.org. Those links were then run through the BeautifulSoup scraper. Therefore, not all the articles from a publication's homepage were included in this data set because only articles that were prominently placed were scraped. This data set is part of a larger data set that has 2.7 million news articles spanning from 2013 to early 2020. The data set was then specified from 2016 till July 2017 and specified to include major publications like New York Times, CNN, Business Insider, etc while still attempting to get a range of political alignments with a mix of print and digital sources. However, there were no specific sampling strategies that were used. The data that we will be using are all directly observable from various publication's websites. This dataset was created by Andrew Thompson from his personal interest and inspiration for topic modeling and sentiment analysis.

This dataset has been distributed through websites like components.one and kaggle. The author has updated the dataset on 4/3/2020 on the components.one website and released the dataset on Kaggle in 2017. This dataset was not created for the purpose of distribution to third parties but for the author's personal interest in topic modeling and sentiment analysis. The dataset also has not been distributed under a copyright,

intellectual property license, license and other regulatory restrictions apply to the dataset. This dataset has not been updated since 3 years prior, but there is a larger version of the dataset containing 2.7 million articles. Since this dataset consists of archival data, there is no concern over whether it is obsolete, especially since we are not asking a question that requires current news publications.

Methods

To answer our research question of how the press & many other media outlets may be politically biased, we decided to first specify the dates of our interest. We chose to start from March 18th, 2015 as this was when Trump first announced his campaign and we chose to look at articles up to April 1st, 2017 which was the first 100 days of Presidency for Trump.

Then, we needed to identify articles that mentioned Trump or Hillary. We initially thought that it would be required to use NER (Named Entity Recognition) on the entire article to identify the democratic and republican candidates from different publications. After attempting this, we realized that this process could be simplified by applying regular expression matching on the titles of these news articles. We found from close reading that we were successfully able to identify articles that mentioned Trump and Clinton with a few exceptions when Clinton referred to the Clinton Foundation. Additionally, we wanted to ensure that each publication had a sufficient amount of articles that mentioned Trump and Clinton. We found that certain publications, like Buzzfeed News and the Atlantic, had an underwhelming amount of articles about Trump and Clinton. Therefore, we dropped publications that had less than 200 articles about each political candidate.

After running sentiment analysis using VADER, we realized that this process was not only demanding for our computers but also ineffective as some articles largely focused on topics irrelevant to our presidential candidates. To overcome this issue, we searched for sentences that included the presidential candidates' names and took the next sentence to perform sentiment analysis. This allowed us to perform sentiment analysis on all of our data from March 18th, 2015 till April 1st, 2017.

Then, we applied sentiment analysis using VADER on articles that mentioned the candidates Hillary and Trump. This provided us with information on the compound, positive, negative, and neutral scores. We focused on the positive and negative sentiment scores as we found these scores would be the most indicative of each media outlet's political bias.

While we believe that VADER will be able to effectively generate positive, negative, and neutral sentiment scores, we wanted to highlight some of the limitations with VADER. VADER was created and trained using social media text. This meant that the context in which VADER was trained was more informal than some of our major publications like the New York Times. VADER also struggles with complex sentiments like irony, sarcasm, and mockery. However, from some close reading, we were able to find that the majority of our articles did not include these complex sentiments.

After calculating the sentiment scores for each article, we aggregated the articles by date by averaging the negative and positive sentiment scores to examine how the sentiments of each candidate have changed over time. We performed a similar analysis while grouping for publication to examine potential biases in each

publication. We also aggregated the articles by publication to find the average negative and positive sentiment scores, which we will dive into in our analysis.

Analysis & Discussion

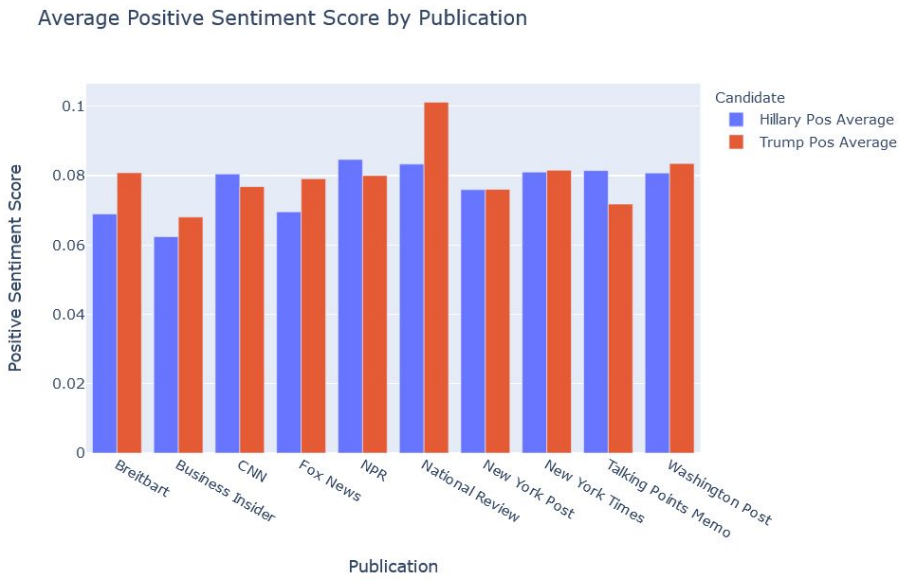

Figure 1

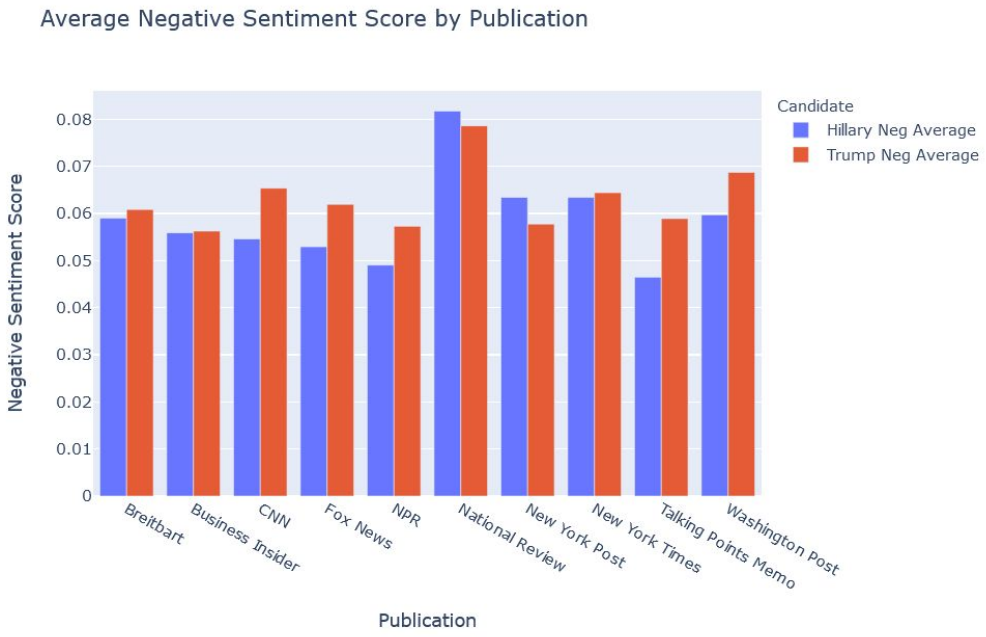

Figure 2

Figure 1 and 2 shows the average positive and negative sentiment scores respectively. The publication with the most positive sentiment towards Trump is the National Review, which is to be expected due to their right-leaning focus. According to the figures, the publications that supported Clinton more so than Trump includes CNN, NPR, and the Talking Points memo. In all three of these publications, their sentiment scores showed that they had a higher positive sentiment score regarding Clinton as opposed to their positive sentiment towards Trump. These publications also had a lower negative sentiment score regarding Clinton compared to the scores for Trump. From what we know about the image of these publications' and their assumed biases, these three publications (CNN, NPR, & the Talking Points Memo) all have an image of being left leaning. Therefore, these scores seem to support what they are already known for.

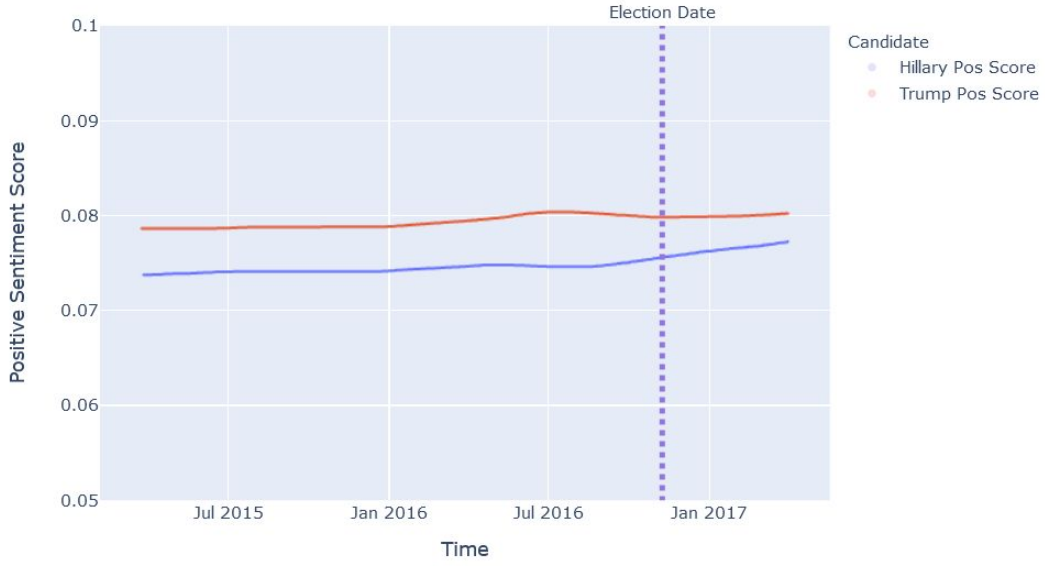

Figure 3

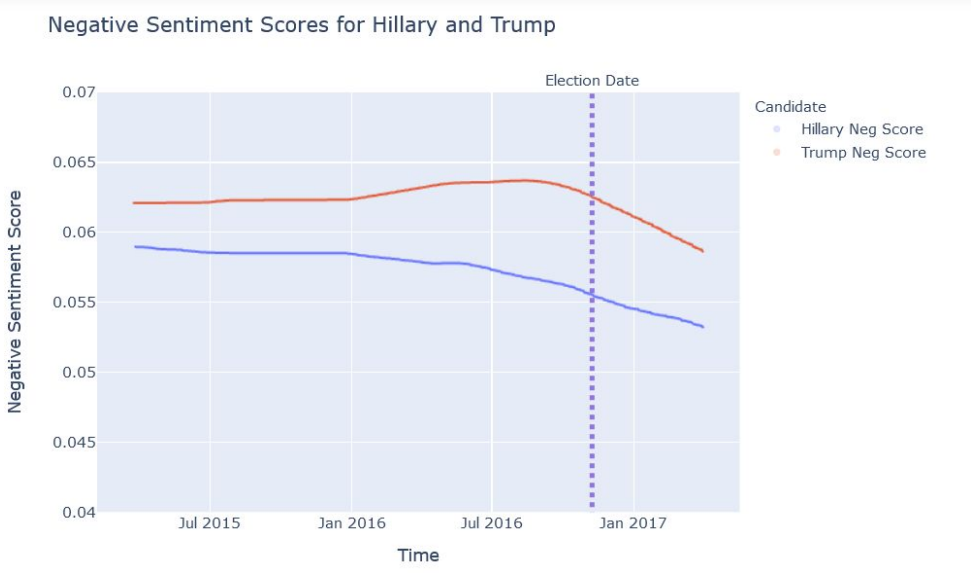

Figure 4

Figures 3 and 4 show the average positive and negative sentiment scores over the specified time period. These results are notable in that both candidates' sentiment scores remained relatively constant. It was not until after the election that there was a dip in negative sentiment and a positive sentiment. There are a couple of possible explanations for this phenomenon. It could be that since the election is over and not as hot of an issue, many publications toned down their rhetoric and support of either candidate in favor of other stories. In other words, since Trump got elected, the authors and reporters no longer had much reason to put out pieces portraying either candidate negatively.

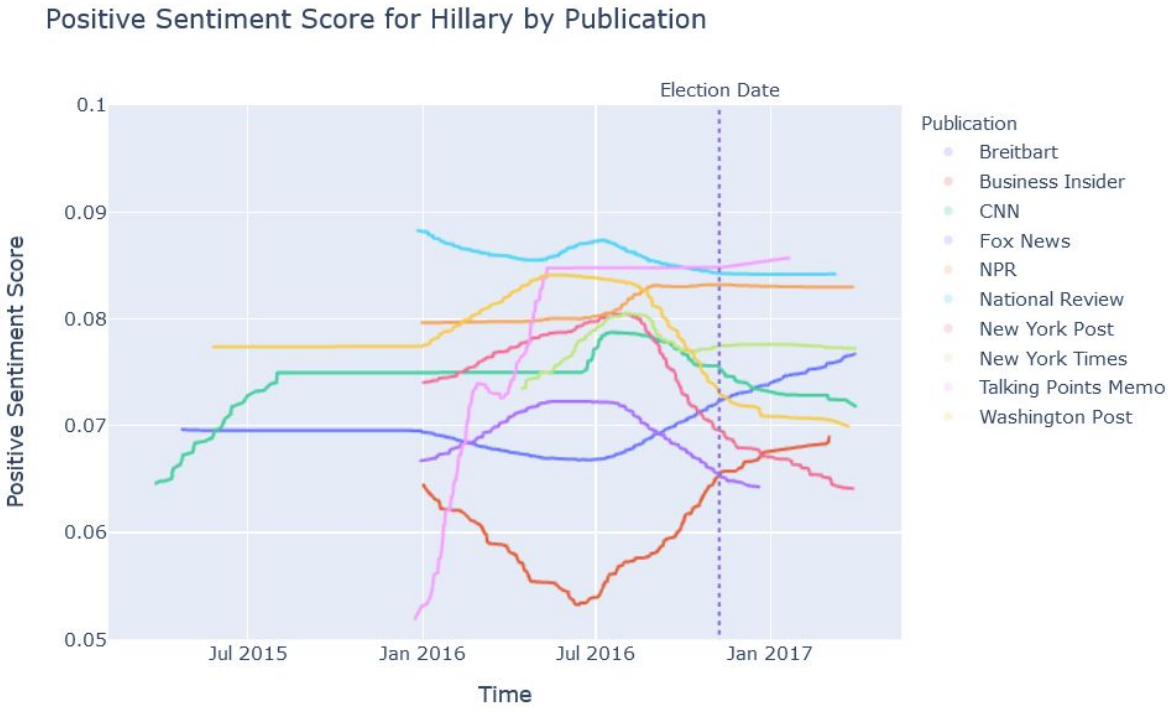

Figure 5

Figures 5 and 6 show the average positive sentiment scores for Clinton and Trump, respectively, by publication. Figures 7 and 8 show the average negative sentiment scores with the same parameters. It is interesting to see in figure 6 that Trump's positive scores remained relatively constant throughout the campaign and election cycle, while, as seen in figure 5, positive sentiment towards Clinton fluctuated far more. Based on figures 5 and 7, it appears that the publication 'Talking Points Memo' made it a point to be very supportive of Clinton, as seen in the sharp rises in positive sentiment and dips in negative sentiment in regards to Clinton. However, this change in sentiment towards Clinton was not seen in their sentiments regarding Trump. As seen in figures 6 and 8, the positive and negative sentiment of 'Talking Points Memo' towards Trump was not nearly as significant as changes towards Clinton.

It is also worthy to point out that there was a shift in sentiment towards Clinton around June 2016, which is when she won the Democratic Party nomination to be president. Because of this, many publications, who may

have been more supportive of other democratic candidates, switched their focus towards Clinton. This is reflected in figure 4 as a decrease in overall average negative sentiment beginning about June 2016. It can also be seen in figures 5 and 7. Clinton winning the democratic nomination had little impact on sentiment towards Trump.

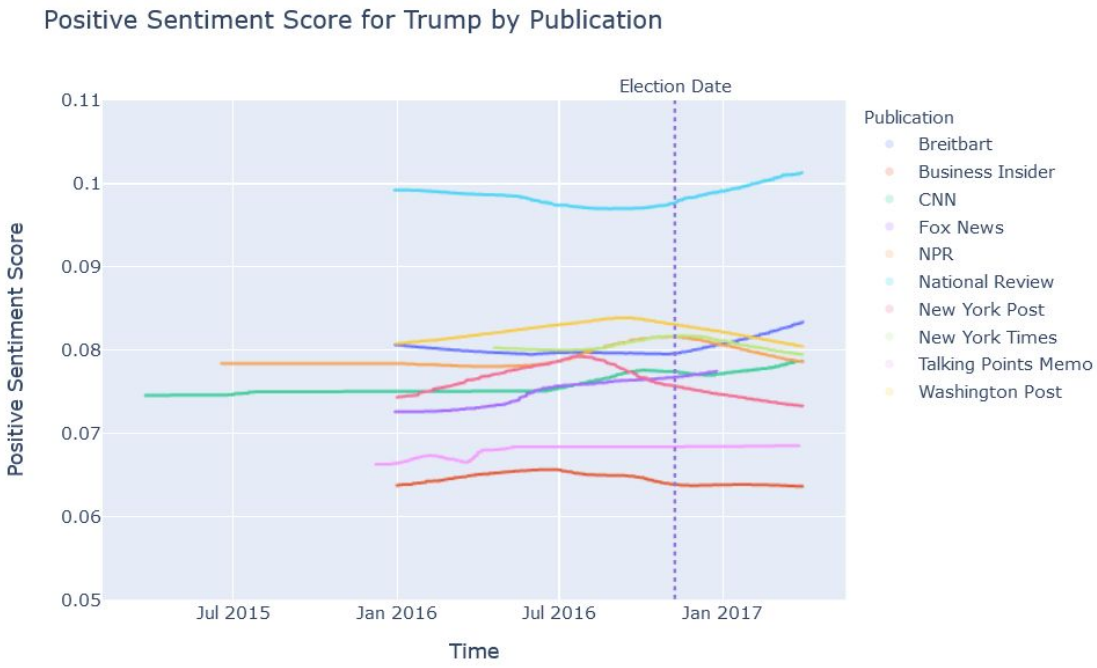

Figure 6

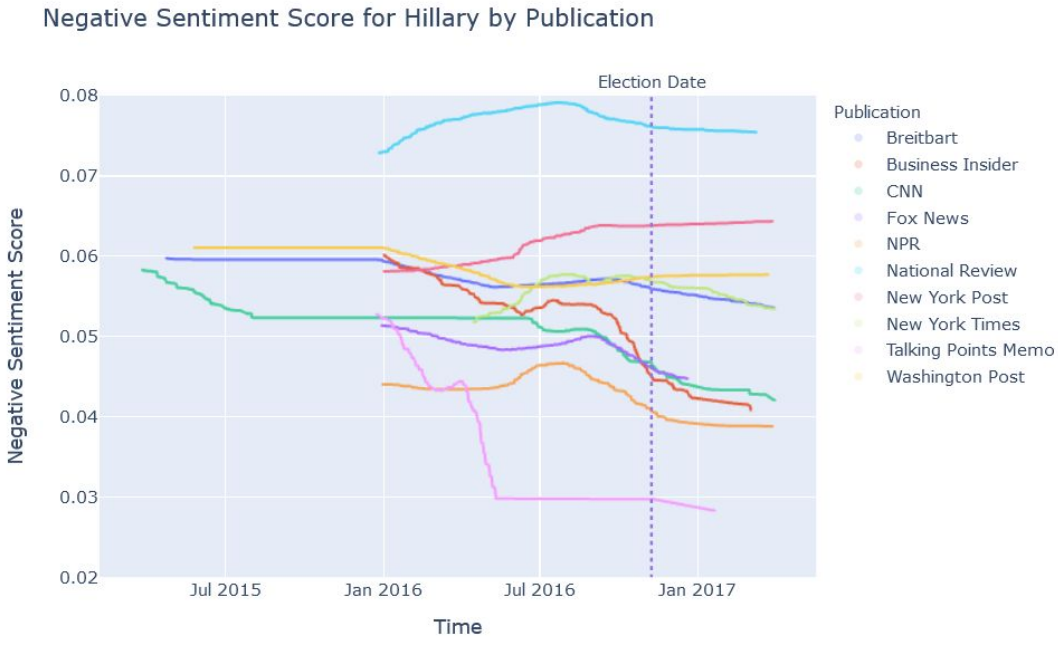

Figure 7

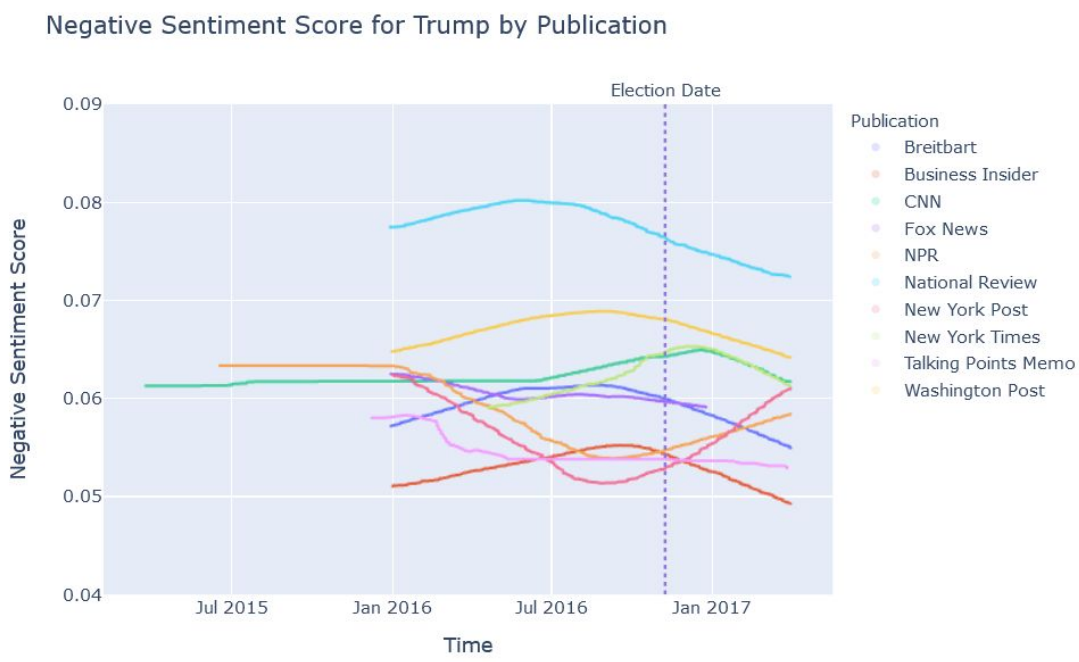

Figure 8

Based on the results, we can say that 'National Review' was the publication that was most favorable to Trump, while 'Talking Points Memo' was most favorable towards Clinton. This makes sense since 'National Review' calls itself a "leading conservative magazine and website" (figure 9). 'Talking Points Memo', on the other hand, has a more liberal streak (figure 10). It is also run by Josh Marshall, an American blogger and journalist, and his voice may have been the most vocal and what led to the large increase in positive sentiment towards Clinton.

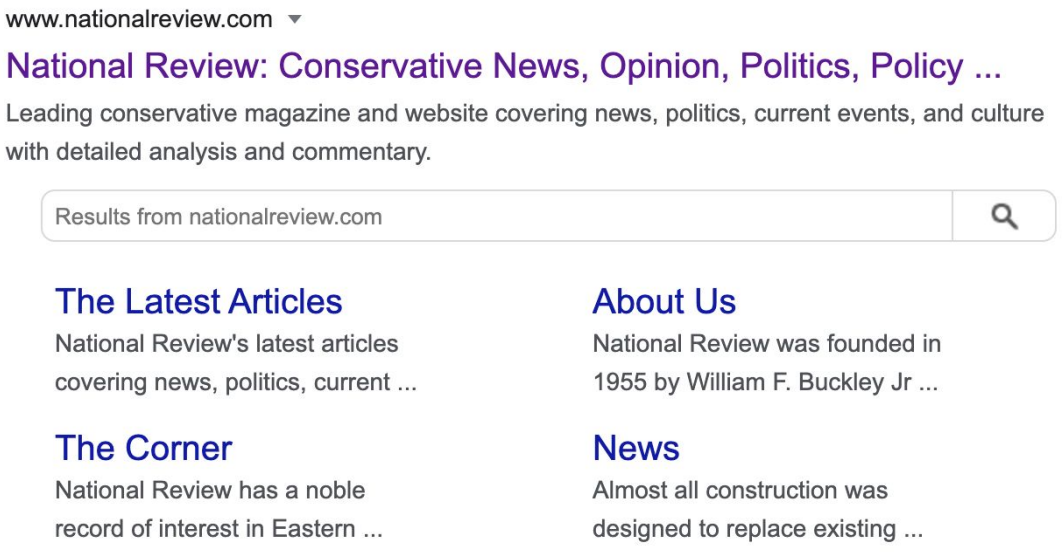

Figure 9

Figure 10

Conclusion

As is the case with many sentiment analysis projects, this project was limited by performance of the sentiment analyzer, VADER. Of course, this limitation was mitigated through having a very large corpus as well as going back and reading through some of the articles that were given more polarizing scores. Since we did not want to use all of the content within the article, we only did sentiment analysis on the sentence where the candidate was mentioned and the sentence following it. This could have led to some inaccurate sentiment scores and because the corpus is so large, it is unfeasible that we could close read the entire corpus. Thus, it could be possible that we missed certain elements that, if close read, would have led us to disagree with the sentiment score.

It is also important to note that certain publications do not have many articles, or any articles at all, prior to January of 2016. This is due to how the data was collected. It only pulled & scraped the most prominently placed articles from each of these publications. In other words, it only pulled articles primarily from the home page & received many views. Therefore, there may be articles that speak of Trump & Clinton that just weren't pulled due to the exclusivity of the data extraction.

Possible next steps would be to try a different sentiment analyzer to see whether the results are altered as a result. Another possible question to consider is whether the title of an article differs in sentiment in regard to its content. For example, in order to garner more clicks and views, authors or publications may have used more polarizing titles as compared to the actual content of said article.

This project of analyzing political bias among various different publications is important work because we want to be aware of the different biases fed to us by the source of where we get our news. While certain publications may not have strong evidence leaning towards one specific bias, there are some, like National Review and Talking Points Memo, that strongly lean one way or the other. Therefore, should one receive news sources from them, they should be aware of their biased tendencies in order not to be accidentally misled by the sentiment of their content.