Math 189R Midterm Writeup

Andy Liu, Justin Jiang

July 2020

1 Introduction

The upcoming 2020 U.S. Presidential Election will be one of the most important events in the year, and as such, the competition between incumbent Donald Trump and Democratic Nominee and former Vice-President Joe Biden (as well as Kanye West, as of July 4) has attracted substantial amounts of news coverage. Media opinions have also had an increasingly high effect on the election results in recent years, with Hillary Clinton going as far as to claim that media coverage was a significant factor in her 2016 loss to Trump.

As such, our project uses text sentiment analysis and topic modeling to analyze public opinion and "media opinion" to see whether media coverage of the 2020 U.S. Presidential Election is biased toward either candidate. We further refined this by the usage of topic modeling to consider media coverage of the U.S. Presidential Election on specific political topics, in order to consider which aspects of each candidate's policy were preferred.

In terms of the specific progress made for the midterm writeup, we were able to train a topic model on a news dataset using Amazon Comprehend. Additionally, we scraped news headlines and articles from the Google News "Joe Biden" and "Donald Trump" topics in order to represent the most current media sentiment toward Donald Trump and Joseph Biden. Finally, we used a NLTK text sentiment analyzer in order to analyze the sentiment of this news coverage regarding Trump and Biden, which combined with the other aspects that we worked with, allowed us to estimate the general sentiment of news coverage about the two major party candidates surrounding each of the topics found with our topic model.

In the current political atmosphere, many say that the news media is often disconnected with the general public. We intend to the veracity of that

belief by using methods to estimate the sentiment both the public and the media hold against controversial topics important in the 2020 Presidential Election. For example, say we were to estimate sentiment on gun control, a longstanding partisan issue. If we analyze sources that are representative of both the public and the media, then we can estimate the sentiment that each group holds towards gun control. Say we find that the media is overwhelmingly negative on gun control (i.e, supporting widespread gun reform), but we also find that the general public is actually more moderate (supporting some gun reform, but not too much), then we can say that there is a gap between media and the general public. We can scale this method to important topics in the upcoming election to calculate a final "distance" score of how separated the media is from the public.

1.1 Midterm Progress

We did not work on all aspects of our project before the midterm presentation, instead aiming to focus on analyzing media coverage of the 2020 election. Because of this, we did not work on topic modeling or text sentiment analysis for public opinion, instead opting to focus on media coverage.

Our progress was primarily in the areas of topic modeling of media coverage, text sentiment analysis of media coverage, and synthesis of our results from these two aspects of our project in order to find Trump vs. Biden media opinion on the topics found with our topic model.

2 Topic Model

2.1 Data Sourcing

In order to create our topic model, we first needed a data source consisting of news articles from the last few years, in order to see what topics the media was most concerned about. Luckily, we were able to locate a public dataset of over 2.7 million news headlines and articles, called All The News 2.0, containing articles from 2016 to 2020.

2.2 Topic Model

Amazon Web Services (AWS) provides a pre-trained topic modeler in its Amazon Comprehend service. Amazon Comprehend is a fully-managed natural language processing service provided by Amazon that allows users to abstract their models by providing a simple-to-use GUI. Another huge advantage of Amazon Comprehend is that it uses AWS servers to run analysis jobs; therefore, with a large data set like ours, we save computation time because we can take advantage of Amazon's vast computing resources.

Amazon Comprehend has a topic modeling feature that allows users to input documents for topic modeling analysis. Since the data set we are using, All The News 2.0, contains 2.7 million articles and Amazon Comprehend has a input limit of 1.0 million articles, we had to trim the data set to be able to use Amazon Comprehend. We took 1.0 million of the most recent articles in the data set, which we achieved by importing the entire data set into Python3.7 and using pandas to splice the last 1.0 million rows. We further preprocessed the data by removing new lines from the main text of each article. We had to do this because Amazon Comprehend takes as input one document per new line, so therefore, if we did not strip the new lines from each document, Amazon Comprehend would interpret the input incorrectly.

3 Text Sentiment Analysis

3.1 VADER

Unfortunately, our dataset was not labelled with positive or negative sentiment, and due to time constraints, we were unable to write our own text sentiment model with labelled data. As such, we utilized a pretrained model from Python's Natural Language Toolkit, which is known as VADER (Valence Aware Dictionary and sEntiment Reasoner). We chose to do so because it was especially strong with "NY Times Editorial"-like text, and could tell us how strongly positive or negative statements were. However, we intend to try out other models if time permits, especially since VADER is usually better-equipped to deal with more informal text (e.g. social media) than news articles.

VADER calculates individual valences v for each word of an article, and outputs positive, negative, and neutral scores, which represent the propor-

tion of words in the article that are positive ($v \geq 0.05$), neutral ($0.05 > v > -0.05$), and negative ($v \leq -0.05$). The sum of the valences, after normalization, is the compound score. For each topic, we computed a weighted sum of each article's compound score multiplied by the probability (from our topic model) that that article was about that topic. We then divided this sum by the sum of the probabilities that each individual article was about that topic, thus recovering a weighted average of the topic's compound score.

3.2 Data Sourcing

In order to find articles specifically relating to Trump and Biden, we searched our All The News 2 dataset for articles with "Trump" or "Biden" in the headline. Admittedly, this is a rather strict filter, and one that loses many articles referring to Trump or Biden. For the final project, we will either use exclusively sources pertaining to Trump or Biden, as determined by Google News, or write our own classifier to determine whether an article's content refers to Trump, Biden, both, or neither.

By filtering our existing dataset, we were able to source 27158 articles relating to Trump and 2111 articles relating to Biden. These articles covered a total of 58 topics. Unfortunately, the large time frame of our dataset, as well as the higher relevancy of Trump for much of the 2017-2020 dataset, meant that our data disproportionately talked about Trump, and much of the Biden coverage was in 2020, which possibly could've had lower sentiment overall than 2017-2019.

4 Results

We used a Jupyter notebook to synthesize the results of our topic model and sentiment analysis. We generally found that articles in our Trump dataset had a more positive sentiment. However, we suspect this may be caused by sampling error, as we had far more Trump data from a generally wider time frame.

Figure 1: Average sentiment per topic for Joe Biden.

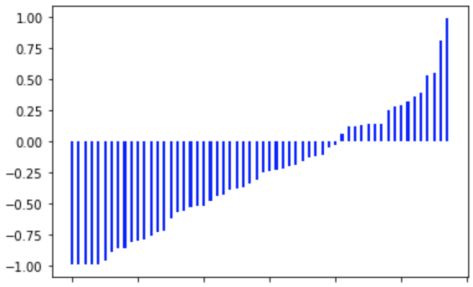
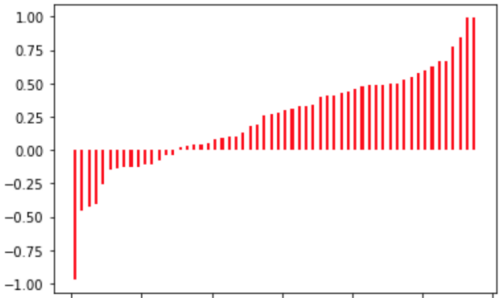


Figure 2: Average sentiment per topic for Donald Trump.



By calculating the difference between average sentiment for Joe Biden and average sentiment for Donald Trump for each of our topics, we were able to find the topics that Biden and Trump were strongest in (relative to their 2020 opponent). We then found key words from our topic model for each topic number in order to estimate the areas in which Biden and Trump receive the most positive coverage, relative to their opponent.

Figure 3: Top topics for Joe Biden (blue).

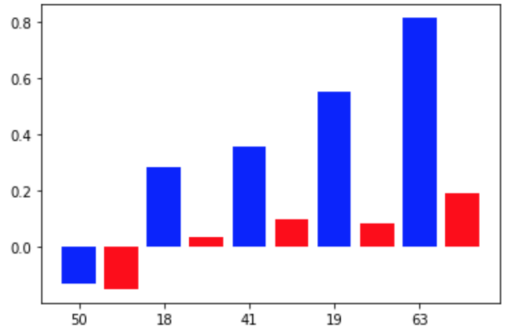
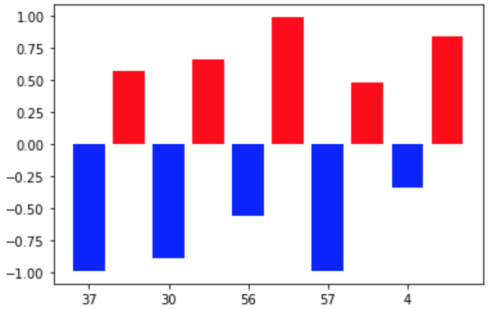


Figure 4: Top topics for Donald Trump (red).



We found that Biden's strongest areas were Topics 50 (which had key words "European", "London", "Brexit"), 18 ("Facebook", "World", "Open"), 41 ("Woman", "Mother", "Family), 19 ("Russia", "Putin", "Intelligence"), and 63 ("Apple", "iPhone", "App"), indicating a general strength in foreign policy, tech policy, and family values.

Meanwhile, Trump's strongest areas were Topics 37 ("Abortion", "Assault", "Allegation"), 30 ("Iran", "Saudi", "Nuclear"), 56 ("Survey", "Poll", "Voter"), 57 ("Field", "Game", "Score"), and 4 ("Election", "Vote", "Political"). This indicates a general strength in abortion, Middle East policy, sports, and elections.

Figure 5: Negativity (red), neutrality (grey), and positivity (green) likelihoods for Biden.

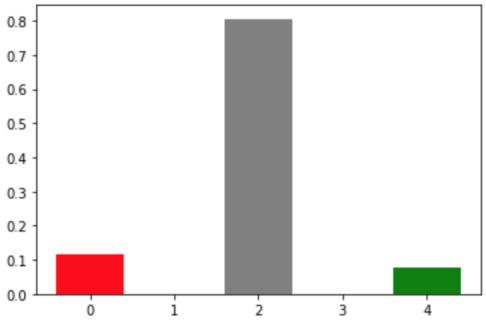
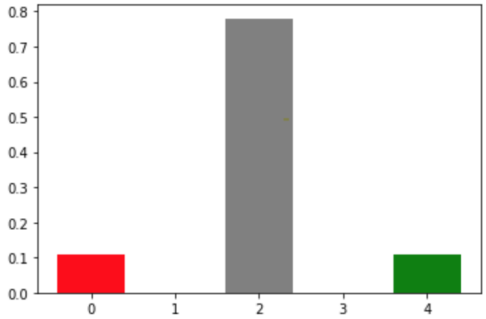


Figure 6: Negativity (red), neutrality (grey), and positivity (green) likelihoods for Trump.



We were also able to calculate the individual probability that an article describing Biden or Trump would be positive, neutral, or negative. For both Biden and Trump, news coverage was overwhelmingly neutral, although Trump had a slightly higher rate of positive articles.

5 Future Plans

5.1 Public Opinion Model

Due to time constraints, we were only able to consider media opinion, while the original scope of our project considered media opinion relative to public opinion. One direction we can take this project is by scraping discussion

boards such as Reddit using a wrapper package such as PRAW. For instance, subreddits such as r/political-discussion have a high level of activity and frequently reference both Trump and Biden in their posts. One thing to be careful about with this approach, of course, is that we find a representative sample of public opinion, as many subreddits tend to skew conservative or liberal.

5.2 Live Data

For the final project, we hope to have completed a live continuous web scraper running 24/7 on an AWS instance. This scraper will continuously download news articles from the Google News-curated topics of Joe Biden and Donald Trump at a minute resolution. Then, we will feed this article data into our sentimental analyzer, which will provide a sentiment score. Then, using our topic modeler, we will assign the article relevant topics. Take this (hypothetical) article headline "Donald Trump ruins US-Saudi relations, causing global oil collapse": this article will definitely have a negative sentiment, and the topic modeler will assign topics like "foreign relations", "energy markets", and "oil" to this article. Thus, we will then update Trump's sentiment on those three models, and downgrade his rating on all three of them due to the negative sentiment in the article.

With this stochastic data, we hope to create a time series for each topic related to each candidate, and from these time series, we hope to draw correlations to worldwide futures data and public polling data.

With the futures data, we might be able to predict market movements fast enough to act on them. Furthering the above example, such an event would most likely cause a dip in oil futures, but to be profitable, our model would need to predict such a movement before the market corrects itself. Thus, powerful computing services and lightning fast internet speeds such as those provided by AWS are a must need if we want our model to be profitable.

With the polling data, such an event would most likely cause public opinion of Trump to go down, since rising oil prices cause high gas prices (which no one likes) and other negative economic events. With time series data about polling, available from websites such as fivethirtyeight, we can find correlations between polling time series and sentiment time series. Thus,

when an event like a Saudi oil crisis occurs, our model should be able to predict the corresponding shift in public opinion.