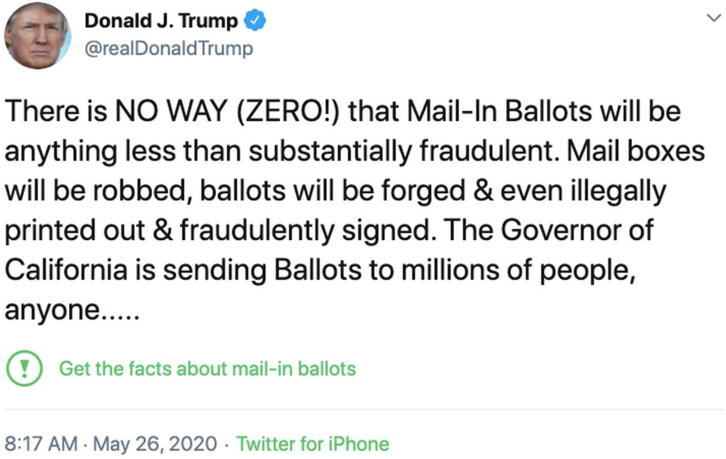
US Politics by Social Media

Team members: Rachel Ng Min Yee (CSCI 4502), Xinyi Lu (CSCI 4502), Anuragini Sinha (CSCI 4502)

Introduction - Background



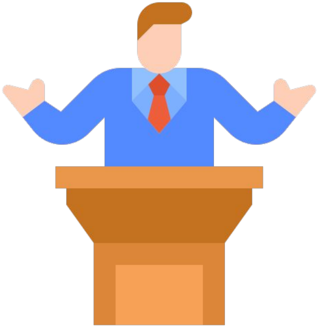
Donald J. Trump
@realDonaldTrump

There is NO WAY (ZERO!) that Mail-In Ballots will be anything less than substantially fraudulent. Mail boxes will be robbed, ballots will be forged & even illegally printed out & fraudulently signed. The Governor of California is sending Ballots to millions of people, anyone.....

Get the facts about mail-in ballots

8:17 AM · May 26, 2020 · Twitter for iPhone

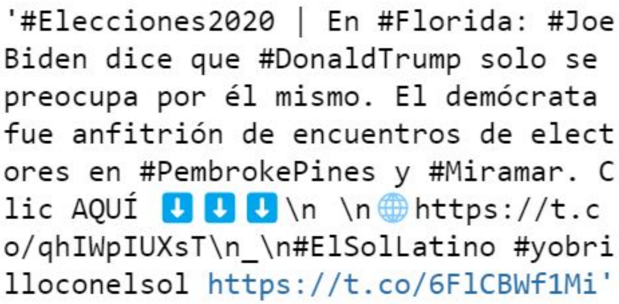
Introduction - Why is this interesting?



- Election candidates can predict their own and their competitors' odds of success.
- Election parties can tailor their campaigns to the sentiments of the people on the ground.

- Voters can be critically aware of the influence of social media in politics and make more objective decisions.
- Voters can better leverage on social media to push forward their political ideologies/causes in general.

Introduction - Challenges of our project



```
'#Elecciones2020 | En #Florida: #Joe
Biden dice que #DonaldTrump solo se
preocupa por él mismo. El demócrata
fue anfitrión de encuentros de elect
ores en #PembrokePines y #Miramar. C
lic AQUÍ ⬇⬇⬇\n \n🌐https://t.c
o/qhIWpIUXsT\n_\n#ElSolLatino #yobri
lloconelsol https://t.co/6FlCBWf1Mi'
```

1. Tweets in different languages

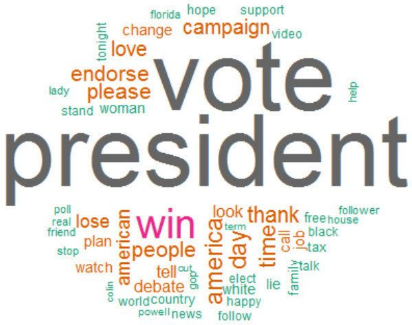
2. Large Dataset

- Original: 1 million rows
- Need to find a data reduction method or work on a subset of data
- Computational intensive code

3. Evaluation

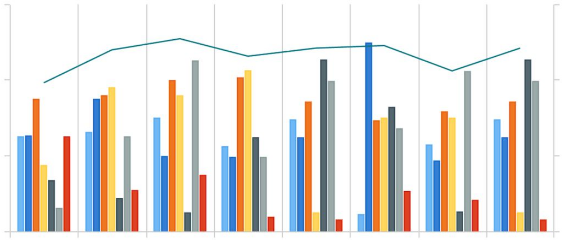
- Deriving common metrics to compare across different elections (ie. 2016 US Presidential Election vs 2020 US Presidential Election)

Introduction - Possible contributions/new insights



Find out how the opinions of Twitter users change across the campaign period

Real-time monitoring of key topics discussed by Twitter users in the next US presidential election



Find out which groups of Twitter users have certain opinions. Candidates can change their campaign strategy to address their concerns

Related work

1. A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election:
    a. Sentiment analysis of accessible tweets and tweets being removed from Twitter across time.
    b. **Insights:** removed tweets posted after the 2020 US Election Day sided with Joe Biden while those before Election Day were more favorable about Donald Trump.
2. Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 presidential election:
    a. Analysis of the tweets posted during the 2016 US elections and classification of sentiments into 6 groups: Trump supporter, Hillary supporter, whatever, positive, neutral and negative
    b. **Insights:** political homophily level rises when there are close connections and similar speeches

Related work

3. Analysis of political sentiment orientations on Twitter:

a. Long Short Term Memory (LSTM) classification model to predict the sentiments and results of the elections

b. **Insights:** dominance of support for a single party on Twitter in the 2019 General Elections of India

Proposed work: Dataset

Dataset Chosen: Collection of Tweets from the 2020 US presidential election related to Donald Trump and Joe Biden.

Dataset Key Features:
- Two subcollections (one for Trump, one for Biden).
- Total of approx. 1.72 million rows of data (970,919 rows for Trump and 776,886 rows for Biden).
- 394,400 or roughly 20% of the Tweets are made from the US.
- Total of 21 columns
- Notable columns: created_at, tweet, likes, reteweet_count, user_followers_count, user_location, lat, long, city, country, continent, and state_code

Proposed work: 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*

Proposed work

01

Text Classification

- Name Entity Recognition
  - Which parties are involved in this tweet?
  - Which candidates are involved in this tweet?
- Keyword Extraction
  - Which keywords are the most important/relevant in the tweet?
- Sentiment Analysis
  - 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?

Proposed work

02 Opinion Analysis

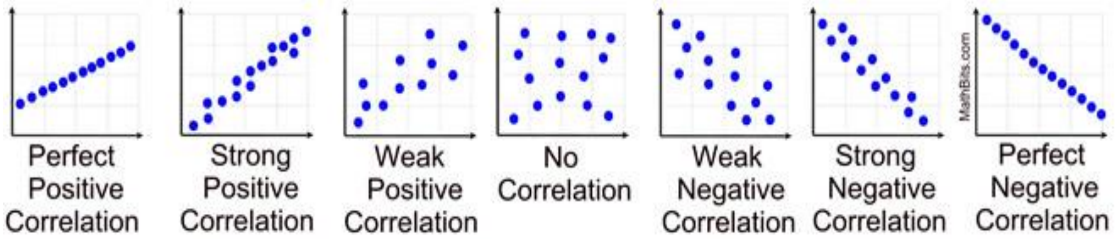
- Temporal Analysis
  - What are the predominant opinions over time?
- Categorization of Opinions according to:
  - State
  - Country
  - In the US vs outside of the US
  - Democratic vs Republican

03 Case Study on Donald Trump

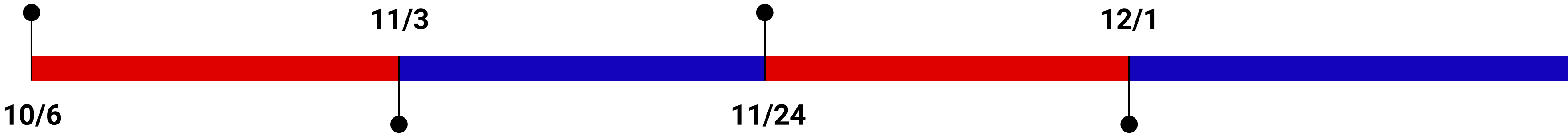
- How did opinions change over time?
- Did the timeline coincide with certain events?
- How did Twitter specifically help him/prevent him from swinging favor?
- What factors helped him garner his large voter share?

Evaluations

- Once we have found a strong correlation between sentiments made by Twitter users
- More specifically we plan to analyze possible correlations between Tweets made by twitter users during the 2016 and 2020 election periods regarding political candidates.
- We can model these findings in a correlation model, if we find that certain attributes during our research were useful.
- When completing our evaluation we also hope to consider different scenarios and see if we are able to find general patterns
- Doing so, will also allow us baseline the performance and build those results into our current correlation model



Milestones



10/6

Data Preprocessing

Data cleaning, text mining, natural language processing (text classification, topic modeling)

11/3

Data Analysis

Opinion analysis:

- Categorization: Which groups of people have these opinions?
- Temporal: How do opinions change over time?

11/24

Data Visualization

Charts: bar, line, choropleth, scatterplot maps to show results

12/1

Evaluation

- Evaluation of results
- Documentation: report writing, presentation slides