Michael Caballero
December 14, 2020
Data H195A
Final Prospectus

Polling Public Opinion with Social Media Data

I. Introduction

Over the course of this past semester, I have been conducting independent research for my data science thesis. I have refined my topic to polling public opinion of presidential candidates throughout the 2020 presidential election using social media data. Due to the difficulty of collecting social media data, I have chosen to simplify my project and only use Twitter data. My main question for this project is: can social media replicate the accuracy of national polls and properly forecast the popular vote of the election?

II. Motivation

This thesis idea stems from an interest that I have always had in politics and government. While it is unrelated to my domain emphasis of economics, it is related to my professional interests. I believe utilizing data to understand sentiment will prepare me for a career in tech where there is value in understanding user sentiment. Furthermore, this research allows me to combine my data science skillset with my passion and informal knowledge of politics.

This research is quite relevant to the space of political science. From talking to my likely thesis advisor, David Broockman, I have learned a significant amount about the importance of this topic. A robust and repeatable methodology for polling public opinion of campaigning politicians would innovate polling and campaign strategy. A method that had the accuracy to replace traditional polling methods would be extremely important. In some races, like those for state senators or governors, it would be the only reliable method for polling support as those races traditionally do not have much polling data. And this method could even be important for races with extensive polling data. Traditional polling methods are often expensive and time-intensive so a cheap, efficient model that relies on social media data would provide valuable complementary data. In sum, Twitter and other social media platforms can provide a huge sample size of data that could allow cheaper polling in real-time.

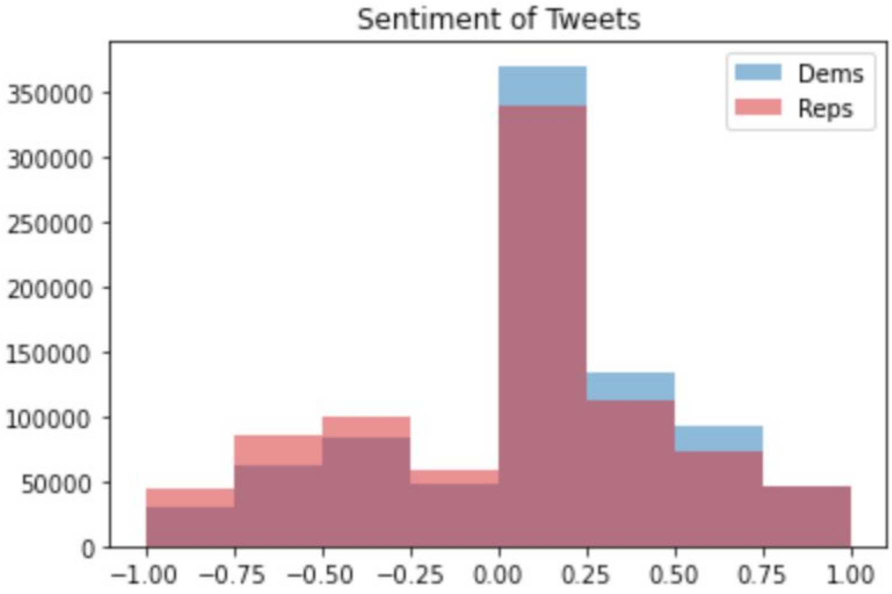
While the specific case of electoral prediction using social media data is interesting and would be an important contribution to political science, I believe this research is also important for all social sciences. This type of prediction foreshadows a transition in social sciences as these disciplines are undergoing a shift from data scarcity to abundance. As more interactions happen on digital platforms, the capacity of social scientists to predict the attitudes and behaviors of society should grow. Thus, this field of electoral forecasting using social media data is a representation of a larger transition in the social sciences due to growing data availability.

III. Data

During this project, one of the biggest challenges I have overcome is the scarcity of social media data. Though I wanted to perform prediction using multiple platforms, the most effective method in academic research, I chose to only use Twitter as it was the easiest platform to collect data from (Skoric et al. 10). Even though Twitter has the most accessible data, it still provides its own challenges and obstacles. Most free Twitter data is only accessible as historical tweets from specific accounts or tweets collected in real-time while searching for keywords, hashtags, phrases, or account mentions. Unfortunately, I did not plan for this impediment during the 2020 election as I was still trying to understand this field of research. But, I still was able to collect important data. The data I use in my research will consist of four datasets of Twitter data; three I found online, and the last I collected. All data was gathered using Twitter's API. As Twitter owns the data, I cannot post most of it online -- under Twitter's Terms of Service for developers I am only able to publicly post the ids of the tweets.

My first dataset 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 more robust -- it is a 20 million tweet data set collected from 7/1-11/11 (Sabuncu). This dataset, found on IEEE Dataport, is composed of tweets that were searched by using party names, their abbreviations, candidates' names, and election slogans (see Appendix A). 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 to weeks after the election. This is the only dataset I found online which collected tweets that mentioned specific keywords and tweets from specific accounts (see Appendix B).

This last dataset from Emily Chen inspired my self-collected dataset as the accounts that this repository followed contained almost solely politicians. For my own dataset, I thought it would be useful to follow multiple types of accounts. Related, one problem that previous researchers have encountered is that the demographics of Twitter do not represent the demographics of the voting population. From exploratory data analysis on the IEEE dataset, I was able to confirm that the data was biased as such.



The histogram above represents the sentiment of one million tweets randomly sampled from the IEEE dataset. The tweets that contain Republican keywords in the sample are more likely to be negative while the tweets that contain Demographic keywords are more likely to be positive. This portrays how the demographics of Twitter do not match the voting population -- Twitter users are more likely to be left-leaning and thus Democratic. Therefore, I wanted my collected dataset to also address this systematic bias of Twitter data.

I collected the fourth Twitter dataset by scraping historical tweets from multiple prominent accounts. I wrote a python script that utilized the Twitter API to scrape the most recent tweets of 57 accounts. (This script is in my public GitHub repository and linked here.) This dataset consists 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 are all balanced so that there were 10 right-leaning accounts and 10 left-leaning accounts.

Hopefully, in measuring the proportional features of Tweets from these accounts, I can factor out some of the innate bias in Twitter data. While the datasets do overlap somewhat, all provide unique data and I will use multiple datasets in my final model.

In addition to Twitter data, I also collected polling data of the two major candidates. With this polling data, I can perform supervised learning to train my model. To find accurate polling data, I compared popular polling aggregators that provide robust estimates of public opinion. I then scraped the RealClearPolitics average of polls from 9/31/2019 to the day of the election (realclearpolitics.com). I believe RealClearPolitics's polling aggregation will be sufficient but may collect data from other polling aggregators as well.

IV. Methodology

For academic research regarding predicting public opinion with social media data, there are two main types of analyses. These analyses look at the sentiment or structure of social media data, sometimes analyzing both together (Gayo-Avello 654-59). Sentiment approaches are based on extracting the preference of user's social media posts, while structural approaches represent the social structure of conversations on the platform. The most basic type of approach is a volumetric analysis where researchers predict outcomes based on the number of party or candidate mentions on a platform. While this is a simplistic method, it has been shown that number of tweets can correlate with electoral performance. One specific study showed that this correlation in Congressional elections in 2010 and 2012 (DiGrazia et al.). This basic volumetric approach was one of the first preferred methods by researchers until sentiment analyses were utilized. Sentiment approaches generally collect posts, score the preference using lexicon or machine learning-based models, and then predict considering the number of positive and negative tweets regarding candidates and parties. This sentiment analysis is the most popular method of electoral forecasting with social media data and has even found success in predicting the 2012 US presidential election using Twitter data (Choy et al.).

While there are studies that showed that social media forecasting using these volumetric or sentiment analysis methods had high accuracy, there are also many studies that show these methods are not reliable. Thus, structural approaches have become more complex by looking at other metrics besides total volume of related posts.

One study found a positive relationship between electoral outcomes in the 2015 Finnish parliamentarian elections and Facebook likes on candidates' official Facebook pages (Vepsäläinen et al.). Studies have progressed including all of these elements together to generate more accurate and predictive results. One study of the 2020 US presidential election utilized the sentiment of tweets, the number of retweets, and the number of people who posted these tweets (Sabuncu et al.). In my research, I will try to add important features to the structural analysis while still utilizing the traditional methods that have found some success.

In this thesis, I propose using supervised machine learning methods to predict the public opinion of both candidates. I will be analyzing the daily features of both the general conversation on Twitter and the response to prominent left or right-leaning accounts. Likely using Emily Chen's dataset, I will look at the daily mentions of a candidate or party, the sentiment of those mentions, and the response to those mentions -- re-tweets or likes. Using my collected dataset, I will also be analyzing the daily volume, response, and change to these metrics for tweets from prominent accounts grouped by determined political leaning. I plan on performing regression supervised machine learning using python package scikit-learn in Jupyter Notebook to see the accuracy of linear regression, regression forests, and regression neural networks. The label for each day will be the polling values from RealClearPolitics. I foresee a problem with the accuracy of Donald Trump's polling data, as he outperformed the polls by ~3% points in actual vote share, but am still working to find a solution to this problem. I am planning on only training the algorithm until a specified time before the election -- a few weeks to a month -- so that I can compare the algorithm's results to the polling data during that unsupervised time.

My research would contribute to the field of polling public opinion using social media data in three main ways. First, I would be incorporating many structural features of tweets that have not appeared in research to the best of my knowledge (Bilal et al.). The features like the ratio of retweets or likes to followers may be very useful in prediction. I also am following major accounts' popularity over the course of the election. While some studies have made similar efforts, I believe I go further as I will analyze the response to candidates, other politicians, political pundits, and media outlets. Lastly, I believe I will contribute to this field as I have not come across a study that performs supervised learning over the course of the election using polling data. I think this allows my model to be much more accurate than other methods. As I do not see these new methods for polling public opinion replacing traditional polling methods anytime soon, I find no problem with performing this supervised machine learning.

V. Ethical Issues & Outcomes

As of now, I do not see any prominent ethical issues with my research. There is a small possibility that developing a better method of polling could have negative impacts on groups without a fair share of political power. By developing better methods of polling, like with social media data, those who benefit could be those in power who would utilize these methods first. It would give the powerful an unfair advantage in influencing democracy thus having negative repercussions on those with less power in the political sphere like those with lower socioeconomic status and racial or ethnic minorities. I find this ethical concern to be small because this is only one possibility of developing better polling methods. This development could also be beneficial to those without power by evening the playing field of politics -- making accurate polling data cheaper and thus more accessible to all.

Regarding outcomes, the best possible result from this research would be an unsupervised method that predicts presidential vote share accurately and works out-of-sample. This will be hard to verify as I do not have data from previous presidential elections and will have to wait until 2024 for another. A reasonable outcome would be a model that accurately follows polling data and predicts vote share using these new feature engineering methods and only some polling data. This model would display how social media and polling data can work together to develop accurate forecasts and the importance of using a variety of structural features when analyzing tweets. Generalizing this methodology so it could be used in other political races would greatly add to the effectiveness of my research. But, my baseline outcome is to develop a completely supervised model that can accurately predict vote share using Twitter data.

Appendix A

IEEE Dataport’s dataset consists of tweets that were identified and collected because they contained these keywords, phrases, or hashtags: “#USAelection”; “#NovemberElection”; “@DNC”; “@TheDemocrats”; “Biden”; “@JoeBiden”; "Our best days still lie ahead"; "No Malarkey!"; “#MAGA2020”; “@GOP”; “Trump”; “@POTUS”, “@realDonaldTrump”; “Pence”, “@Mike_Pence”, “@VP”; "Keep America Great".

Appendix B

Emily Chen’s dataset followed accounts and specific keywords over the course of the election. The accounts followed are linked here and the keywords followed are linked here, both on the dataset’s public GitHub. Please see the repository’s README.md for more information on the dataset.

Works Cited

Bilal M., A. Gani, M. Marjani and N. Malik, "Predicting Elections: Social Media Data and Techniques," 2019 International Conference on Engineering and Emerging Technologies (ICEET), Lahore, Pakistan, 2019, pp. 1-6, doi: 10.1109/CEET1.2019.8711854.

Chen, Emily. *2020 US Presidential Election Tweet IDs*. GitHub, 29 Nov 2020. Web. 29 Nov 2020. <https://github.com/echen102/us-pres-elections-2020>

Choy, M., Cheong, M., Laik, M. N., & Shung, K. P. (2012). *US presidential election 2012 prediction using census corrected twitter model*. arXiv preprint arXiv:1211.0938.

DiGrazia J, McKelvey K, Bollen J, Rojas F (2013) *More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior*. PLoS ONE 8(11): e79449. https://doi.org/10.1371/journal.pone.0079449

Gayo-Avello, Daniel. “A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data.” Social Science Computer Review, vol. 31, no. 6, Dec. 2013, pp. 649–679, doi:10.1177/0894439313493979.

General Election: Trump vs. Biden. RealClearPolitics, 3 Nov 2020. Web. 29 Nov 2020. <https://www.realclearpolitics.com/epolls/2020/president/us/general_election_trump_vs_biden-6247.html>

Hui, Manch. *US Election 2020 Tweets*. (Version 19). Kaggle, 8 Nov 2020. Web. 29 Nov 2020. <https://www.kaggle.com/manchunhui/us-election-2020-tweets>

Sabuncu, Ibrahim. *USA NOV.2020 ELECTION 20 MIL. TWEETS (WITH SENTIMENT AND PARTY NAME LABELS) DATASET*. IEEE Dataport, 16 Nov 2020. Web. 29 Nov 2020. <https://ieee-dataport.org/open-access/usa-nov2020-election-20-mil-tweets-sentiment-and-party-name-labels-dataset#files>

Sabuncu, Ibrahim & balcı, Mehmet & Akgüller, Ömer. (2020). *Prediction of USA November 2020 Election Results Using Multifactor Twitter Data Analysis Method*.

Skoric, Marko M., Jing Liu, and Kokil Jaidka. “Electoral and Public Opinion Forecasts with Social Media Data: A Meta-Analysis.” *Information* 11.4 (2020): 187. *Crossref*. Web.

Vepsäläinen, T.; Li, H.; Suomi, R. *Facebook likes and public opinion: Predicting the 2015 Finnish parliamentary elections*. Gov. Inf. Q. 2017, 34, 524–532.