Original Dataset: https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

Github Repository (with cleaned/modified datasets and scripts): All data and code is accessible via this link https://github.com/ArjunaBazaz/2020_Twitter_Sentiment_Analysis

Preprocessing Steps:

1. Column Selection: Since not all columns are directly relevant to the analysis, unnecessary columns were removed to simplify the dataset. The following columns will be retained for immediate or potential future analysis:
    a. created_at: Used to filter tweets from the day before (October 21) and the day after (October 23) the debate, and to examine sentiment changes over time.
    b. tweet: The main text for sentiment analysis.
    c. about: Specifies whether the tweet is about Trump or Biden, enabling candidate-specific sentiment analysis.
    d. Additional columns retained for potential future analysis:
        i. likes and retweet_count: Can be used to measure engagement with tweets (optional).
        ii. state and country: May be used for demographic or geographic sentiment analysis (e.g., swing states, optional).
        iii. user_followers_count: Potentially used to assess the influence of user reach on sentiment trends.
2. Handled missing values:
    a. Categorical variables: For user_location, state, and country, missing values will not be imputed. Data points with missing values will be excluded from any analysis involving location since assumptions about user location cannot be made.
    b. Quantitative variables: For user_followers_count, missing values will be replaced with the mean value (18,418) under the assumption that the distribution of followers is consistent enough to allow for mean replacement without introducing bias.
3. Text cleaning: The main focus of sentiment analysis is the "tweet" column, which underwent basic cleaning to improve the accuracy of the VADER sentiment tool. We removed irrelevant components such as URLs, emojis, hashtags, mentions (e.g., "@username"), and extra white spaces. We retained punctuation and capitalization since VADER uses these features to enhance sentiment detection.

Unit of Observation:

Each row in the dataset represents a single tweet related to the 2020 U.S. presidential debate.

Data Dictionary:

Column Name	Data Type	Details	Potential Usage
created_at	Quantitative (String)	The time of the tweet in Date-Time format	Can be used to show sentiment changes over time
tweet	Categorical (String)	The tweet body	Can be used in text analysis to gain insight from the tweet itself
likes	Quantitative (int)	The number of likes	Can be used to measure which kind of tweets got more likes
retweet_count	Quantitative (int)	The number of retweets	Can be used to measure which kind of tweets got more retweets
user_followers_count	Quantitative (int)	The number of followers a user has	Determine whether people with more followers lean more to one candidate
user_location	Categorical (String)	The location of the user	Determine which locations had the most favorable opinions of each candidate
country	Categorical (String)	The nation of user's account	Can be used to filter out tweets to just American tweets
state	Categorical (String)	The state of the user's account	Can be used to determine if swing states favored one candidate more
about	Categorical (String)	Whether the tweet was about Trump or Biden	Used to classify tweets by candidate to do statistical analysis
Compound_Score	Quantitative (Float)	The overall sentiment score of the tweet, calculated using VADER's compound score	Used to measure sentiment intensity towards a candidate
Sentiment	Categorical (String)	The categorized sentiment of the tweet: "Positive," "Negative," or	Can be used to compare overall sentiment trends between candidates

Column Name	Data Type	Details	Potential Usage
created_at	Quantitative (String)	The time of the tweet in Date-Time format	Can be used to show sentiment changes over time
tweet	Categorical (String)	The tweet body	Can be used in text analysis to gain insight from the tweet itself
likes	Quantitative (int)	The number of likes	Can be used to measure which kind of tweets got more likes
retweet_count	Quantitative (int)	The number of retweets	Can be used to measure which kind of tweets got more retweets
user_followers_count	Quantitative (int)	The number of followers a user has	Determine whether people with more followers lean more to one candidate
user_location	Categorical (String)	The location of the user	Determine which locations had the most favorable opinions of each candidate
country	Categorical (String)	The nation of user's account	Can be used to filter out tweets to just American tweets
		"Neutral"	

Figures:

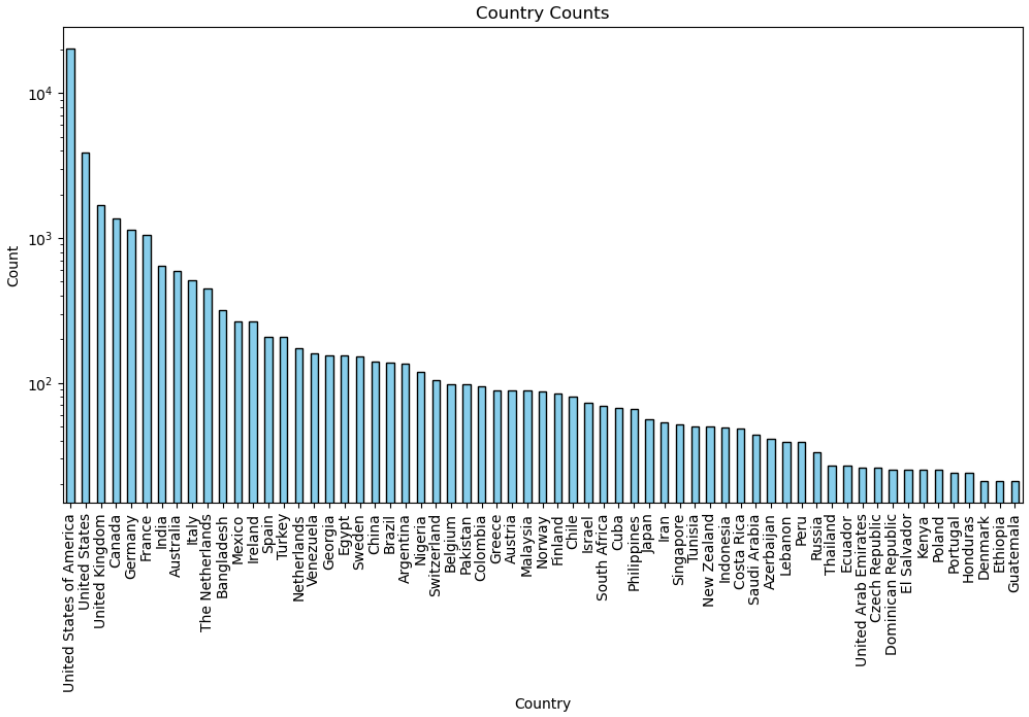

Figure 1 - The number of tweets by nation (minimum 10 tweets) on a log scale

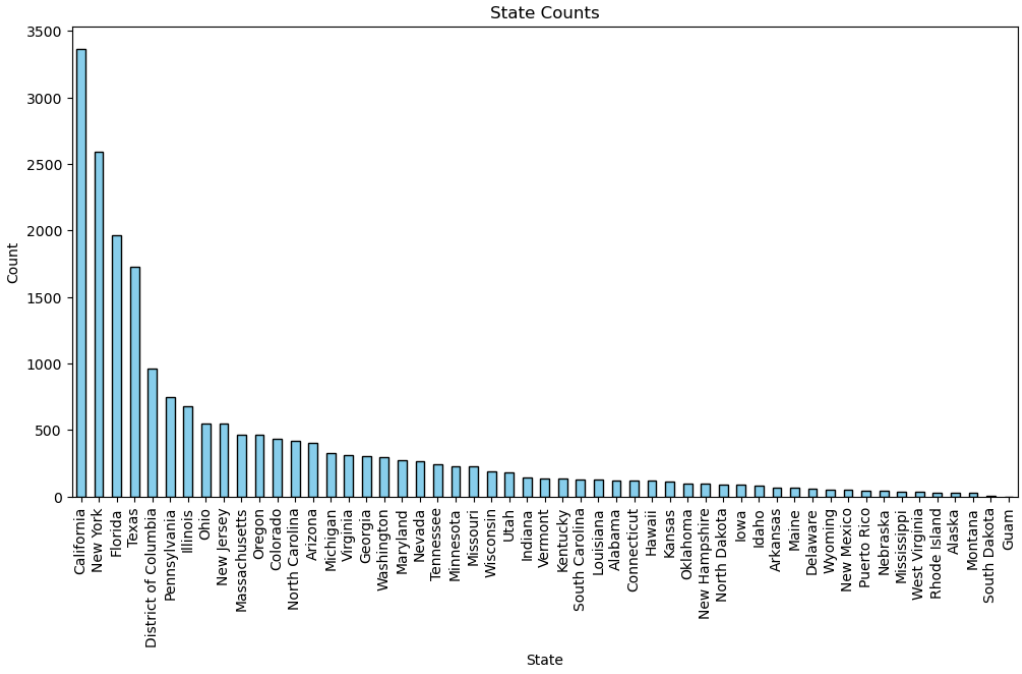

Figure 2 - The number of tweets by state

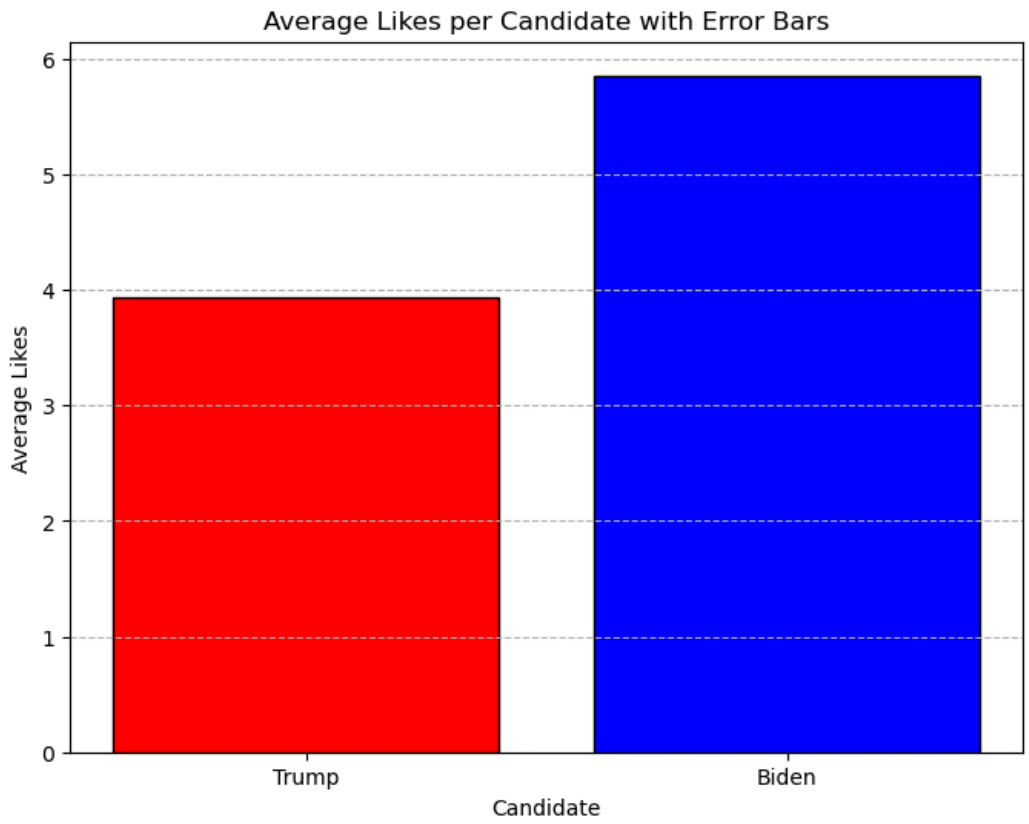

Figure 3 - The average number of likes per candidate from October 21st - October 23rd

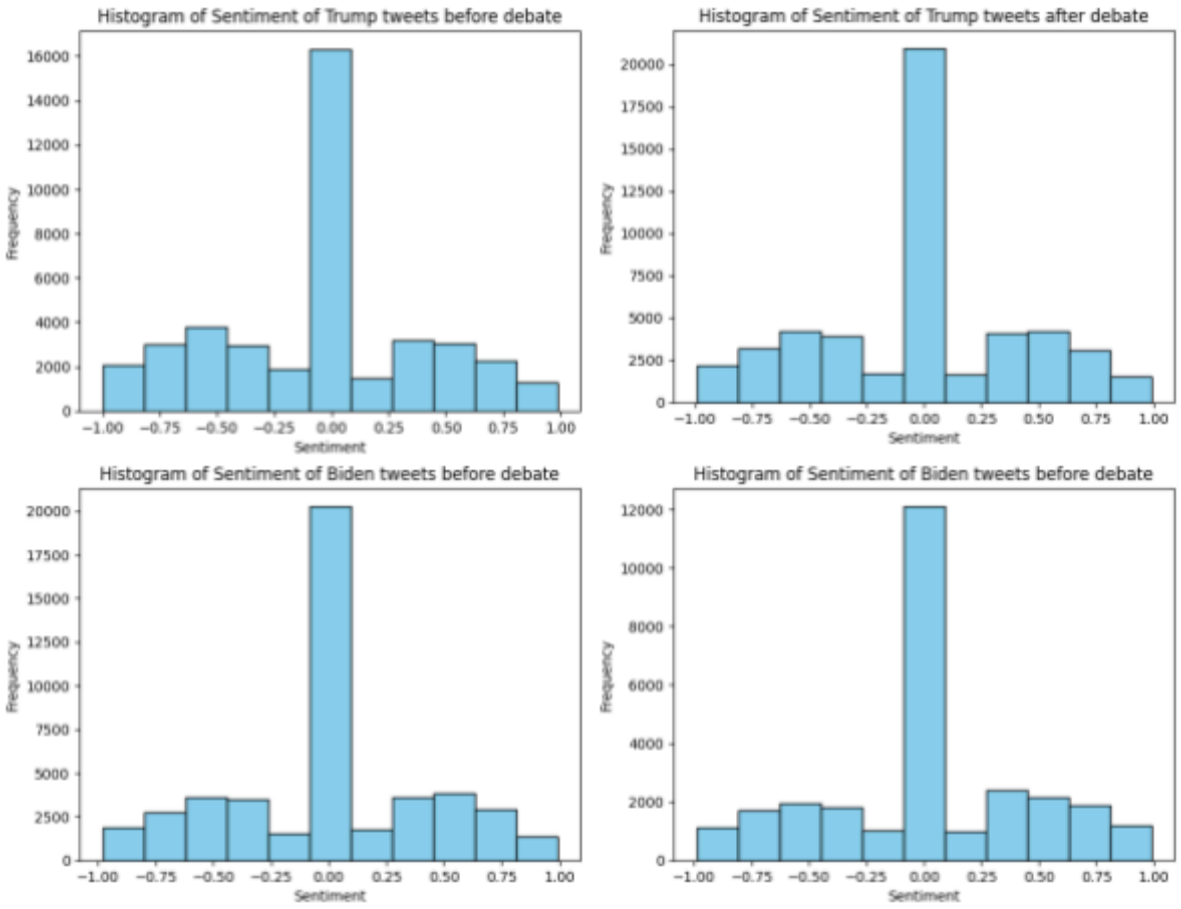

Figure 3 - Histograms of sentiments of Biden and Trump before and after the debate

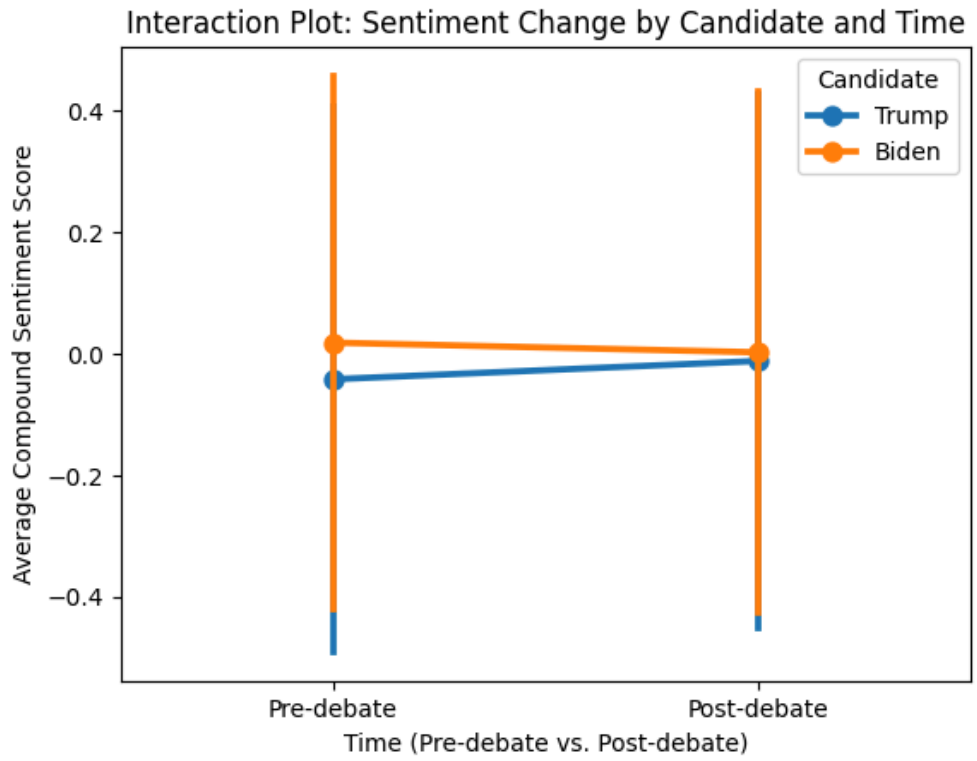

Figure 4 - A lineplot to visualize interaction of candidate and time on sentiment

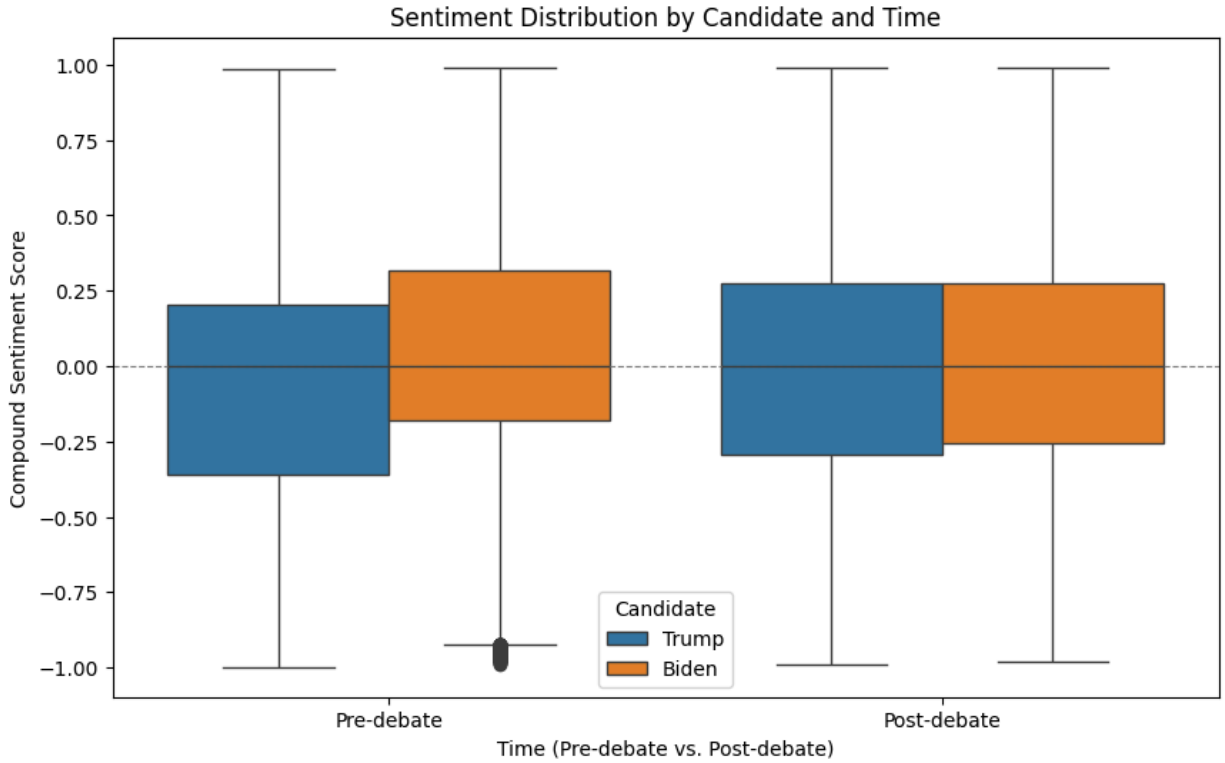

Figure 5 - Boxplots of sentiment distributions for Trump and Biden pre and post debate.

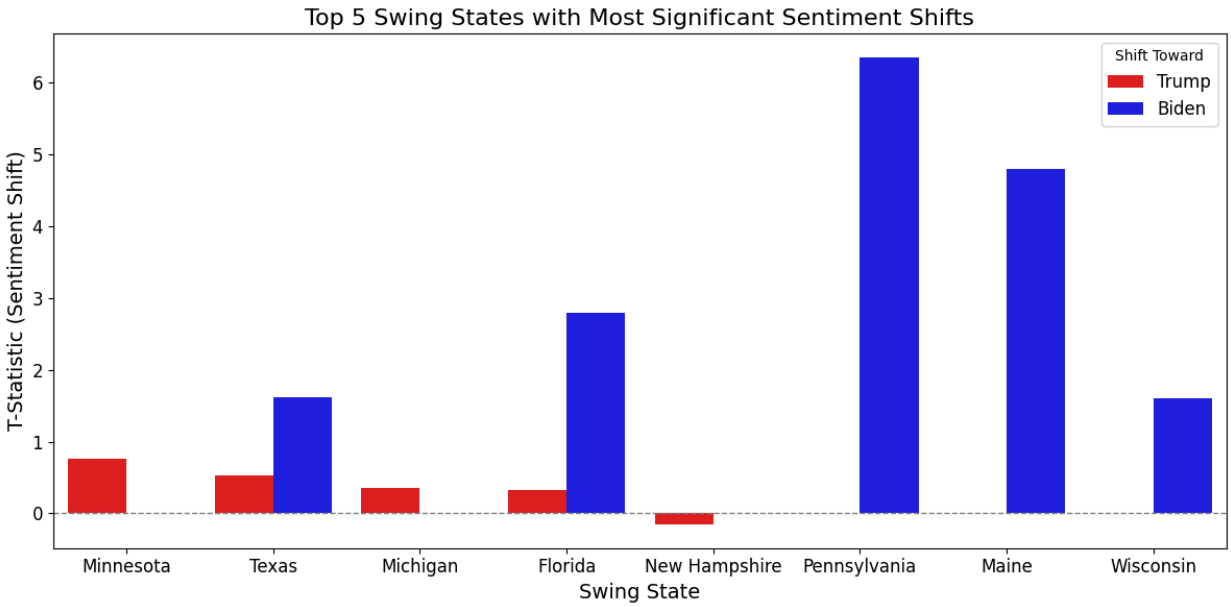

Figure 6 - A bar chart showing the top 5 states with the most significant sentiment shifts toward Trump and Biden (based on the T-statistic).