`War on "Fact Check" -- the Path to Magic 270

Rodrigue Rizk, Dominick Rizk, Vijay Srinivas Tida, and Sonya Hsu

Introduction

Twitter has been the social media platform because the President of the United States likes to share his opinions on many ongoing affairs. The President's Tweets have been "fact checked," deleted, rebuckled, supported and coutertweeted, the flow of information has been phenomenal, especially, in this very historical election 2020. Up to the morning of November 5th, the Electoral votes are still being counted as we finish this abstract, i.e., Pennsylvania, Georgia, North Carolina, Nevada, and Arizona.

	Electoral Votes	Total Vote in (%)	Biden (%)	Trump (%)
Pennsylvania	20	88	50.3	48.6
Georgia	16	99	49.5	49.2
North Carolina	15	94	48.7	50.1
Nevada	6	76	49.4	48.5
Arizona	11	88	50.5	48.1

Twitter data are collected from October 15th to November 2nd 2020 to see if there are comparable reflections to the votes in 50 States, mainly battleground states, with the Tweets regarding both Republican and Democratic parties. Tweets are most likely related to COVID-19, economy recovery, unemployment, mask-wearing, black lives mattered, racial divide, voting fraud, mail-in ballots and etc. Specifically, this project looks at the locations where the Tweets initiate, and retweets. Deep dive into the content where the origin of sources are, for example from the bi-parties, news sources, then, positive or negative of content. From the Tweets of the election 2020, the researchers intend to discover the influences of the social media platform -- Twitter, in relations of the election outcomes. The below research questions are derived from the previous studies as well as the trends of the uses of social media.

Research questions:

- Detect any correlation between the sentiment of users on Twitter and the eventual election results.

- Detect potential election manipulation.
- Predict the winner candidate from tweets.

Methodology

Many previous works leverage sentiment analysis techniques on social media for predicting election results. However, these works failed to account for the importance of the geolocation of Tweets, further possible influencing the election outcomes. In nearly every state, either candidate who gets the most general votes wins the “electoral votes” for that state. Within 50 states, the magic 270 is the path to the presidency. In this research project, we present a geolocation based multiclass classifier for sentiment analysis to make the prediction model and then, compare to the outcomes of 2020 US presidential elections.

741,000 tweets were collected from October 15th to November 2nd, we first filtered and classified based on states and sentiments. Then, we determine the polarity and subjectivity metrics which help in classifying users’ stance towards each candidate.

One of the main challenges in applying sentiment analysis to social media content is its dynamic change. To mitigate this problem, we propose a geolocation-based hybrid model that consists of a Multinomial Naive Bayes machine learning classifier and a sentiment lexicon for rule-based sentiment. The intuition behind this approach is to account for the multi-stance characteristics of tweets that are dual positive-negative paradoxes. For example, the same tweet on a topic can be negative for one candidate while positive for another. To solve this negative-positive Tweets, special tokens are generated to be biased more towards the subject than positive-negative bases.

Our model consists of four modules: acquisition, preprocessing (cleaning and vectorization/ tokenization), processing, and visualization. The model will be trained as a joint dataset of the collected tweets on both presidential candidates. Then, we will train our model on each candidate dataset separately. Below is the proposed research model:

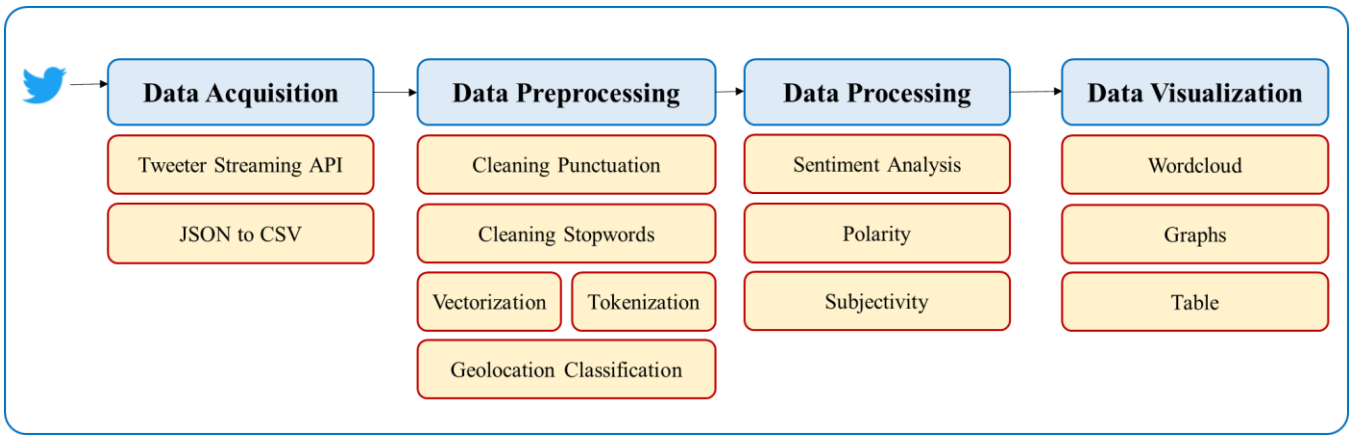

Fig. 1. Proposed Framework

Preliminary results

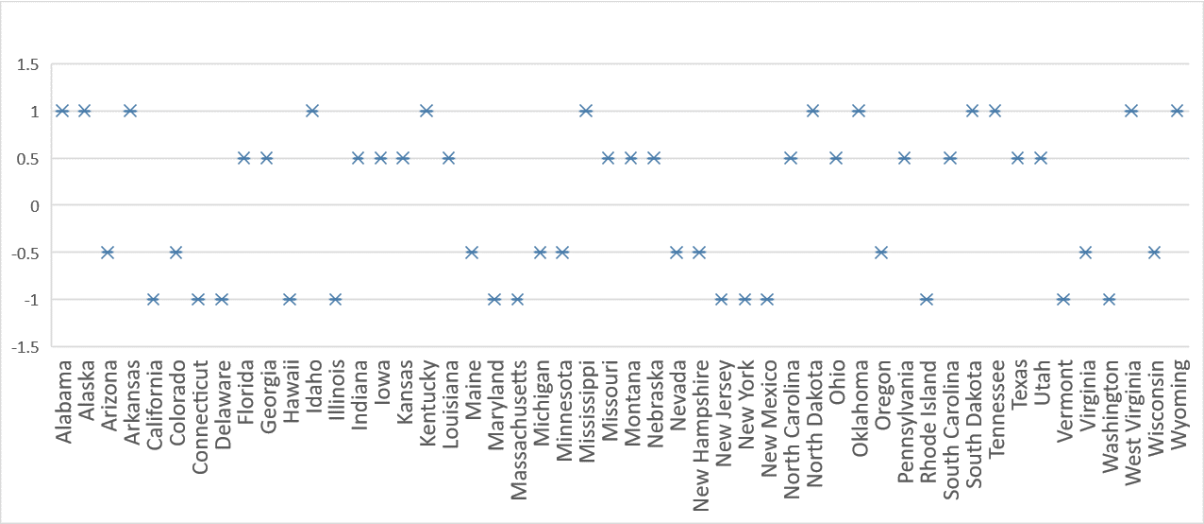

Fig. 2. State-based predicted results where 1: Strongly Republican, 0.5: Somewhat Republican, -0.5: Somewhat Democratic, and -1: Strongly Democratic