US Politics by Social Media

Project Proposal

Rachel Ng Min Yee
University of Colorado, Boulder
Rachel.Ng@colorado.edu

Xinyi Lu
University of Colorado, Boulder
xilu2783@colorado.edu

Anuragini Sinha
University of Colorado, Boulder
ansi3987@colorado.edu

1 Introduction

Social media platforms are widely used by citizens during the US election period to express their opinions. This project examines the sentiments of Twitter users towards the different political parties and how Tweets can affect the outcome of the US presidential elections. With the use of natural language processing techniques such as keyword extraction, text classification, topic modeling and opinion analysis, we aim to extract key information from social media data and study the political behavior of Twitter users. This will be helpful for election candidates to assess their odds of success based on previous case studies and to help them tailor their campaigns based on the sentiments on the ground.

2 Related Work

Multiple projects have explored the sentiments of tweets related to the US election. Ali et al. proved that removed tweets posted after the 2020 US Election Day sided with Joe Biden while those before Election Day were more favorable about Donald Trump. This was done by performing sentiment analysis on tweets that were subsequently removed from Twitter and comparing their results to accessible tweets and accounts across time [1]. In another research project, Caetano et al. discovered that political homophily level rises when there are close connections and similar speeches. This was accomplished by analyzing the tweets of users during the 2016 US elections and classifying sentiments into six groups: Trump supporter, Hillary supporter, whatever, positive, neutral and negative [2].

Finally, Ansari et al. found out the dominance of support for a single party on Twitter for the 2019 General Elections of India by using the Long Short Term Memory (LSTM) classification model to predict the sentiments and results of the elections. Ansari et al. also compares the results from the LSTM model with other machine learning models [3].

Our approach is different from past projects as we aim to perform opinion analysis using natural language processing techniques to answer key questions. Some questions we can possibly answer are: which groups of people have these opinions? How do opinions change over time? Subsequently, with insights from the data, we can find out the possible factors that led to Trump's large voter share despite the controversies surrounding him in the 2020 US Presidential Election.

3 Proposed Work

3.1 Dataset

The dataset we chose to work on is a collection of tweets containing #DonaldTrump, #Trump, #JoeBiden, #Biden hashtags from the 2020 US presidential election [4]. There are two subcollections in the dataset - one for the tweets related to Donald Trump and one for Joe Biden. With 970,919 and 776,886 rows respectively, there is a total of approximately 1.72 million rows of data in our dataset. Of all the tweets, 394,400 or roughly 20% of them were tweets from the US.

The data set has a total of 21 columns and a Data Dictionary describing the names and definitions of the columns of the dataset can be found in the appendix. The specific columns that we found interesting were columns created_at, tweet, likes, reteweet_count, user_followers_count, user_location, lat, long, city, country, continent, and state_code.

3.2 Subtasks

With the above dataset, we hope to perform text classification, opinion analysis as well as a case study on Donald Trump and his relatively successful voter share despite his abrasive personality and the serious controversies surrounding him. These analyses can be further split into the following subtasks:

Analysis	Subtasks
Text Classification	<u>Name Entity Recognition</u> • Which parties are involved in this tweet? • Which candidates are involved in this tweet? <u>Keyword Extraction</u> • Which keywords are the most important/relevant in the tweet? <u>Sentiment Analysis</u> • Is the tweet a positive, negative or neutral one?

	• What identity is the tweet supporting / criticizing? Topic Modeling • Grouping tweets that share common topics • Grouping tweets that share the same sentiment • Which topics were most discussed? • What were the topics that supporters of each party cared the most about? • Which party's supporters were more vocal about their opinions? • Which party's supporters generally had the bigger following on Twitter?
Opinion Analysis	Temporal Analysis • What are the predominant opinions over time? Categorization of Opinions according to: • State • Country • In the US vs outside of the US • Democratic vs Republican
Case Study on Donald Trump	• How did opinions change over time? • Did the timeline coincide with certain events? • How did Twitter specifically help him/prevent him from swinging favor? • What factors helped him garner his large voter share?

We believe the work to be sufficient and feasible for our group size of 3 at the undergraduate level as we are approaching the topic of the 2020 US presidential election from multiple different angles. If time permits, we could possibly look into adding other datasets and comparing the 2016 election to the 2020 election that we are currently analyzing.

4 Evaluation

Once we have completed our analysis for the project, we plan on evaluating the results by using the following methods:

- For the correlations we find, we plan to further validate them against Tweets made during the 2016 US presidential elections, if possible.
- We will calculate the correlation coefficient for the correlations to measure their strength.
- If a strong correlation is found, then we will create a correlation model with the attributes that are found to be useful from our research.
- We also plan to analyze baseline performances.

5 Milestones

Task	Start Date	End Date
Data Preprocessing Text mining, natural language processing (text classification, named entity recognition, topic modeling)	10/6	11/2
Data Analysis Opinion analysis (categorization, temporal)	11/3	11/23
Data Visualization Charts (bar plot, line charts, scatter mapbox, choropleth map)	11/24	11/30
Evaluation Evaluation of metrics, documentation (report writing, presentation slides)	12/1	12/8

REFERENCES

[1] Ali, R. H., Pinto, G., Lawrie, E., & Linstead, E. J. (2022). A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-022-00633-z

[2] Caetano, J. A., Lima, H. S., Santos, M. F., & Marques-Neto, H. T. (2018). Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election. Journal of Internet Services and Applications, 9(1). https://doi.org/10.1186/s13174-018-0089-0

[3] Ansari, M. Z., Aziz, M. B., Siddiqui, M. O., Mehra, H., & Singh, K. P. (2020). Analysis of political sentiment orientations on Twitter. Procedia Computer Science, 167, 1821–1828. https://doi.org/10.1016/j.procs.2020.03.201

[4] Hui, M. (2020, November 9). *US election 2020 tweets*. Kaggle. Retrieved October 6, 2022, from https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

APPENDIX

Data Dictionary

- created_at: Date and time of tweet creation
- tweet_id: Unique ID of the tweet
- tweet: Full tweet text
- likes: Number of likes
- retweet_count: Number of retweets
- source: Utility used to post tweet
- user_id: User ID of tweet creator
- user_name: Username of tweet creator
- user_screen_name: Screen name of tweet creator
- user_description: Description of self by tweet creator
- user_join_date: Join date of tweet creator
- user_followers_count: Followers count on tweet creator
- user_location: Location given on tweet creator's profile
- lat: Latitude parsed from user_location
- long: Longitude parsed from user_location
- city: City parsed from user_location
- country: Country parsed from user_location
- state: State parsed from user_location
- state_code: State code parsed from user_location
- collected_at: Date and time tweet data was mined from twitter