Michael Caballero
Data H195A
September 30, 2020

Presidential Electoral Forecasting with Social Media Data

Because it is an election year and there is a vast amount of political data created currently, my idea for this data science honors program revolves around electoral prediction using social media data. I want to analyze social media data from this electoral cycle and develop an algorithm that uses that data to poll the popularity of presidential candidates Joe Biden and Donald Trump. While the specific case of electoral prediction using social media data is interesting and could introduce new methods to polling and campaign strategy in the future, I believe this research is even more important than one may initially think. This type of prediction foreshadows a transition in the social sciences as these disciplines are undergoing a shift from data scarcity to data 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 from growing data availability.

This project stems from a deep obsession with politics and government that I have had from a young age. While it is mostly unrelated to my domain emphasis of economics, it is related to my professional interests, namely the ability to use social media data to understand users (a skill that is useful in the field I am interested in, product management). As of now, there is no linkage to human contexts and ethics but as I move forward, questions may arise about the ethical concerns of using social media data to forecasts elections.

The main question I would like to address in this research is: can social media replicate the accuracy of national election polls and forecast the popular vote ratio of the presidential election? Under the umbrella of electoral prediction using social media data, there are a number of factors that determine the scope and direction of the research including the baseline for presidential opinions, social media platforms used, techniques used as predictors, and how success is measured. As far as the baseline for

public opinion regarding presidential candidates so that I can train and test my model, I will be using the RCP average of the national general election polls from RealClearPolitics.com. But, the final and most important test of my algorithm will be its ability to predict the popular vote of the presidential election. Regarding the social media platforms I will use for prediction, I believe I will start by using Twitter data in my algorithm. A meta-analysis on this subject found that using multiple social media platforms resulted in the highest predictive power, but using multiple platforms would be a reach goal for this project. While the top four platforms ranked by predictive power were blogs, Twitter, forums, and Facebook, respectively, I will begin my research by using Twitter data as it is much easier to access, collect, and process than blog data. Regarding the techniques used, the same meta-analysis found that sentiment and structural approaches had the highest predictive power when combined. 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. Though I am still figuring out the exact methods I will use, I believe combining sentiment and structural approaches will lead to the best algorithm. Lastly, a large point of contention in the scientific community working on this topic is how the success of results should be reported. I will be measuring the accuracy of my predictor using MAE (Mean Average Error) and R squared.

My starting dataset will consist of tweets pulled from Twitter using the Twitter API, and the daily change of the RCP average for national general election polls. While I do not have familiarity with the Twitter data, I will be doing EDA to find out what are the characteristics of tweets best for sentiment analysis (like the text of the tweet) and what characteristics are best for structural analysis (like the number of retweets or the number of followers the user has). I am still developing an exactly timeline for when and how I will collect the data, but hopefully will be able to collect data from after the DNC until the election. Moving forward, I believe the highest priority will be to figure out exactly how I am collecting data, then find a professor I would like to work with, and tertiary, figure out which methods are best for analyzing the Twitter data.