Michael Caballero

Methodology Writeup

For my thesis, I am attempting to poll public opinion of presidential candidates using social media data. The first step in my project was deciding the social media platform to use in order to focus on then accessing and utilizing that platform's data. The most recent meta-analysis of social media data discusses the various methodologies employed in this field of research and the decisions that were made along the way, including the decision of social media data sources. This study found that gathering data from multiple sources was most effective, with blogs being the most accurate single source of data. Twitter was ranked the second most accurate source of data; due to the difficultly of collecting data from blogs, I chose to use data from Twitter for this project. I could include data from other platforms, like blogs or Facebook, but currently do not believe I will have the time to take those steps.

On a side note, going forward I will discuss multiple ways that I will be approaching this project. I believe that I can improve the methodologies discussed in this field but am not sure which improvements will be most effective. Thus, I plan on trying out multiple methods and finding out which ones are most effective.

For my data, I decided to gather my own dataset and use multiple datasets I found online. All data was gathered using Twitter's API. Because Twitter's API only provides limited access to historical data, it was difficult to gather my dataset for this project (I did not foresee this being a problem during the 2020 election as I was still trying to understand this field). The first dataset I found online is a 1.7 million tweet data set from Kaggle user Manch Hui which scraped all tweets from 10/15-11/8 that included #DonaldTrump, #Trump, #Biden, or #JoeBiden. The second dataset I found online is a more robust version of the first dataset -- it is a 20 million tweet data set collected from 7/1-11/11. This dataset, found on IEEE Dataport, is composed of tweets that were searched by using the names of the parties, their abbreviations, the names of the party candidates, and the election slogans. The third and most robust data set I found online is a repository from Cornell's Emily Chen that collected 868 million tweets from 5/20/2019. This repository collected tweets both from specific accounts and in real-time tweets that mention specific keywords. The accounts that this repository followed contain only politicians. The fourth dataset that I plan to use is one I collected. It is a dataset of the last 3200 tweets of the presidential candidates and their running mates, 20 news organizations, 20 political pundits, and 20 popular politicians. The accounts of news organizations, political pundits, and politicians were all balanced so that there were 10 right-leaning accounts and 10 left-leaning accounts. While these datasets may overlap, I also may use all of them in exploring which methodology is best so I thought I should mention each one. I also have scraped the RealClearPolitics average of polls from 9/31/2019 to the day of the election, which I will use as the dependent variable in my machine learning methods.

I currently am planning on mainly using Jupyter Notebooks, NumPy, and pandas, to process and model the data. Due to the large computational difficulty of working with datasets that are approaching a billion tweets, I may have to employ other tools but I am not sure if that will be necessary yet. I plan on documenting the process by uploading all of my data and Jupyter Notebooks to Github with each iteration I go through. Right now, I only see one area for introduced bias which I will discuss later.

Though I have only read part of the literature on this field of public opinion forecasts with social media data, there seem to be two main types of polling public opinion. These two main types are measuring the volume or sentiment of Tweets. A volume analysis is basically examining the relative frequency of the mentions of political parties (or canidates) in Twitter messages posted during course of the election (like this study). A sentiment analysis goes one step further to classify the sentiment of the tweet and then assign it as positive or negative towards one candidate (like this study). A few of these methods' implementations are performed using supervised machine learning where the model will be trained on some type of polling data. Most of the purely unsupervised models are not very accurate and even many of the supervised models do not claim to have reliable methods to forecast public opinion. Additionally, many of these machine learning methods are very simple..

The highest goal this project could achieve is creating a unsupervised machine learning method for polling public opinion of presidential candidates. While this goal is possible, I do see many large obstacles in the way of achieving it. Firstly, the demographics on Twitter do not accurately represent the demographics of voters in America. I believe this is why many unsupervised models are quite inaccurate with their prediction of public opinion or electoral outcomes. Second, the noise on Twitter contributes to inaccurate data and thus the model is inaccurate because the data is inaccurate. Overcoming this issue of noise is quite difficult with factors like a large presence of bots on Twitter's platform. Nonetheless, I believe I can contribute to this field of research without completing this large goal. As I do not see this method for polling public opinion replacing traditional polling methods anytime soon, I find no problem with performing supervised machine learning. Thus, I will be working on two separate methods to see which is more promising.

The first method I will work on is attempting to find the most accurate supervised or semi-supervised model. As much of the research in the field uses very basic machine learning, I am hoping to contribute by implementing a variety of machine learning methods from scikit-learn. I will be testing multiple machine learning methods using a variety of features from tweets to find the most accurate model. Using the data from RealClearPolitics as the baseline, I will see how accurate these models can get. I also am considering using deep learning with TensorFlow but am not as well-versed with deep learning. After doing purely supervised machine learning, I would like to attempt semi-supervised machine learning. I could potentially accomplish this by training my model on the polling and twitter data until a month before the election, then switching to only training my model on the twitter data. Overall, this method is designed to employ complex machine learning methods to the traditional analyses in the field, attempting to maximize accuracy with polling data. I believe this method is a step towards performing this analysis without polling data.

The second method is one which I believe would also be a new contribution to the field. From my analysis, most of the research in the field operates off measuring the effect of politics on Twitter. While I believe this is a viable method, there is another way to view Twitter data. In a sense, opinions, campaigns, and media outlets on Twitter help form public opinion -- they shape or "cause" the public's view of presidential candidates. Few papers actually implement a methodology that aligns with this view of measuring the response to those shaping public opinion. While some have done basic methods of this, like measuring candidates' likes on Facebook over the course of campaigning, I propose taking this method to the extreme. By measuring the popularity of left and right-leaning candidates, pundits, media outlets, and other politicians, it may be possible to measure the public's opinion of the presidential candidates. I believe my collected dataset enables me to conduct this type of analysis unlike the other datasets. I am hoping my equal set of conservative and liberal accounts will allow me to factor out the demographic bias on Twitter. Though I was mainly aiming for popular accounts, this method is biased by my own political beliefs as I picked out the 20 media outlets, politicians, and political pundits. If I was able to create this unsupervised machine learning model based on following the popularity of handpicked accounts and it was accurate, I believe it would be a large contribution to the field.