US Politics by Social Media

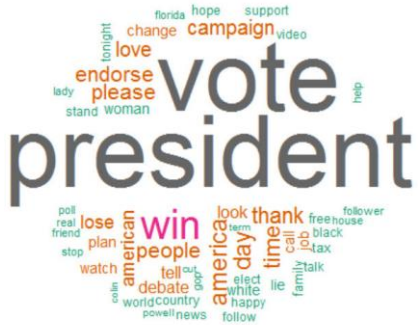
Team members: Rachel Ng Min Yee (CSCI 4502), Xinyi Lu (CSCI 4502), Anuragini Sinha (CSCI 4502)

Introduction - Background

Why is this interesting?

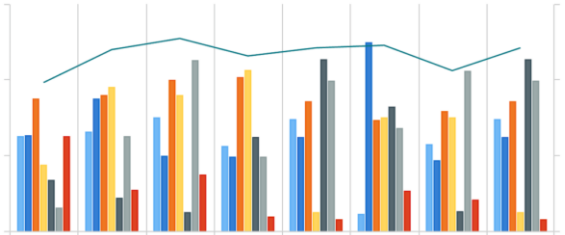
1. Election candidates can predict their own and their competitors' odds of success.
2. Voters can be critically aware of the influence of social media in politics and make more objective decisions.

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
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

Dataset

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

	created_at	tweet_id	tweet	likes	retweet_count	source	user_id	user_name	user_screen_name	user_description	...	user_followers_count
0	2020-10-15 00:00:01	1.316529e+18	#Elecciones2020 | En #Florida: #JoeBiden dice ...	0.0	0.0	TweetDeck	3.606665e+08	El Sol Latino News	elsollatinonews	Noticias de interés para latinos de la costa...	...	1860.0
1	2020-10-15 00:00:01	1.316529e+18	Usa 2020, Trump contro Facebook e Twitter: cop...	26.0	9.0	Social Mediaset	3.316176e+08	Tgcom24	MediasetTgcom24	Profilo ufficiale di Tgcom24: tutte le notizie...	...	1067661.0
2	2020-10-15 00:00:02	1.316529e+18	#Trump: As a student I used to hear for years,...	2.0	1.0	Twitter Web App	8.436472e+06	snarke	snarke	Will mock for food! Freelance writer, blogger,...	...	1185.0
3	2020-10-15 00:00:02	1.316529e+18	2 hours since last tweet from #Trump! Maybe he...	0.0	0.0	Trumpytweeter	8.283556e+17	Trumpytweeter	trumpytweeter	If he doesn't tweet for some time, should we b...	...	32.0
4	2020-10-15 00:00:08	1.316529e+18	You get a tie! And you get a tie! #Trump 's ra...	4.0	3.0	Twitter for iPhone	4.741380e+07	Rana Abtar - رنا أبتر	Ranaabtar	Washington Correspondent, Lebanese-American ,c...	...	5393.0

Sample dataset containing tweets related to Donald Trump

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

- Named 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

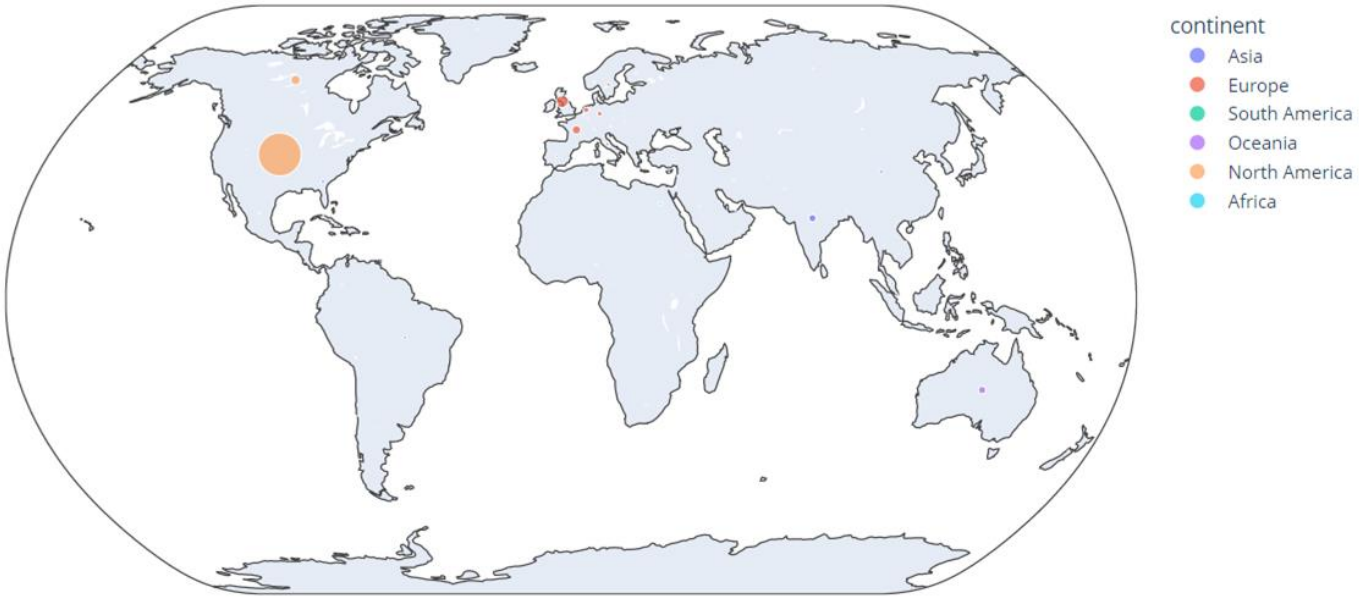
- How did opinions change over time?
- Did the timeline coincide with certain events?
- How did Twitter specifically help him/prevent him from swinging favor?
- What factors helped him garner his large voter share?

Completed Work

Data Preprocessing

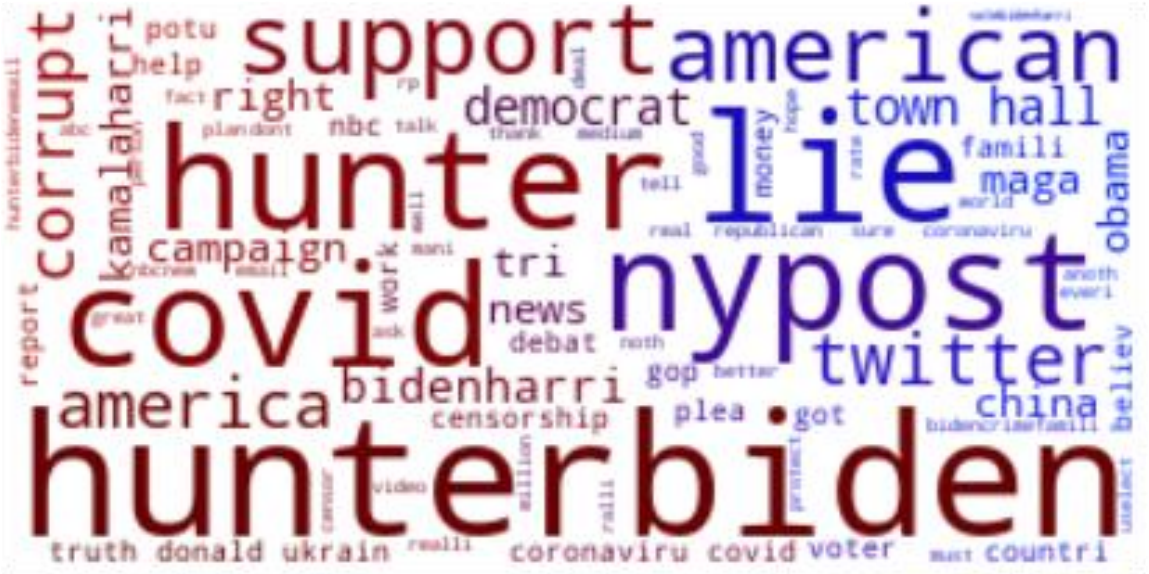
- **Data Cleaning**
  - Numeric, categorical columns
    - Drop rows with NA country
    - Convert date columns to datetime type (columns: created_at, user_join_date, collected_at)
    - Convert numeric columns to integer type (columns: tweet_id, likes, retweet_count, user_id, user_followers_count)
    - Standardise country names (eg. United States of America → United States)
  - Text column: tweet
    - Detect language of tweets → filter for english tweets
    - Text cleaning: remove punctuations, numbers, tokenization, remove stopwords, stemming and lemmatization
- **Sampling**
  - 20,000 rows of raw data (10000 from each dataframe)
  - Cleaned data: 8175 rows

Exploratory Data Analysis



Distribution of tweets across continents

Exploratory Data Analysis



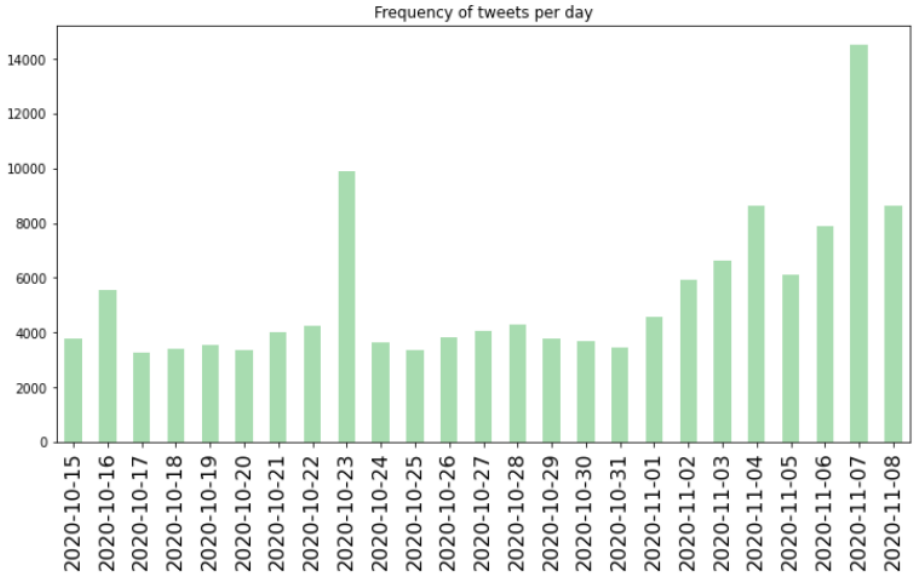
Word cloud of all tweets

Exploratory Data Analysis

Word cloud of all trump tweets

Word cloud of all biden tweets

Exploratory Data Analysis



Daily frequency of tweets for the full dataset

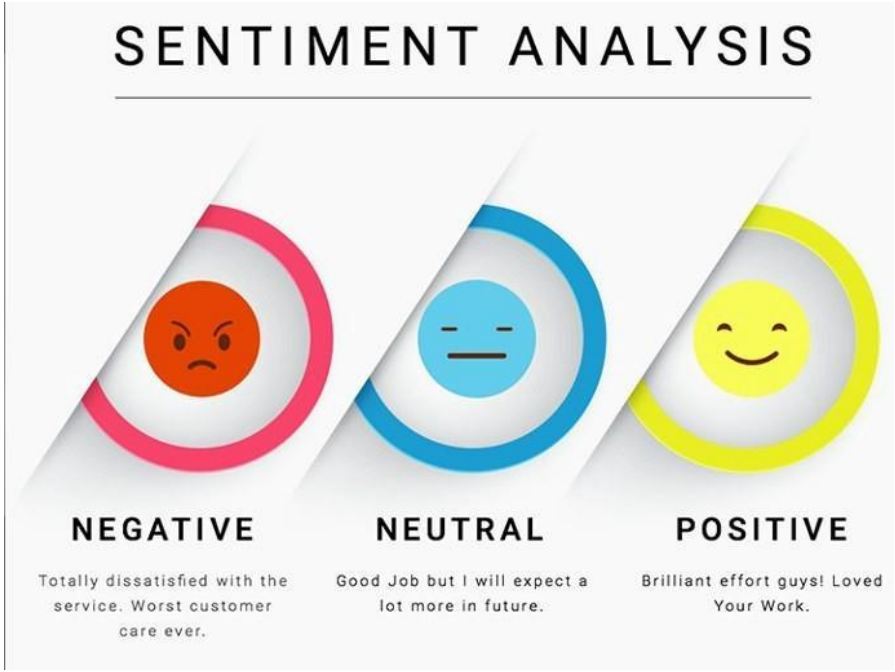
Work to be completed

Sentiment Analysis

Sentiment Analysis is one aspect of the project that is still being developed.

Sentiment Analysis effectively utilizes different Natural language processing techniques to extract and find these correlations between tweets made before and during the election.

This process will be done utilizing different sentiment analysis tools to find if the sentiments from these tweets are positive, negative, or neutral.



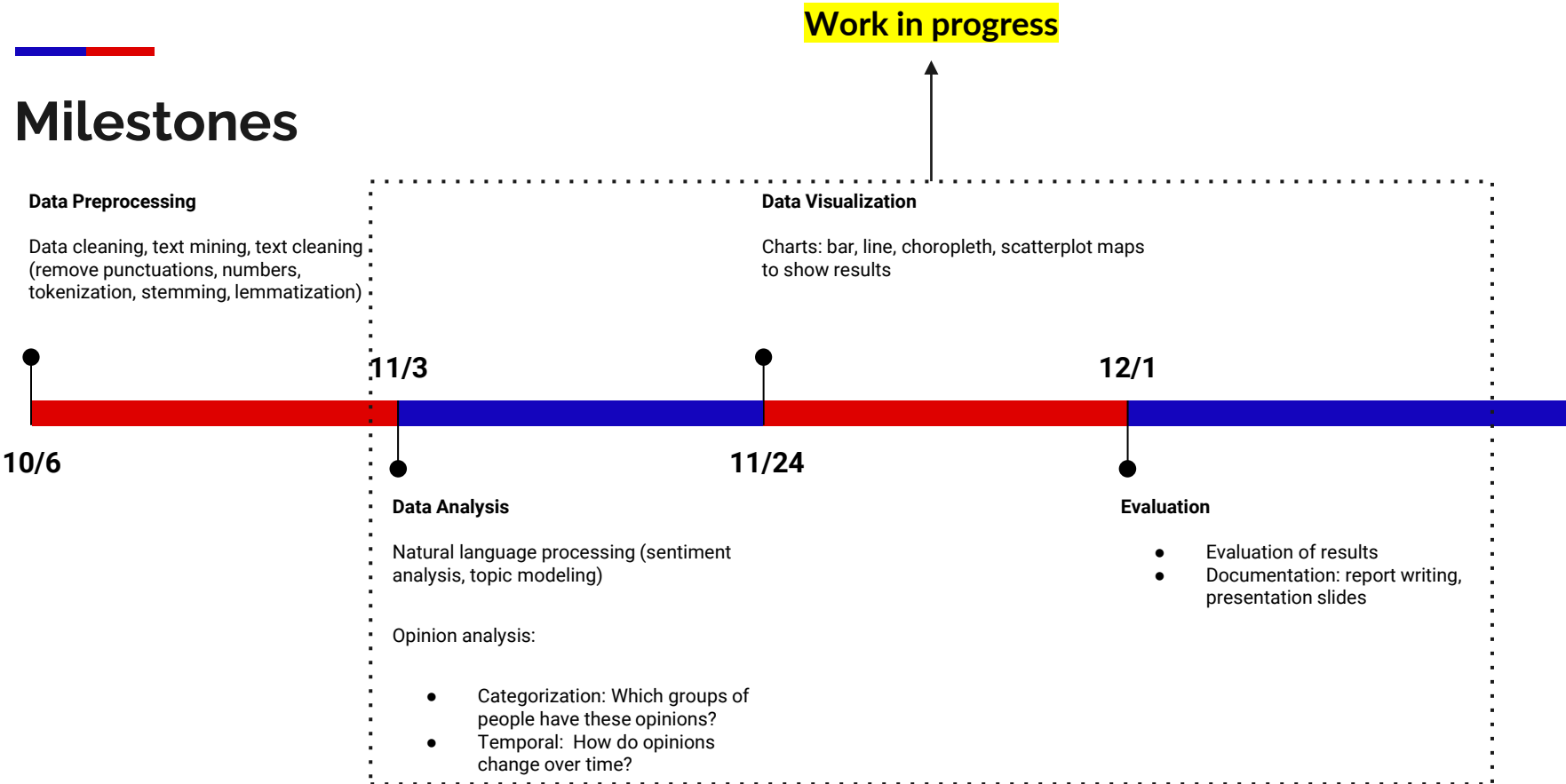
Opinion Analysis

- How do the opinions change **over time**?
- **Categorization** of opinions based on:
  - State
  - Country
  - In the US vs outside of US
  - Democrats vs Republicans

Case study on Donald Trump

- How did US election tweets help him/prevent him from swinging favor?
- What factors help him garner votes?

Milestones


Work in progress
Data Preprocessing
Data cleaning, text mining, text cleaning (remove punctuations, numbers, tokenization, stemming, lemmatization)
Data Visualization
Charts: bar, line, choropleth, scatterplot maps to show results
11/3
12/1
10/6
11/24
Data Analysis
Natural language processing (sentiment analysis, topic modeling)
Opinion analysis:
Categorization: Which groups of people have these opinions?
Temporal: How do opinions change over time?
Evaluation
Evaluation of results
Documentation: report writing, presentation slides

Thank you

Q&A