Thomas Bagley
Roxanne Baker
Daniel DeVeau
Jeffrey Ho
Jason Novack

Sentiment Analysis of Political News

CS216 Group 6 Final Report

Github: https://github.com/jasoncode/cs216project

Introduction

Fake news has been a hot topic of the past year, largely due to its uprising during the 2016 US election. Originally meant to raise awareness about false sources of information, fake news has become a politically charged term that is used to discredit opposing news outlets. Misinformation has become so prevalent that social media sites have launched efforts to combat its spread. Our project goal is to perform sentiment analysis on a wide range of real and fake news sources in order to investigate and compare the language they use. We chose to restrict our data to political news, because of our initial interest in the topic.

We hypothesized a significant difference in the language of real and fake news sites, specifically that fake news contains more sensational and negative language than more reputable news sources. Our data collection and analysis was originally planned around investigating this hypothesis, but took us in other directions as well. Although we were met with challenges in the process, our findings generally aligned with our initial expectations.

Data Collection

In order to perform a significant analysis of political news outlets, our goal was to acquire articles from as many sources as possible, representative of the entire political spectrum. After many hurdles, we acquired over 10,000 articles from around 200 sites. Our reliable news sources were scraped manually, the details of which are in our Scraping section. The majority of our fake news was acquired from the Kaggle dataset 'Getting Real about Fake News'. This turned out to be a major challenge of our project, as the dataset was riddled with noise, satire, non-political news, and non-English news to exclude. We were able to hand-clean this dataset, leaving us with 185 new sources of fake news.

Precisely defining fake news turned out to be more difficult than we anticipated due to the wide variety of sites within the Kaggle dataset. Luckily, the dataset assigned an attribute to every site, e.g. 'bias', 'bs', 'hate speech' which helped us weed out sites that did not belong. When absolutely necessary, we manually investigated outlier sites to identify intentions of deception. For example, a site that proclaims itself as satire is certainly not fake news, whereas a site that hides a disclaimer about 'invented facts' likely intends to deceive.

Real News Sites (Manually Scraped)

- NYT
- The Guardian
- Economist

- News Examiner
- Breitbart
- Reuters
- The Atlantic
- Politico
- National Review
- One America News
- Washington Times

Scraping

The first step to acquiring news articles is getting a list of URLs to articles we want to use. For our real news sources, we acquired 500 articles from each source (although a few sites did not have that many, in which case we used all that they had). Since we are mainly looking at averages between websites, we felt that 500 articles was enough to acquire a reliable average from a specific site. For some sources this is possible through an API call, such as the New York Times. But for most we needed to scrape these URLs from the website itself. This became challenging for a variety of reasons. Some sites don't keep articles in categories that are applicable to us, so any site that did have have a politics category or something similar we skipped. Additionally, some sites had a politics category, but did not have any way to access articles older than a few days. Even for some sites that did have enough articles, sites that used infinite scrolling or required button clicks but did not update the url are very difficult to scrape without advanced methods. Therefore, we chose to skip these as well since we were able to find a reasonable number of sources without this problem. Websites that had a politics section, and also had an archive style page with page numbers were used for our project.

In order to perform the scraping, we used the python library lxml, in combination with xpath, in order to scrape data. These tools allow us to easily access specific sections of the html and acquire all text from that section and child sections. For example, many articles are constructed with sequential <p> tags, often with child tags for things like links. Lxml and xpath allowed us to acquire all of the text in just a few lines. We then encoded the text to normalize formatting and write it to a text file. The text file was then used as input for the sentiment analysis. Once we scraped URLs from a site, the next step is to scrape the article body text for each article. This step is fairly straightforward, but requires some individualization for each website.

While scraping articles is in an effective and relatively simple method for getting articles in bulk from a specific site, there are some challenges for a project like this where we need a variety of sources. These mainly revolve around the fact that every website is a little bit different, so it is difficult to code a scraper that works for every website. In order to make this simpler, we wrote a util that manages requests and error checking, as well as some of the logic behind scraping articles in bulk and formatting them to text files. However, since every site has different HTML and text formatting, every site needs to be examined manually using browser dev tools in order to manually write the xpath. Additionally, we occasionally need to do some extra text processing for sites that have sections that shouldn't appear in the article text.

Methodology: Sentiment Analysis

We used a Python library called textblob which enables sentiment analysis based on its extensive corpora. The documentation for the library is at the following link for additional information on the library as a whole: https://textblob.readthedocs.io/en/dev/. The key aspect for our project is the sentiment feature, which returns a two-element tuple consisting of 'polarity' on a [-1,1] scale and 'subjectivity' on a [0,1] scale. Polarity can be defined as the emotion or opinion expressed in the text, with a -1 value being most negative and 1 being most positive. Subjectivity measures how opinionated text is, with 0 being most objective/factual and 1 being most subjective/opinionated. The file textblob_test.py includes several intuitive example sentences demonstrating the functionality.

Our analysis file computes the sentiment for every article, and outputs three files which report the data in different ways:

- Results_ByArticle.csv - This file contains a row for each article containing the source name, polarity, and subjectivity, as well as summary statistics for each overall category of news
- Results_BySource.csv - This file lists results article by article but listed under a clear header for its source, as well as summary statistics for each source and each overall category of news.
- Results_Source_Summary.csv - This file contains a row for each source, listing the number of articles from that source and summary statistics. Because this only looks at summary statistics, sources are only included if they contain at least five articles.

As noted above, much of the data output hinges on summary statistics. The statistics we chose to output are as follows:

- Mean
- Absolute Value Mean - This one was included because polarity is on a scale from -1 to 1. Thus, if a source had a lot of articles that were polar but on opposite ends, the mean could come out close to 0, which would not accurately reflect the polarity. In this case, if most of a source's articles
- Median
- Absolute Value Median - Included for the same reason as absolute value mean

Analysis and Visualization

Our primary form of visualization is scatter plots. These plots place subjectivity on the x-axis and polarity on the y-axis, creating a clear picture of the sentiment analysis of the articles we have gathered. We were able to make several scatter plots visualizing different facets of the dataset. We also constructed word clouds for real and fake news to look for any interesting differences in 'buzz words'.

Sentiment Analysis by Article

The most comprehensive plot was of the sentiment analysis of all articles that we have scraped (see Figure 1). Since there are far too many data points within a small range for a simple scatter plot to effectively display it, we opted for a heat map showing what regions of the subjectivity-polarity space contained more fake or real news articles: redder squares have more real news, while bluer squares have more fake news, and gray squares have an even distribution of both.

One interesting finding is that there's a higher proportion of fake news towards the more subjective end of the scale (although there are real and fake news articles all across the spectrum). This is in line with our hypothesis that fake news is more likely to use emotionally charged language than real news.

Fake news also seems to have a greater spread of polarity; while most real news is centered fairly close to 0.0 with some articles trailing off, fake news has a lot of data points across the polarity scale. However, unlike with subjectivity, it is fairly difficult to tell from visuals alone if there's a significant difference in the polarity across real and fake news sites.

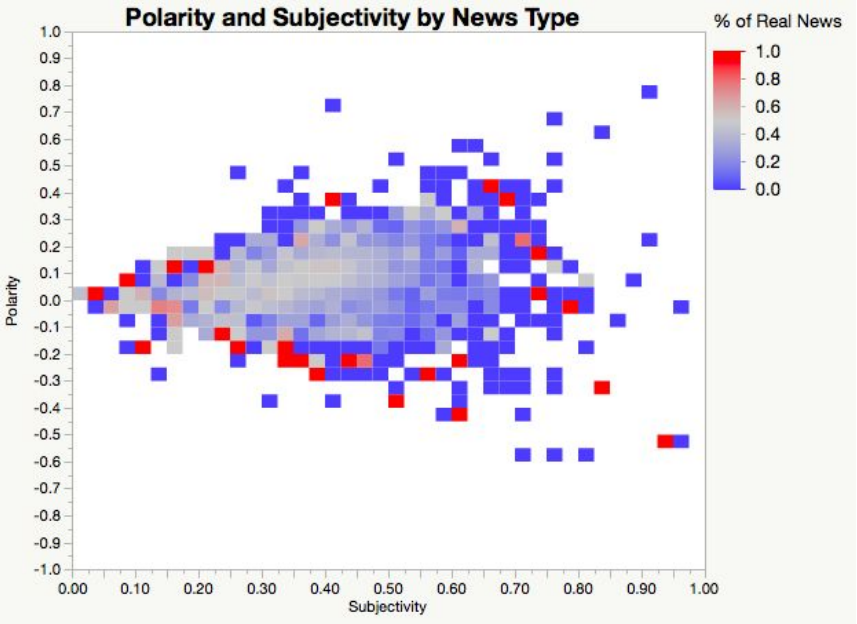


Figure 1: Heat Map of Polarity and Subjectivity by News Type

Sentiment Analysis by Source

We also plotted the mean of the subjectivity and polarity for each source (Figure 2). Since we had far fewer data points, we were able to use a traditional scatter plot for this visualization. We also restricted the fake news sites to those for which we had at least 95 articles to reduce the impact of outliers.

While there are too few sources to draw any statistical conclusions, we still see interesting trends. While the polarities are mostly pretty close together (with the exception of the two fake news sources with polarities over .10), the subjectivities of most of the fake news sources are higher than those of most of the real news sources. This is again in line with our hypothesis that fake news would be more sensational than real news.

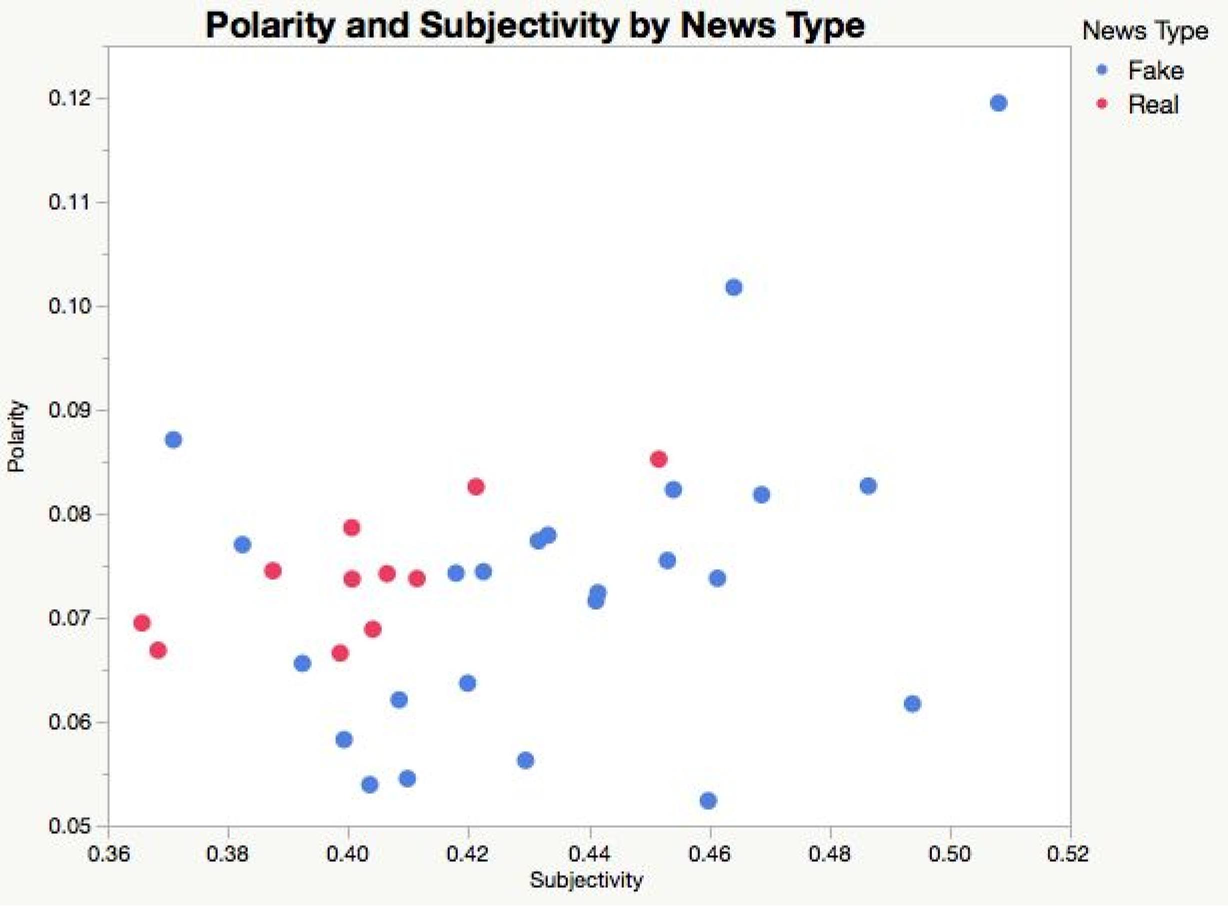


Figure 2: Mean of each source’s polarity and subjectivity, with blue dots representing fake news sources and red ones representing real news sources.

For our real news sources, we also compared left-leaning sources with right-leaning sources, as this was another axis of comparison. We again did this with a heatmap, with blue representing left-leaning sources and red representing right-leaning sources. For the most part, the blue and blue tinted squares are closer to the center of the heatmap, while the red squares are all around the edges. This suggests that the left-leaning sources are densely packed around the center, while the right-leaning sources are less densely clustered and cover a wider range of values.

The center of the data region is right around 0 polarity. Given the spread of right-leaning articles, this suggests that articles from right-leaning sources tend to be more polar than left-leaning sources, but not necessarily more positive or negative. The less easy-to-understand result is the subjectivity; the center is around .375, and there are red squares on both the left and right edges of this. One possible explanation is a wider variety of right-leaning sources in our dataset; because there are fewer major right-wing news sources, the ones that we were able to use are less standardized than the left-leaning sources we had access to.

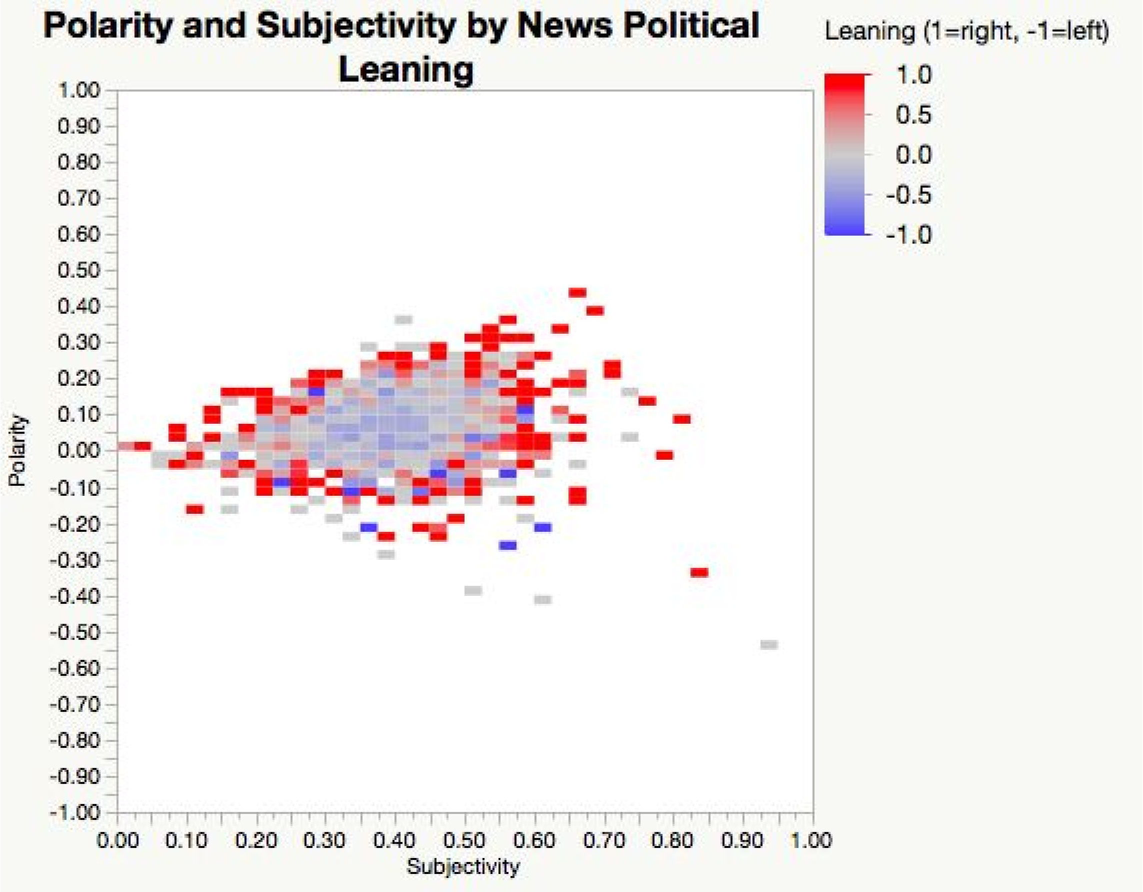


Figure 3: Heatmap of each article's polarity and subjectivity by political leaning of its source

Word Clouds

We also created word clouds; one using the entirety of the fake news data, and one using the entirety of the real news data (see Figure 4 and 5 respectively). This was done using the 'wordcloud' Python library (see https://github.com/amueller/word_cloud) on its default settings, showing the top 200 words after removing common stop words (e.g., 'the', 'of', 'is', etc.). This yielded some interesting results. Some of the differences are likely to result from the gap between the article dates. Most of the fake news articles are from October or November of 2016, meaning that there are a lot of articles focused on the election; meanwhile, our real news articles are mostly from the last month. This likely explains why "Hillary Clinton" appears only on the fake cloud.

Other differences, like the inclusion of "Russia" and "Russian" on the fake cloud, or "Mr Trump", "Republican", and "Democrat" on the real cloud seem less likely to result from this. US relations with Russia are fairly tense, and fake news may be more likely to capitalize on this to create more interesting stories. Meanwhile, real news sources may have a stronger emphasis on respect and thus be more likely to refer to Donald Trump as "Mr. Trump". The inclusion of "Republican" and "Democrat" on the real cloud may indicate that real news is more likely to discuss party differences as a whole instead of focusing on specific individuals.

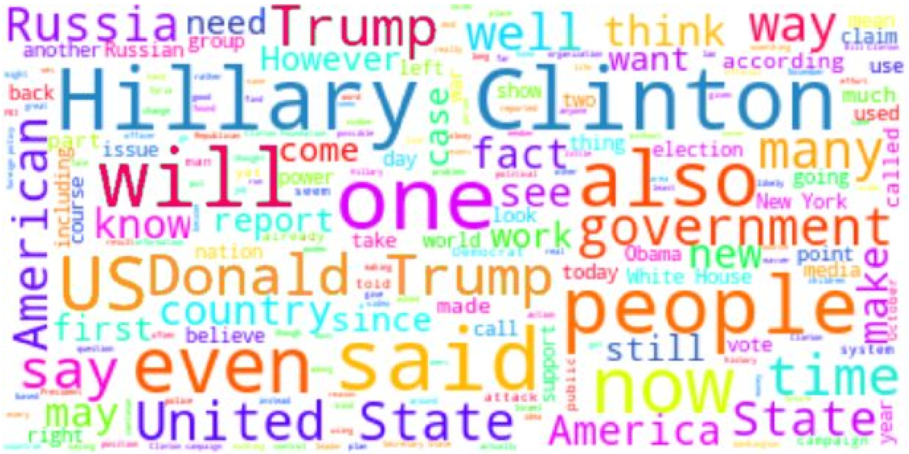


Figure 4: Fake News word cloud

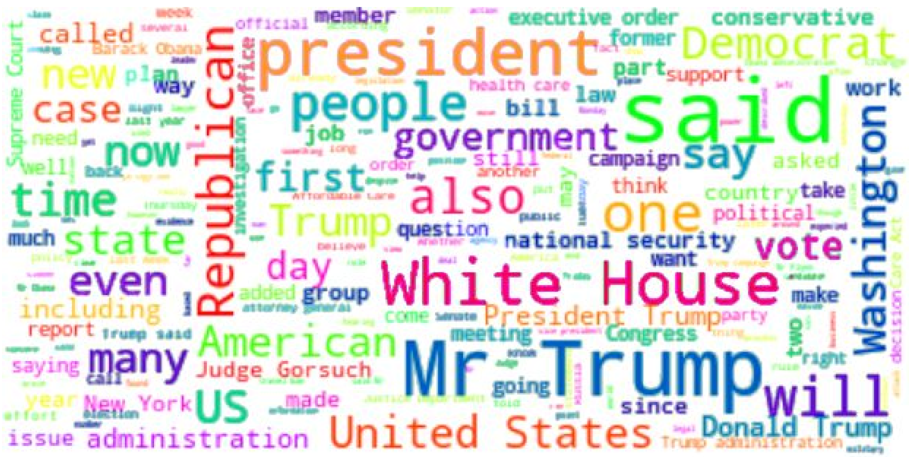
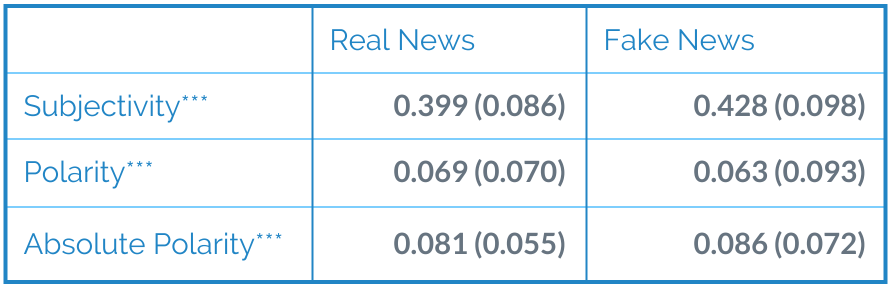


Figure 5: Real News word cloud

T-test Analysis

When comparing between the sentiment analyses of real news and fake news, numerous differences appeared statistically significant. The results of two-tailed t-tests that compared the mean subjectivity, polarity, and absolute polarity are illustrated in Table 1 below. All three sentiment analysis parameters differed at a statistically significant level when measured at an alpha level of 0.0001. Thus, the data suggests that the language used in real news articles differs on at least some level from the language found in more mainstream, real news sources.



	Real News	Fake News
Subjectivity***	0.399 (0.086)	0.428 (0.098)
Polarity***	0.069 (0.070)	0.063 (0.093)
Absolute Polarity***	0.081 (0.055)	0.086 (0.072)

Table 1: Mean and Standard Deviation of Sentiment Analysis by News Type
(***p < .0001)

For instance, the mean subjectivity score for fake news articles was significantly greater than the mean subjectivity score found in real news articles. The results suggest that the language in fake news articles is generally more opinionated and subjectively charged than the language found in real news articles. Similarly, fake news articles generally scored lower polarity scores than real news articles, thus suggesting a more overall negative tone in fake news article language. The results are not particularly surprising, suggesting that real news tends to adopt a more positive overall tone in language, whereas fake news tends to be more negative and critical. Lastly, the two news types also differed significantly in regards to absolute polarity. In general, fake news sources scored higher in absolute polarity. Since absolute polarity measures the extremity in language (in either a positive or negative direction), the results suggest that fake news tends to be more sensationalist in nature. Regardless of whether the news is positive or negative in tone, fake news articles tend to use more extreme and emotionally charged language when compared to real news articles.

Conclusion

Overall, the results of our analysis appear to support a lot of our prior hypotheses. In general, we found the variance in sentiment analysis scores to be much greater amongst fake news sources. Whereas real news sources tended to cluster around similar subjectivity and polarity scores, fake news articles tended to vary greatly in both subjectivity and polarity. In addition, when compared through two-tailed t-tests, fake news sources, on average, scored higher in subjectivity, lower in polarity, and higher in absolute polarity. The results thus suggest that the language in fake news sources is more opinionated in nature, slightly more negative than real news, and more extreme overall. In addition, the results of our analysis also illustrated interesting conclusions when we compared purely amongst our real news sources. In general, our right leaning sources appeared to be less densely clustered in subjectivity and polarity scores, though this may be a consequence more of the sources we selected rather than right and left leaning news in general. Thus, our analysis suggests that the language of a news article could vary substantially depending on its origin, a conclusion which may prove useful in future attempts to identify fake news sources.