ANLY 580: Project Proposal

Topic Modeling and Feature Engineering for Informed Political Understandings

Monroe Farris, Haley Roberts, Sam Pastoriza

Project Description

Our project aims to explore and better understand short-term political predictions by utilizing open-source social media content and current Natural Language Processing (NLP) techniques. The specific use case for this project is the Georgia Gubernatorial election which occurred on November 8th, 2022, uniquely positioning our effort to utilize the actual results of the election for understanding our results.

We look to leverage volumetric analytics, feature engineering, sentiment analysis, and topic modeling to help us better understand if there is a way to create a short term (i.e. less than a month) political prediction of a winner in a particular race using only social media data.

The corpus for the study is built from Twitter to draw insights from sources closer to the voters and individuals participating in the election. Additionally, the short time span of this analysis is an approach not commonly seen with political prediction models. This in conjunction with NLP techniques such as sentiment analysis and topic modeling provides a unique avenue for this study.

Proposed Methodology

Modeling Approach

Given the nature of our topic and intended modeling goal - a predictive political model based on social media data - our approach is primarily investigative in nature to determine the validity of making predictions using social media data. Therefore, we will approach this analysis with a three-pronged methodology consisting of volumetric analytics, sentiment analysis and feature engineering, and topic modeling. These three key areas will allow us to determine the ability of social media data to be used in a predictive manner.

We will first get a volumetric understanding of the data. This will baseline the audience and provide familiarity with the distribution of the data, the amount of Tweets, and which candidate is being talked about the most. Keywords associated with each candidate will categorize each tweet as discussing a particular candidate, both candidates, or the election in general. The resulting dataset will be summarized and the most number of tweets associated with each candidate used as a potential indicator of a candidate's success. Mentions of the candidates over time will also be considered as an avenue of investigation into a candidate's success or failure. For example, perhaps a candidate's increased rate of Twitter mentions is a strong indicator of the outcome.

Then, we will perform sentiment analysis on each Tweet, allowing us to understand the tone that the Tweet is conveying. Two approaches to sentiment analysis will be considered: 1) positive versus negative sentiment found using VADER, a lexicon and rule based sentiment model that performs best on social media data, and 2) an exploration of adding more complex emotional layers to model sentiments such as aggression or joy. This will provide insights into the manner in which each candidate is discussed, and allow for a better composite understanding of the overall tone of the election. For example, one would expect that a candidate who is discussed more positively should be in a better position to win the election. In this analysis, we will investigate if polarizing sentiments provide a good indication of a candidate's success or failure. To bolster the investigation, we will feature engineer the extracted sentiment with the time of the post to better gauge the standing of each candidate going into the election.

Finally, we will utilize topic modeling to provide tailored political insights. Using topic modeling, we will be able to determine the common topics users mention in their tweets. By gathering enough of these topics, we should be able to filter tweets based on these topics and analyze the characteristics of the tweets containing each topic. Theoretically, the informed voter who tweets their preferences have common reasons that can be extracted through topic modeling. Therefore, honing in on the topics that voters are interested in may aid in understanding a candidate's chance of success by considering their stance on each topic indicated.

Combining the approaches of volumetric analysis, sentiment analysis, feature engineering, and topic modeling will provide us with an understanding as to if there is enough signal in the social media data to create an informed prediction of election results. If this is the case, then we intend on combining the pillars of our methodology in a meaningful way to best create probabilistic predictions of election outcomes.

Data

Our data consists of tweets collected from Twitter via the Python Tweepy API. Tweets were collected from October 18, 2022 through the end of election day, November 8, 2022. This collection over 22 days produced a dataset of 56,908 tweets. For each tweet, metadata was collected about the tweet's creation time, author, retweet count, favorites count, and more.

To get tweets specific to the Georgia gubernatorial election, we utilized 20 key filter terms relevant to the two candidates, Brian Kemp and Stacey Abrams, as well as the Georgia Governor's election in general. These filters, when queried via the Tweepy API returned relevant tweets that contained at least one of the keyword strings provided. For example, if using the filter term, “GA election”, tweets with that exact string would be retrieved.

Finally, in addition to the Twitter data, our project is in the unique position to know the results of the election and use this knowledge as a litmus test for how well the different approaches of our analyses perform and aid in developing a thorough understanding of which features produce high signal and predictive power.

Evaluation

Our project is unique in regard to evaluating success in that it is highly exploratory in nature. We intend to investigate the three means of understanding the data explained above with the goal of finding optimal features for predicting election outcomes. These features will be evaluated to understand if they provide a good or strong signal toward this prediction goal. Depending on the feature, this will be done in different ways. Ultimately, we hope to consider the results of these feature investigations independently as well as cumulatively to produce an understanding of which features (textual, sentiment, volumetric, topic, etc.) are best used in predicting an election outcome. However, we are not explicitly building a prediction model. Therefore, typical model evaluation criteria such as accuracy, recall, precision, etc. will not be considered. Instead, we will look at the distribution of true votes accrued from the election (example: 54% candidate A, 46% candidate B) and utilize these percentage points as endpoint goals. We will work backward from this point to explore our features to discover the best inclusion or exclusion of engineered features to reach the true results.

Considering this, we will evaluate our system of exploration, feature engineering, and topic modeling to better understand how social media data such as tweets can be broken down and analyzed to provide information toward understanding election results. Therefore, while no hard accuracy will be evaluated, the success of the project will be measured in insights gained toward social media’s use in understanding the election process.

Hardware

Our project will not require any special computational or hardware resources. With a corpus of approximately 57,000 tweets (and likely less after cleaning), our local computers should be able to handle the processing. If greater processing power is needed, platforms such as Google Collab or another compute-providing resource are available and will be utilized.

Unknowns

Two major unknowns exist for this project based on the nature of the data: 1) quantity and quality of data retrieved, and 2) unknowns related to the validity of using this type of approach for political predictions.

First, we decided to use social media data for this project as it is broadcast to be a close representation of human thoughts and sentiments. However, there are limitations in regard to both the quality and the quantity of data that is collected through this medium. In terms of quantity, Twitter’s API only allows for tweets to be collected from a maximum of 7 days prior to the collection time. Therefore, since we started this project in late October, tweets were only collected for approximately three weeks before the election. This will not allow for political opinions or sentiments from earlier than three weeks prior to the election to be captured or analyzed. Also, there is major uncertainty over the quality of Twitter data in terms of the truthfulness of posts, the propaganda and campaigning nature of posts, and the presence of spam. Considering all of this, it is understood that there are unknowns in regard to the data that will be fleshed out during the cleaning process and through the end of the project.

Second, there are unknowns in regard to the amount of “signal” toward the prediction of a winning candidate that may exist in the open-source Twitter data. We are trying to translate the act of a human voting to features extracted from social media text, which is inherently hard and contains many uncertainties. Because of this uncertainty of social media representation of voter response, we chose to alter the original idea of predicting political race outcome solely from social media text to pivot toward a focus on exploring and understanding how to work with this data to potentially create a type of political prediction. This is why we will consider the three-pronged approach in the analysis portion: to investigate optimal processing with this type of data and report on the experience and results.

In sum, there is much uncertainty resulting from the nature of the data being analyzed and the goal of the analysis. However, we plan to pivot and adjust our study as our explorations provide guidance toward the best way to clean, understand, and optimally utilize this social media data for the purposes of our investigation.

Presentation of Results

To best convey our modeling results, we will conduct an oral presentation accompanied by a slide deck that will walk through the basis of the project, the modeling approach, portions of code methodology, and significant findings.