

2016 Bias Finder

Final Project Report

12 Thursday May 2016

POSTED BY PLATAPOOP IN UNCATEGORIZED

≈ LEAVE A COMMENT

BiasFinder

Website: manticorerays.com

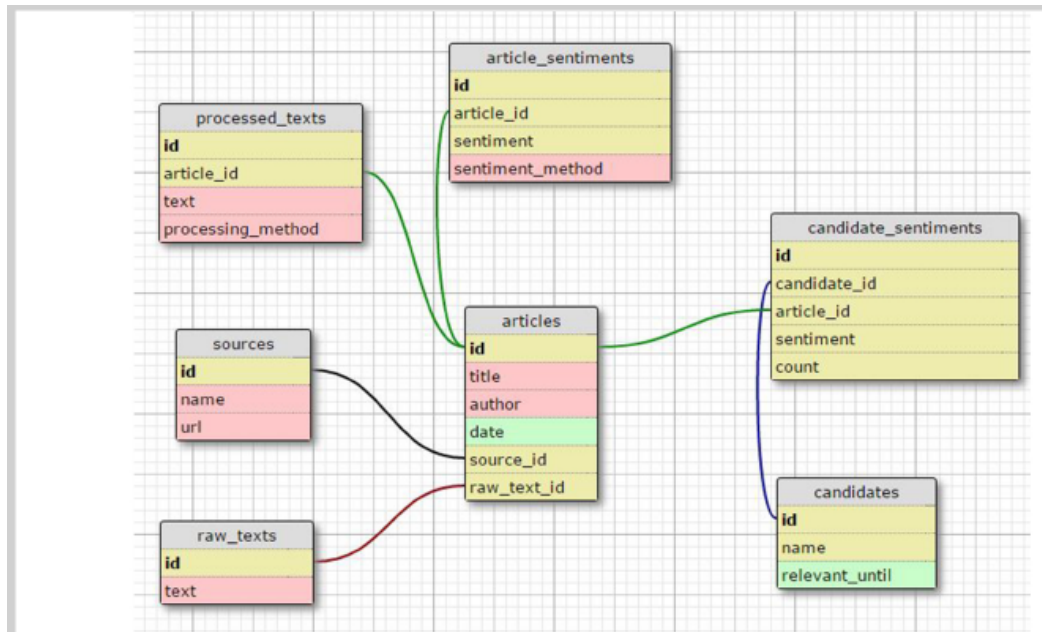
(note that the website links to the IP address as the domain name may not have propagated)

Poster

Summary:

Our initial goal for this project was to find the biases in news articles. Was a given website liberal or conservative, does it support Hillary over Sanders, etc... We have accomplished many of these initial goals and more including finding biases toward a candidate given a website, calculating the average sentiment and political affiliation of websites, determining how often a given website talked about candidates, and finding websites that are similar to other websites. Overall, we believe we have accomplished what we set out to do this semester, but if given more time, we would like to increase the accuracy of our sentiment analysis (e.g. retrain Stanford CoreNLP's deep learning neural network on a more relevant dataset).

Explanation of Data



This is our database schema. It is normalized so there is no redundant data that is being stored in the database. Articles, candidates, sources, and raw_text is populated with the news articles scraped from Google. We run our sentiment analysis algorithm on all of the articles (raw_texts), and fill out article_sentiments and candidate_sentiments. You will notice that we have a table called processed_texts, which would have included articles after post-processing (shortening of sentences or removing stop words). However, we did not have the time to go through and process the articles. sentiment_method is currently unused because we only used one method for sentiment analysis (Stanford CoreNLP).



Figure 2 - Trump National Poll Data and Aggregate Sentiment over Time

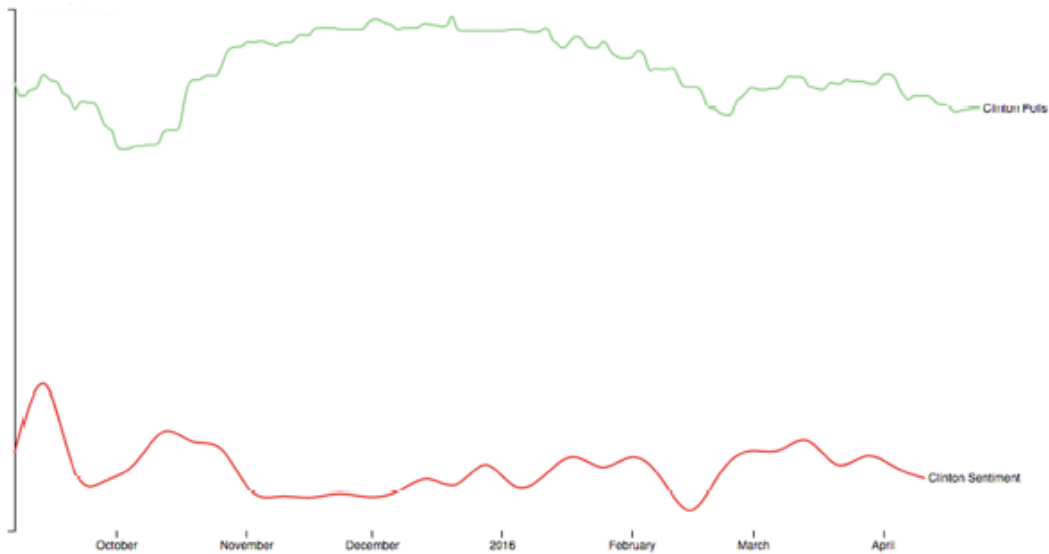


Figure 3 - Clinton National Poll Data and Aggregate Sentiment over Time

These figures show poll results and average sentiment for Hillary and Trump over time. We can see from Figure 3, the graph about Clinton, that there is a correlation between polls and sentiment (this could be sentiment reflecting polls, or another factor driving both, we do not know). Another interesting finding can be seen in Figure 2. We see that Trump's sentiment fluctuates wildly and is fairly

uncorrelated with polls which consistently trend upwards. What is interesting is that his sentiment spikes up drastically after it is apparent that he will be the Republican nominee. The polling data was obtained from [FiveThirtyEight](#). Trump, Kasich, Cruz, Sanders, and Clinton's graphs can be seen on our website.

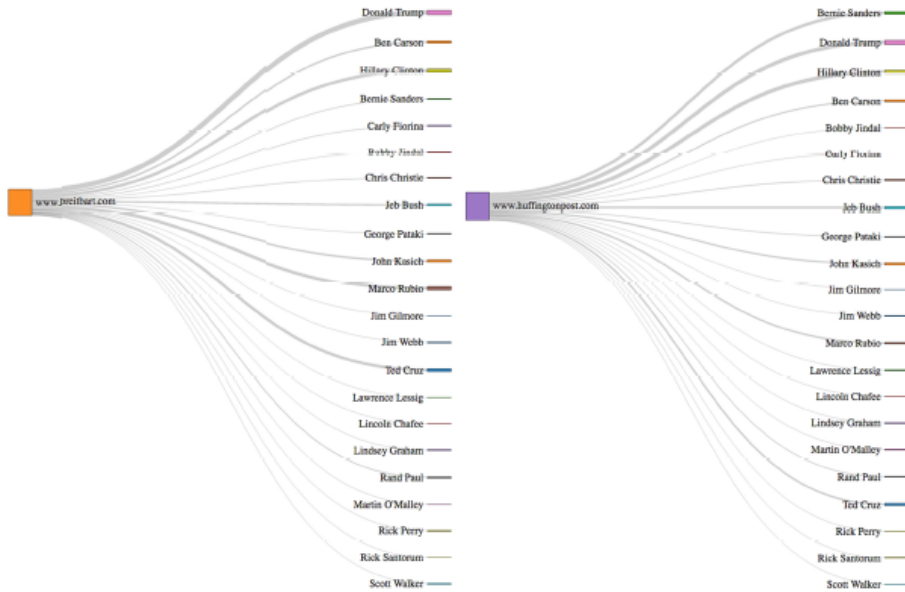


Figure 4 - Breitbart's Candidate Mention Count Figure 5 - Huffingtonpost's Candidate Mention Count

These graphs show websites and how often they talk about candidates. Based on our data, conservative websites talk more about conservative candidates, and liberal websites talk more about liberal candidates, which is consistent with expectations. Data for washingtonpost, breitbart, politico, huffingtonpost, and other news sources can be seen in a more readable format on the website along with more information.

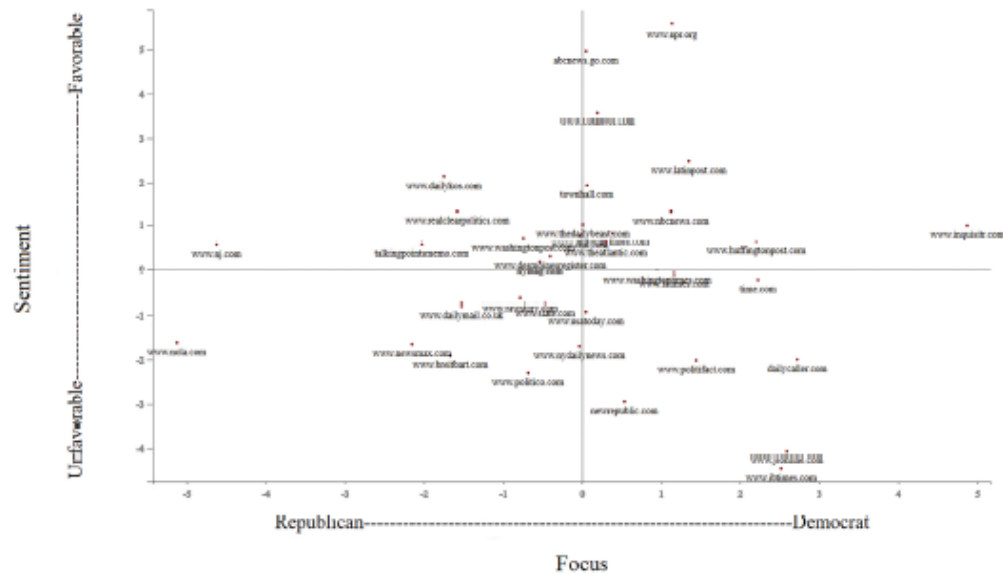


Figure 6 - Content Focus and Average Sentiment for Popular Sources

This graph shows sentiment and focus for 40 or so websites, with the x-axis being which political faction they talk more about (conservative <--> liberal) and the y-axis being general sentiment of the articles (favorable <--> unfavorable). From this, one may also find websites that are politically similar to one another. A tool on our website is included to find similar websites.

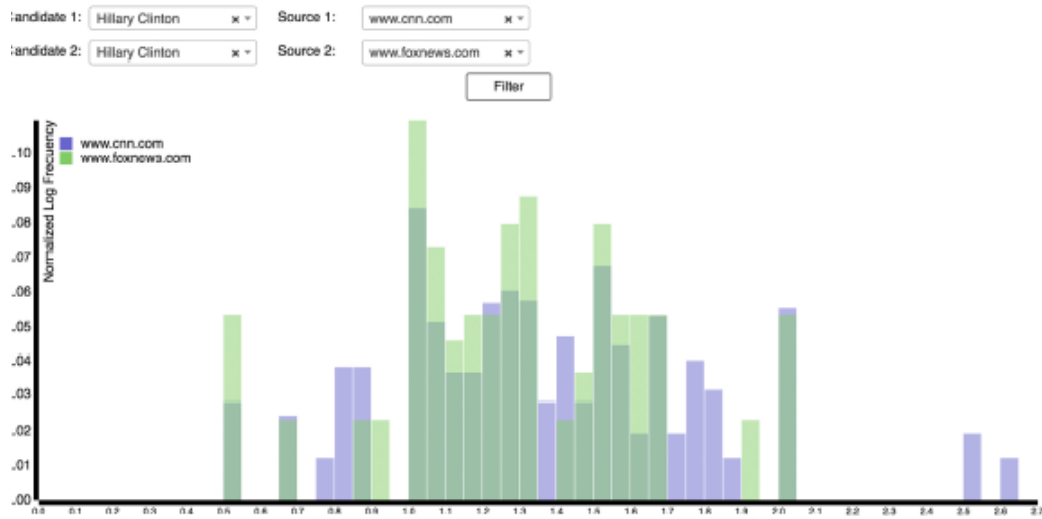


Figure 7 - Comparing Sentiment Counts for Clinton, CNN vs Fox

These graphs show general sentiment toward candidate(s) given website(s). It gives a quick way to see how a website feels about a particular candidate compared to another website (or how two websites differ in opinion about the same candidate). For example, from figure 7, we can see that Clinton is favored more by CNN than Fox News. You can play around with this tool on our website too.

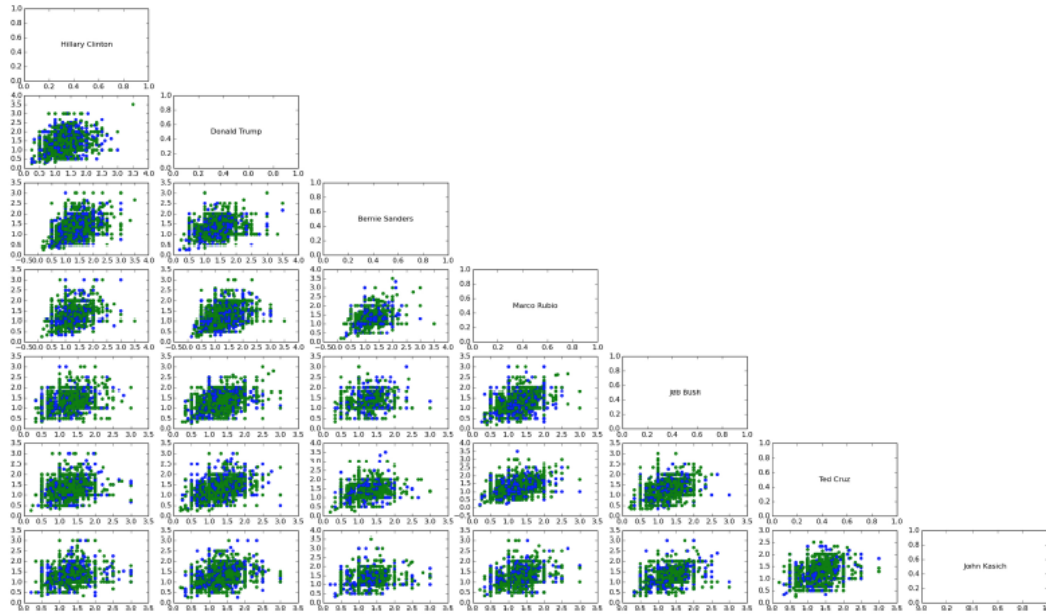


Figure 10 - Pairwise Correlation between Candidate Sentiments in the Same Article

This scatterplot matrix was made to see if there were any notable trends between candidate pairs (i.e. articles that tend to talk positively about Bernie will talk negatively about Hillary) but no such correlations were found.

Hypothesis

The general prediction was that political bias would be present on some news sites. Our hypothesis predicted that Fox News would have a Republican bias, while NYTimes would have a Democratic one. In addition, we believed that some sites would have preferences towards or against certain candidates. Finally, we predicted that sentiment would be correlated with poll results.

Data and Methodology

Our data is an aggregation of over 80,000 news articles from thousands of different websites. This was obtained by scraping Google News for URLs using Dan McInerney's Python Selenium framework, starting from September 2015, all the way to April 2016 for each candidate. Then we used Python's newspaper library to scrape the links we collected. The relevant information for retrieved is the URL and the text, but also included is title and author. Data was initially stored in Sqlite, but then migrated to MySQL. Initially, we used Python's TextBlob to get articles' sentiments and did show that newspaper sources wrote with some biases toward a candidate. However, we could not be certain because articles often referred to multiple candidates. Because of this, we adopted Stanford's CoreNLP framework which does pronoun resolution. We deployed our code on Amazon Web services and ran over 300 compute hours on spot instances of m2.2xlarge (8 vCPUs, 32 GB RAM) over the 80,000 articles to compute sentiments. We had a wrapper in Python that was responsible for creating spot instances at a specified price, and running the java process that analyzes a subset of the articles. The findings (candidate sentiment and sentiment per article) were then put into MySQL.

Results

The Good:

- We saw that some websites focused a lot on certain candidates (e.g. HuffingtonPost likes writing about Sanders).
- Figure 6 shows that our data seems to be consistent with the observations made by the Pew Research Center in 2014 for most websites. Pew Research. Note that this research does not discern between different media outlets.
- Many candidates had sentiment that moved similarly with poll results.

The Bad:

- Google news is biased. A quick Google search shows that many people have complaints that Google News is biased toward the Democratic Party. This could affect the distribution of articles we got each week.
- Stanford's CoreNLP was trained on movie reviews, where it is a lot easier to determine what is positive and what is negative. News sources are rarely as explicit and have different sentence structures which may result in less accurate predictions from Stanford CoreNLP.
- Pronoun resolution is hard. Some larger articles never got analyzed due to memory constraints, this might bias our results in some way.

Achievements

- We obtained meaningful results by analyzing the portrayal of candidates and the political stance of news sources.
- We have a large dataset of political news articles which is something not readily available on the internet. This dataset could be used for interesting things such as machine learning to automatically generate political articles.

More to Explore

- Better sentiment analysis (i.e. better training set, wait for NLP techniques to advance, etc...).
- Visualizing different aspects of the data (3D plot of sentiment distributions over time, most common candidate/sentiment combinations, etc...)
- Poll prediction (using various attributes of the weekly news articles to predict next week's poll results).

[Blog at WordPress.com.](#)