Finding Political Bias in Online News

Authors: Dazheng Zhang, John Yang, Paul Johnson, Kevin Shu, and John Joe Friedmann
CS1951A: Data Science

The purpose of this project was to determine the bias of news sources. It was predicted that some sites would have preferences towards or against certain candidates and that sentiment would be correlated with poll results. In addition, predictions were made about the biases of specific news sources and included Fox News having a Republican bias and NYTimes having a Democratic bias.

The data that was collected is an aggregation of over 80,000 news articles from thousands of different websites. This data was obtained by scraping Google News for URLs using Dan McInerney's Python Selenium framework. The framework scraped URLs for articles that were written between September 2015 and April 2016 for each candidate. Then Newspaper, a Python library, was used to scrape article text, authors, source, and date from the links that were collected. Sentiment was then calculated using Stanford's CoreNLP framework. To determine sentiment on a candidate level, pronoun resolution for each of the presidential candidates was run to give context to sentiment extraction. The code was deployed on Amazon Web Services and ran over 300 compute hours on spot instances of m2.2xlarge (8 vCPUs, 32 GB RAM) to calculate sentiments for the more than 80,000 articles. A script using boto3 was written for creating spot instances at a specified price and then running the java process to analyze a subset of the articles in a distributed manner. The results were stored in a MySQL database. The features extracted include average sentiment and political affiliation of websites, how often a given website discussed each candidate, and similarities between websites based on candidate biases.

The data that was collected seems be consistent with the observations made by the Pew Research Center in 2014 for most of the websites that were analyzed. In addition, it was observed that the sentiment of many candidates fluctuated with poll results. The large dataset of political news articles that was collected is unique; it is a dataset that is not readily available on the Internet. This dataset could be used for other interesting purposes, such as machine learning to automatically generate political articles.