PyPioneers Final Draft

Giorgio Ramirez

Suprya Azhakath

Ooreoluwa Ayeni

Chengyin Hu

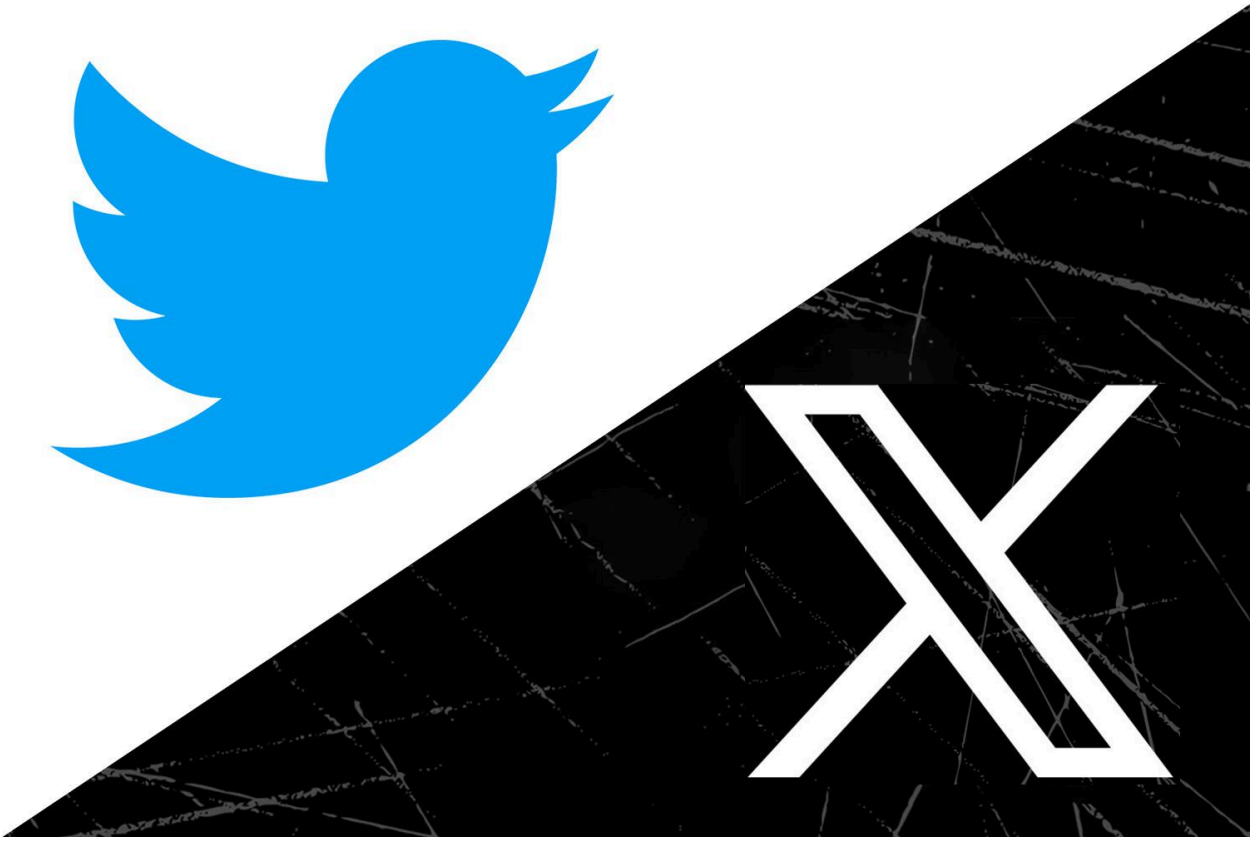


Table of Contents

Abstract/Executive Summary

Purpose:

Sentiment analysis surrounding text is a popular branch of AI that has gained interest in the last few years with the emergence of more robust supercomputing and interest in other NLP techniques. Large datasets with millions of rows and dozens of features can be analyzed at a speed that was not possible even 10 years ago. With this project, we wanted to analyze tweet sentiment during the 2020 US election, focusing on tweets related to Donald Trump and Joe Biden. This was done using a sentiment analysis VADER model. Even though X (formerly known as Twitter) has limited access to its API to paying customers, techniques to disseminate the data into important insights are still useful to those with the resources to access it.

Data:

We downloaded a dataset of tweets mentioning Trump and Biden from Kaggle during the 2020 US election period. The dataset included:

- Tweet content
- Likes and retweets
- User information (e.g., follower count, join date)

Summary of Findings:

1. Sentiment Analysis
2. Predictive Modeling:

- Random Forest was identified as the most accurate model for predicting retweet counts, offering the highest R-squared value and lowest mean absolute error among the models tested.
- Gradient Boosting also performed well but was slightly less effective than Random Forest.
- Linear Regression and Decision Tree models were less effective in capturing the complex relationships within the data.

3. Feature Importance:
    - The number of likes a tweet received was the most significant predictor of retweet counts.
    - Other features, such as user follower count and sentiment scores, also contributed to the prediction but to a lesser extent.

Project Plan

Primary Company Details:

Name: X Corp. (formerly Twitter, Inc.)

Owner: Elon Musk

Founded: 2006 (as Twitter), rebranded to X Corp. in 2023

Address:

Headquarters: 1355 Market Street

Suite 900

San Francisco, CA 94103, USA

Company Communication:

Website: x.com

Customer Support: Accessible via the help center on the website

Business Description:

X Corp. operates a social networking service formerly known as Twitter. It enables users to post and interact with messages known as "tweets." The platform is used for real-time information sharing and social networking, and it plans to expand into financial services and other domains.

Financials:

Valuation: Estimated at $28.5 billion as of August 2023

Revenue: Diverse sources, including advertising, subscription services, and potential future financial services.

Key Executives:

CEO: Linda Yaccarino (since June 2023)

Executive Chairman & CTO: Elon Musk

Major Competitors:

Social Media Platforms: Facebook (Meta), Instagram, TikTok, LinkedIn

Emerging Competitors: Mastodon, Bluesky Social

Business/Analysis Opportunity: Leveraging a dataset of 1.72 million tweets from the 2020 presidential election provided by Kaggle. The dataset lists 21 features, including user information, retweet count, and other geographic metrics, which provide a solid foundation for in-depth sentiment analysis to predict public opinions during the 2020 election. Our data team can build predictive analytical models by investigating key research questions. These predictive models will reveal the relationship between Twitter sentiment scores and outcomes from past 2020 elections. In addition, applying insights gained from these valuable datasets can help predict outcomes for the 2024 election between Joe Biden and Donald Trump.

Research Questions:

Having an endless resource of real-time semi-structured Twitter data is extremely useful to those with the resources to access it. In a US presidential campaign, it is important to have access to data about potential voters, what is important to these voters, and any strategy that may be gleaned from these data.

RQ1: Can sentiment analysis be conducted to predict the electoral college votes and, therefore, the election, given 2020 election tweets and geolocation data? Which model most accurately predicts this?

The importance of the Office of the President of the United States goes without saying. Therefore, the ability to accurately forecast the winner of this office based on accessible data is significant and can be a real asset to an election campaign. By categorizing the words within texts, the research question attempts to optimize the most suitable machine learning model that can most accurately quantify the number of electoral college votes for each presidential candidate based on geolocation data given in the dataset. Four models will be compared: logistic regression, SVM, Random Forest, and Naive Bayes.

RQ2: Can a machine learning model accurately predict the amount of retweets a tweet gets based on sentiment analysis? Which model most accurately predicts this?

Within a tweet, various metrics can be analyzed. Among these are the amount of retweets that a tweet gets. Retweeting can signify different things, such as the level of engagement, the connection among various users and communities, and the endorsement of specific ideas or hashtags in the tweet. Understanding and predicting the number of retweets may offer insight into what posts would be most effective among voters for a presidential campaign. As in the previous question, four models will be compared: logistic regression, SVM, Random Forest, and Naive Bayes.

Hypothesis

H1: Sentiment analysis can be used to predict the winner of a state in a presidential election.

Graded sentiment analysis and geolocation data allow us to analyze the data, separate the tweets into states, and make predictions with various models to determine which one will be the most accurate. This study aims to implement and compare the accuracy of different predictive models in forecasting electoral outcomes based on the sentiment analysis of tweets**.**

H2: Tweets from users with higher follower counts and verified accounts will receive more likes and retweets

The tweet's engagement is typically measured by the number of retweets it receives. The number of retweets is generally influenced by sentiment and the user's social influence, which follower counts can measure. Is a pattern evident in our Twitter dataset, and is it associated with an accurate increase in retweets? We intend to find the most effective machine learning models to test this hypothesis.

Data

The data comes from Kaggle, a user who collected the tweets using the Twitter API. From October 15, 2020, to November 8, 2020, there were 1.72 M tweets with Joe Biden and Donald Trump hashtags, with 21 features. Text from the tweet column will be used in sentiment analysis modeling.

Users

The dataset includes details describing 363,161 unique users. The characteristic variables include user type, user ID, name, user description, user screen name, join date, follower count, and location. Some users have multiple tweets, which will be factored into the modeling process.

Tweets

The dataset contains 1.7 million tweets, including tweet ID, text, creation date and time, number of likes, number of retweets, and the source of the tweet.

Geolocation Data

Geolocation data includes latitude, longitude, city, state, state code, and country parsed from user location information, providing insights into regional public opinion trends.

Check-ins

The check-in portion of the data consists of the timestamps of tweet creation, representing real-time user engagement during the election period. Each tweet includes the creation date and time, showing user activity and daily engagement.

Measurements

Sentiment from tweets can be measured in different ways. A graded approach, on a scale of 1-5, will give more information to the models used than simply evaluating positive, negative, or neutral. After labeling, preprocessing, and feature extraction, the models are trained, and each tweet is given a sentiment label. Our project will focus on these labels and the models used to determine them.

Methodology

RQ1: Predicting Electoral College Votes Using Sentiment Analysis and Geolocation Data

We start by collecting tweets from the 2020 US election dataset to predict electoral college votes, focusing on tweet text, user information, and geolocation data. We clean and preprocess the tweets, removing irrelevant content and labeling sentiment (positive, negative, neutral) using pre-trained models like VADER or BERT. We aggregate these sentiment scores at the state level, incorporating demographic and historical voting data for context. We then train four machine learning models: Logistic Regression, Support Vector Machine (SVM), Random Forest, and

Naive Bayes, splitting the data into training and testing sets. Model performance is evaluated using accuracy, precision, recall, and F1-score, with the most accurate model selected for predicting state-wise electoral outcomes.

RQ2: Predicting Retweet Counts Based on Sentiment Analysis

We collect the 2020 US election tweets, including tweet text, user information, and retweet counts to predict retweet counts. The tweets are cleaned and preprocessed to remove unnecessary content, and the sentiment is labeled using sentiment analysis tools. We train four regression models: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Regressor, using training and testing data splits. Model performance is evaluated using metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared, selecting the model with the best performance metrics for predicting retweet counts.

Computational Methods and Outputs

With classification, we believe AUC/ROC will most effectively reflect which models perform best. Additionally, we may divide the training data based on context-specific training data to be effective for unstructured text. Our outputs for research questions would be predictions for electoral outcomes and retweet counts.

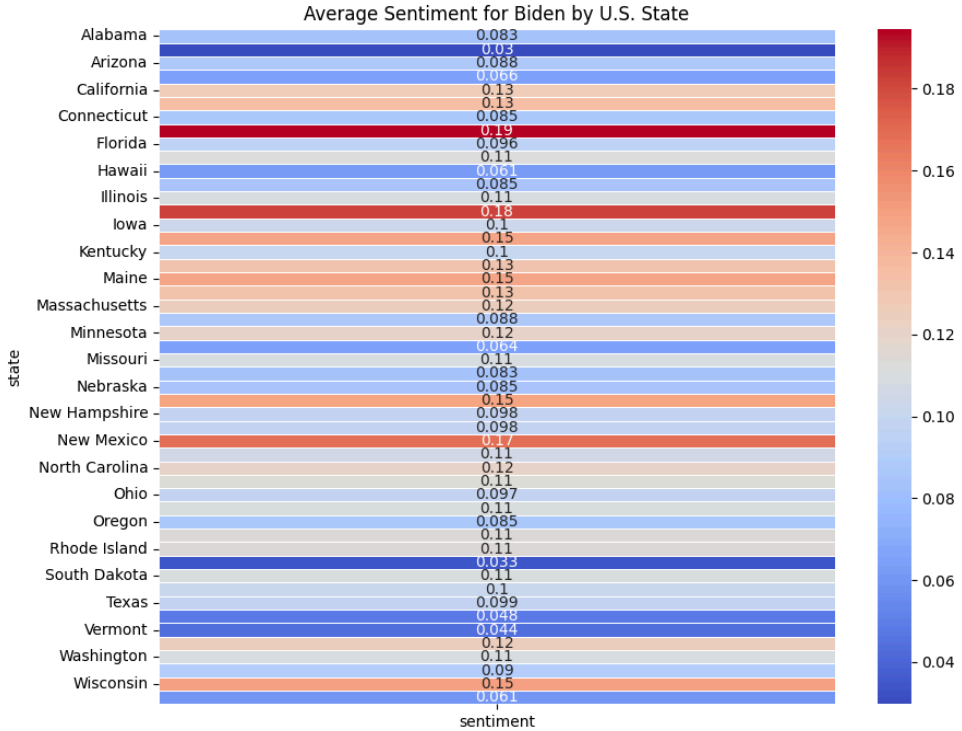
For data preprocessing, we would remove irrelevant content like ads, non-English tweets, and duplicates (if any) during the data cleaning. After that, we will perform model training, using classification models for predicting electoral college votes, such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and Naive Bayes. Then, we would use regression models to predict retweet counts, such as linear regression, random forest regression, and support vector.

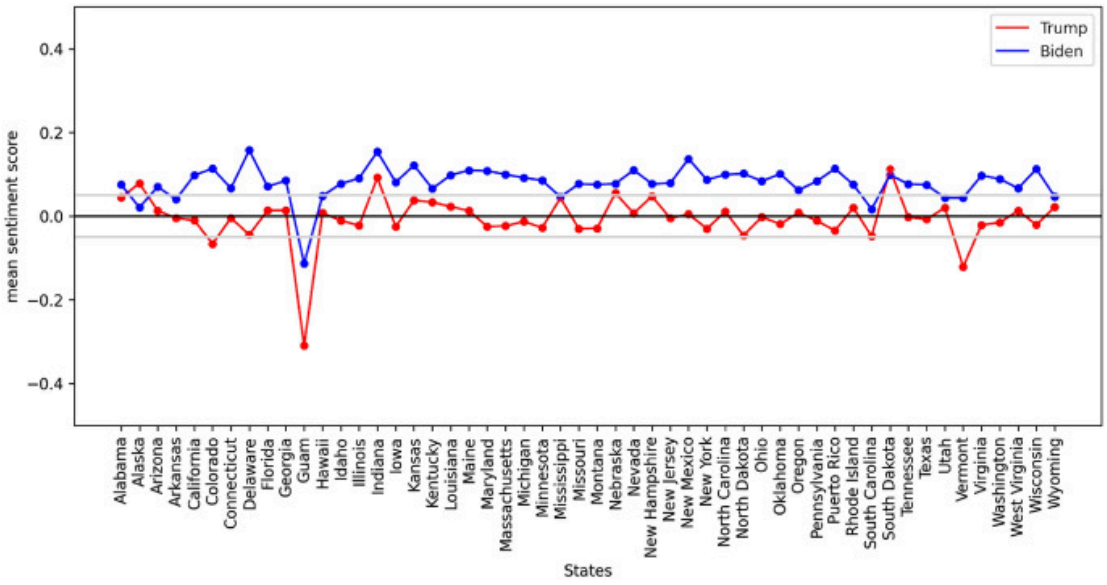
After classification, where we evaluate the model using accuracy, precision, recall, F-1 Score, and the AUC/ROC curve, we will implement multi-fold cross-validation to prevent overfitting. We will then select the best-performing models based on evaluation metrics for final predictions.

Output Summaries

RQ1: Predicting Electoral College Votes

The analysis will identify the states with the highest probability of voting for each candidate. We will include a table listing the top states sorted by the highest probability of voting for each candidate and a heat map using state-level sentiment scores, highlighting regions with solid support for each candidate.

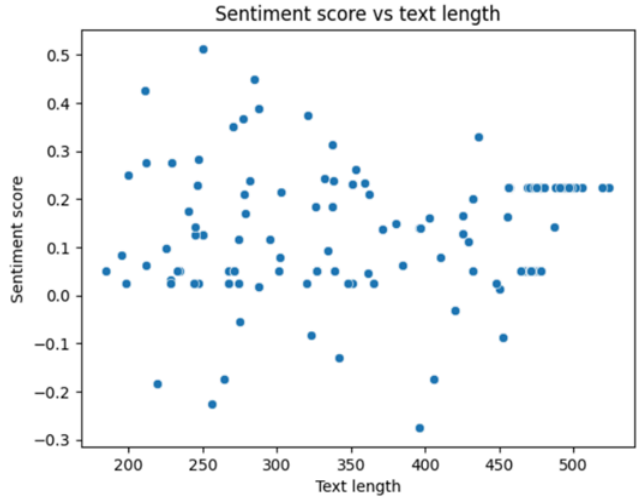




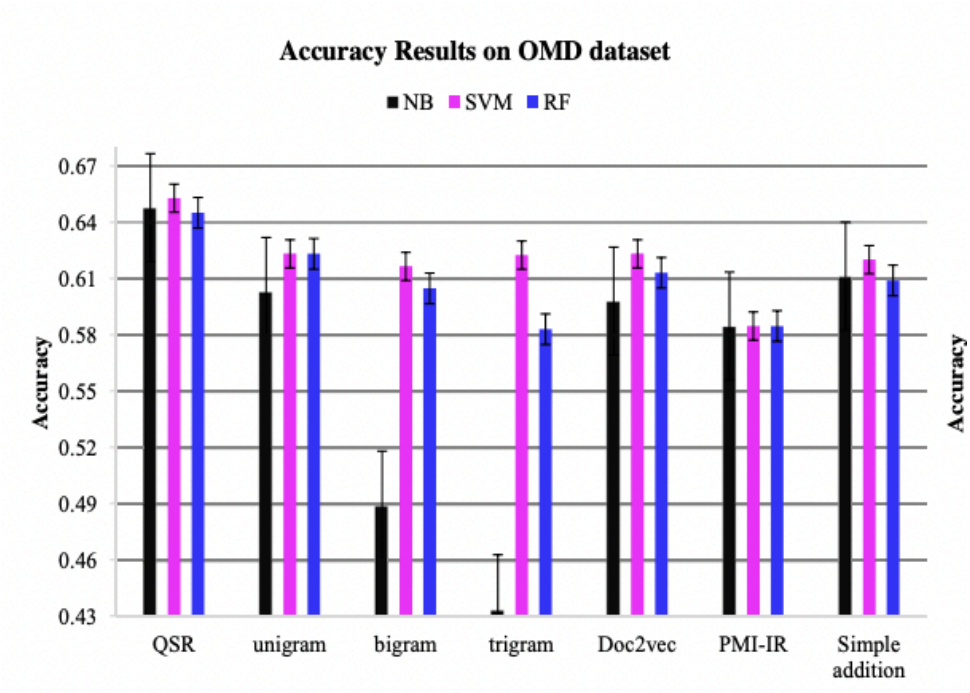
(This sample visualizes the mean sentiment score of tweets about Trump and Biden across different states. It effectively illustrates how sentiment towards each candidate varies by state)

RQ2: Predicting Retweet Counts

Our analysis will identify factors influencing retweet counts. We will include a scatter plot showing the relationship between sentiment scores and retweet counts and a bar chart showing the accuracy of the different models.



(A scatter plot with sentiment score and text length for our data will be # of retweets)



(A bar graph showing the accuracy of models and feature extraction method chosen)

Campaign Implementation

The results from these analyses can guide campaign strategies by identifying regions with solid support and predicting the spread of campaign messages on social media. By understanding the sentiment and engagement patterns, campaigns can effectively tailor their strategies to maximize voter outreach and engagement.

Literature Review

It is estimated that X (formerly Twitter) averages 500 million active monthly users in 2024. With its broad reach and usage, many researchers and companies have used its text data for academic and business purposes. In 2023, however, the X API was converted to a pay-for-access version. Although the pricing makes it prohibitive for many to access the data, election campaign organizations have the resources and still find it helpful to guide campaign strategy. X provides real-time data in the form of posts that campaign strategists can analyze to determine opinions about critical issues, assess campaign strategy effectiveness, and measure how their competitors are faring. Using sentiment analysis and other text analysis techniques and modeling, one can make predictions and conduct ad hoc analysis.

There are many different ways to approach a sentiment analysis problem. Other than the various models to be used, the target sentiment and how it is evaluated are important and relate to the question being answered. The sentiment analysis results may be analyzed through a scale: Most positive to most negative, 5 to 0, or to detect specific emotions: sad, happy, mad, or aspect-based: Positive, neutral, negative. Given the dataset of Trump and Biden tweets from the 2020 election, a graded approach gives more nuanced insights to any election-based questions.

Choosing the most suitable machine learning model for predicting electoral college results based on geolocation data is one straightforward and important problem to explore. However, understanding the importance of modeling other parts of the hashtagged tweets can be just as important to analyze. One metric of a tweet is the amount of retweets it gets. Retweeting can signify many things, including engagement, endorsement, and identifying communities. In creating and comparing machine learning models to predict the number of retweets, one interested in election data can gain insights into current topics and user sentiment.

The Kaggle dataset used in this project is a collection of tweets relating to the 2020 election with hashtags of the two candidates: Joe Biden and Donald Trump. Collected from October 15, 2020, to November 8, 2020, this dataset comprises approximately 1.7 million tweets with 21 features, including tweet text, user information, engagement metrics, and geolocation data. Its size allows for a comprehensive analysis of sentiment and engagement trends, making it a valuable resource for understanding public opinion during the election period.

Several studies, including the work of Yaqub (Yaqub et al., 2020), have demonstrated the immense potential of sentiment analysis in predicting election outcomes. Their research, which involved sentiment analysis on Twitter data to evaluate user sentiment towards presidential candidates in the 2016 US and 2017 UK elections, revealed that sentiment derived from Twitter location data generally reflected on-ground public opinion as evidenced by election results. This study highlighted the power of sentiment analysis in providing insights into the polarized nature of the election discourse on Twitter.

Similarly, a study on the impact of the 2020 US presidential election on Twitter data found significant differences in sentiment and engagement metrics between tweets about Trump and Biden. This research emphasizes the importance of analyzing deleted and suspended tweets to capture the true sentiments during significant events (Lawrie et al., 2).

Furthermore, sentiment analysis of tweets during the COVID-19 pandemic (Sancoko, 2022) revealed fluctuations in public sentiment towards the candidates, underscoring the impact of real-time events on social media discourse.

Various methodologies have been employed in sentiment analysis. These tools analyze text for subjectivity and polarity, helping researchers determine sentiment in tweets. Yaqub's study utilized these tools to map sentiment geographically, revealing that location-based

sentiment analysis can provide a nuanced understanding of public opinion across different regions. Innovative approaches have also been explored, such as integrating real-time data and advanced machine-learning models.

While previous studies have effectively used Twitter data for sentiment analysis and election predictions, some areas urgently warrant further exploration. The need for more granular analysis, such as examining sentiment at the city or county level, remains a significant gap. Additionally, the impact of real-time events on social media sentiment and the role of deleted or suspended tweets in shaping public opinion require immediate attention and more research.

Given the findings from these previous studies, analyzing the 2020 US election tweets using sentiment analysis can offer valuable insights. The ability to predict election outcomes and understand public sentiment through social media analysis underscores the importance of these tools in modern political campaigns. This research can contribute to a deeper understanding of the dynamics during the 2020 election period by comparing the effectiveness of various sentiment analysis models and predicting retweet counts.

Sentiment analysis integration in political research has proven to be a powerful method for gauging public opinion and predicting election outcomes. Building on the methodologies and findings from previous studies, this project aims to explore further the potential of Twitter data in understanding electoral dynamics, choosing the most suitable machine learning model, and providing a comprehensive view of public sentiment during the 2020 US election.

Final Research Questions

When we began this project, our initial research questions were broad and aimed at exploring the overall dynamics of tweet engagement during the 2020 US election. As we delved deeper into the exploratory data analysis (EDA) and performed initial analyses, we identified specific trends, patterns, and gaps that helped us refine our research questions. Based on the insights gained from our analyses and the performance of various models, we refined our research questions to focus more precisely on the predictive aspect and the factors influencing tweet engagement. The geolocation data proved to be insufficient in order for modeling because many of the tweets weren't labeled. Our final research questions are:

RQ1: How does tweet daily sentiment analysis predict overall national polls and what are the relationships between the two?

RQ2: Can a machine learning model accurately predict the amount of retweets a tweet gets based on sentiment analysis? Which model most accurately predicts this?

Initially, we wanted to understand the data by performing EDA and visualizing key variables' basic statistics and distributions. Through EDA and correlation analysis, we identified patterns and relationships between sentiment scores and engagement metrics. We created additional features to enhance model performance and built multiple models to predict tweet engagement. Evaluating model performance provided a clear view of our approach's strengths and limitations, leading to the improvement of our research questions.

Exploratory Data Analysis

For this project, the data collected is from Kaggle based on the US Election 2020 tweets. The dataset comprises tweets that mention key election-related terms and hashtags for the leading candidates, Trump and Biden. Since the research focuses mainly on sentiment analysis for predicting the number of retweets and comparing polling data, removing irrelevant or noisy data is the first step. This included trailing spaces, symbols, and stop words. After these preprocessing steps, the data was reduced to 13 features and 694,046 observations.

The dataset includes the following key columns:

Created_at: Date and time of tweet creation

Likes: Number of likes a tweet has gotten

Retweet_count: The number of retweets

Source: The source used to post the tweet

User_name: The user ID of the person who tweeted

User_join_date: The date the user joined

User_followers_count: The number of followers the user has obtained

We created four new variables based on the 'tweets' and their sentiment scores to assist in exploring the data that would be advantageous to answering the research questions. Those variables are:

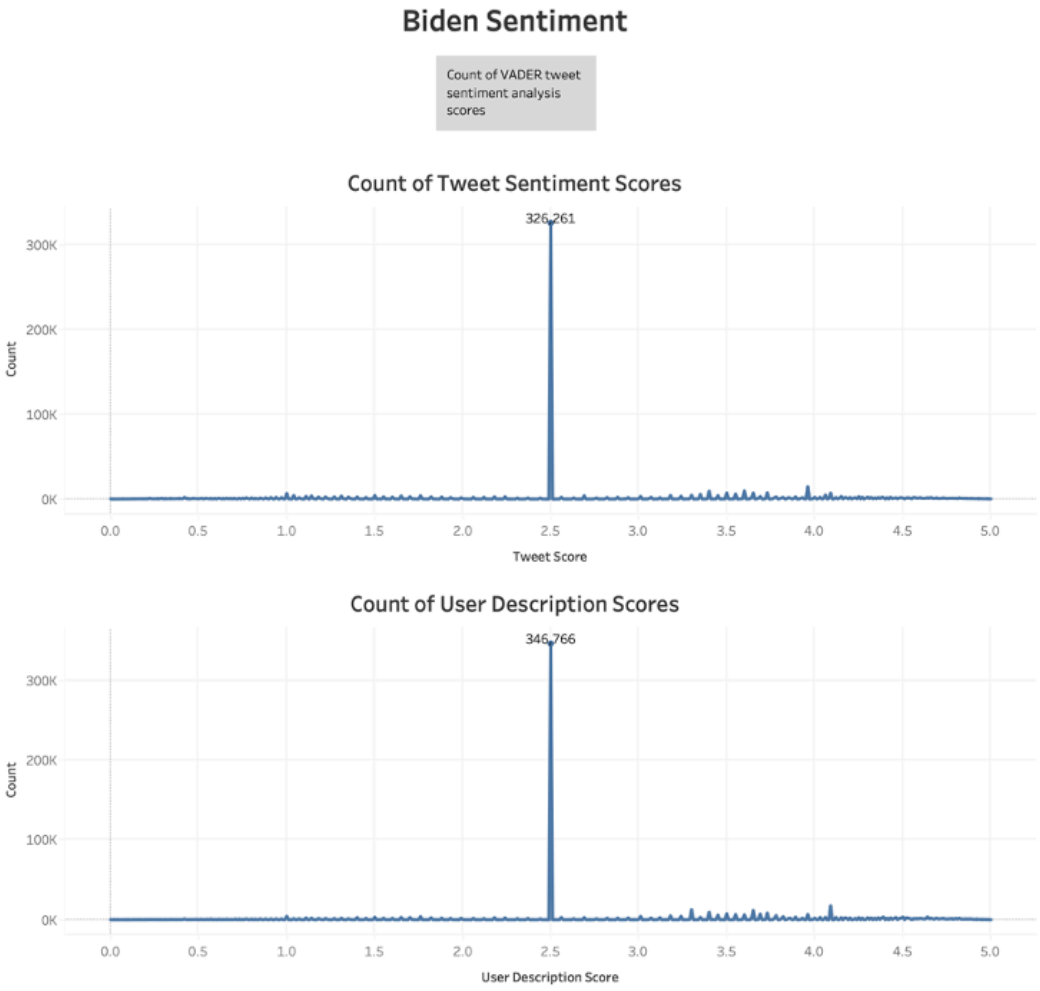
Cleaned_tweet: The cleaned text of the tweet

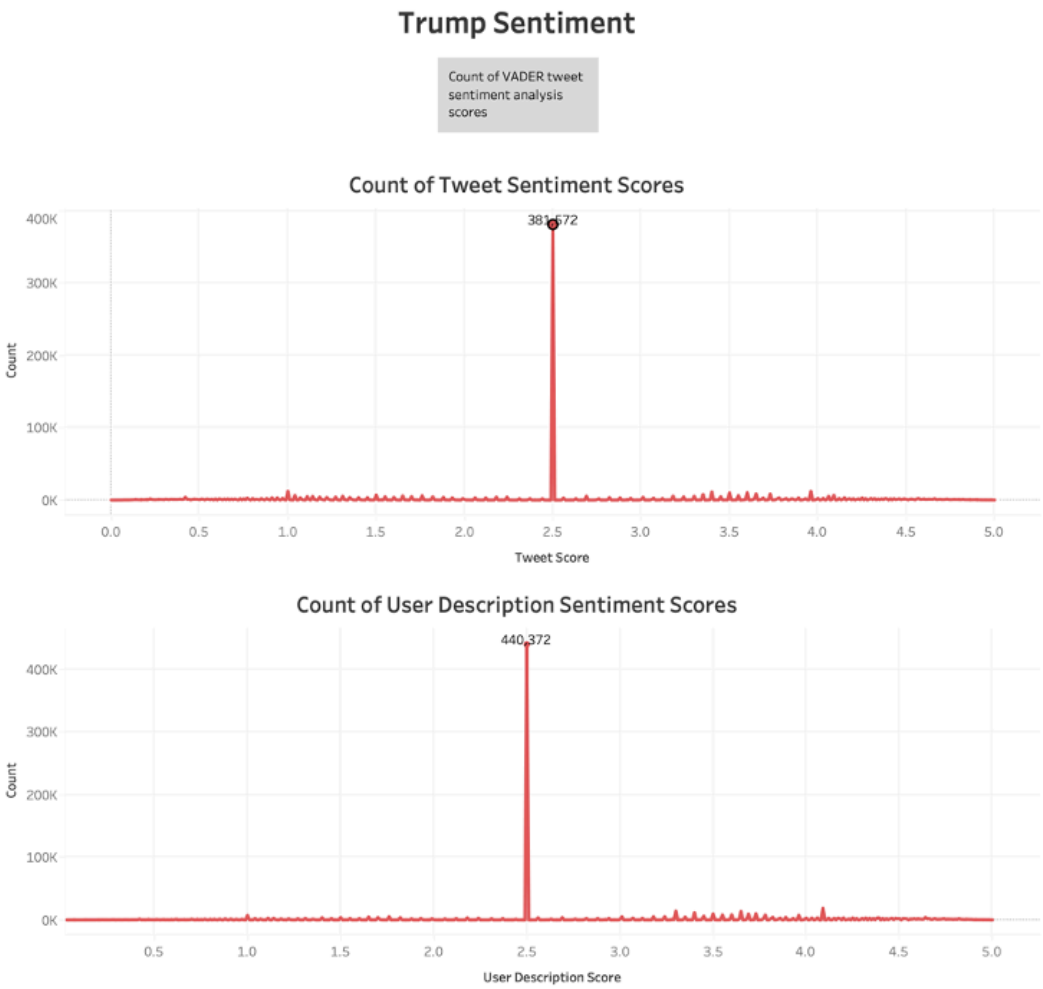
Cleaned_user_description: The description of self by the user

Tweet_sentiment_score: represents the sentiment score of individual tweets

User_sentiment_score: represents the sentiment score of the user's profile

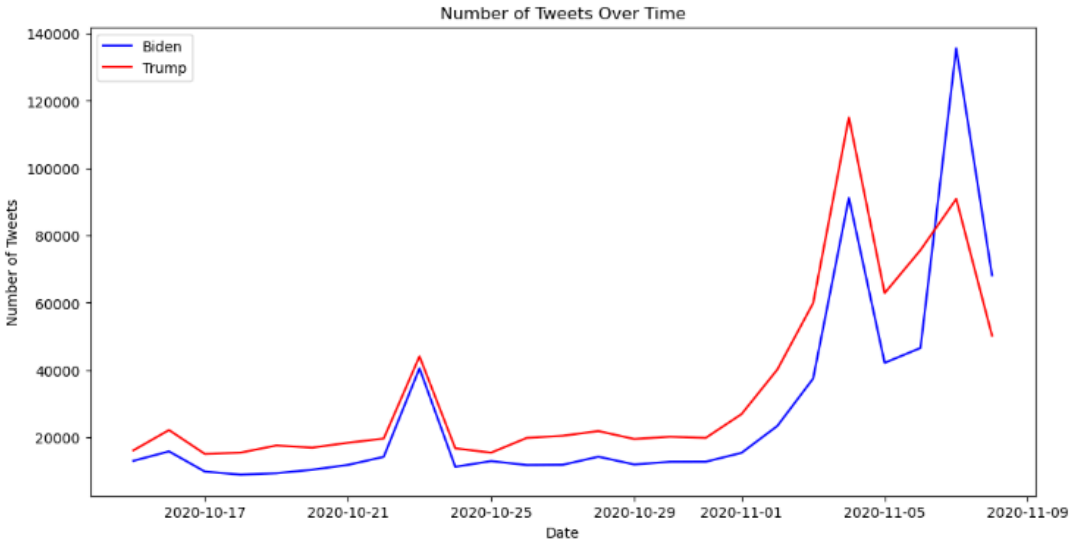
In addition to these new variables, we created a CSV file to store the cleaned and enhanced dataset, which is well-suited for our analysis. First and foremost, the distribution of sentiment scores is a critical starting point for our EDA as it provides a foundation of the overall sentiment within these tweets. In effect, we managed to get the following plots to visualize the spread of the sentiment:



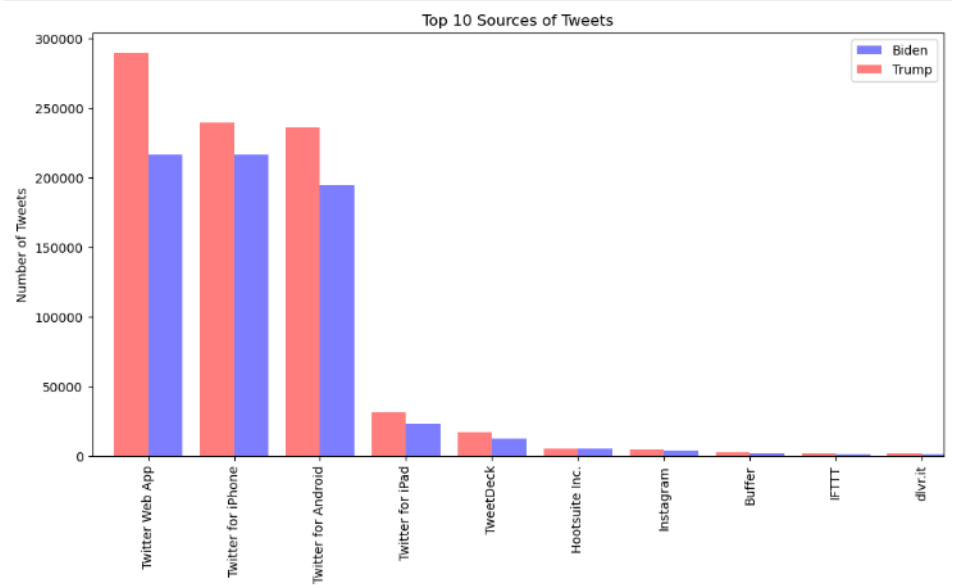


The above graphs show the counts of sentiment scores of Tweets and user descriptions rounded to the hundredths for both Biden and Trump. Of Biden's 694046 observations, 326,261 and 346,766 scored 2.5 exactly neutral for Tweet and User Description, respectively. For Trump, 868,762 observations, 381,572, and 440,372 were scored 2.5 for Tweet and User Description, respectively. Given that the data includes news and de-hashtagged names of the candidates, this amount of neutral sentiment is a reasonable number. Both candidates show more variable sentiment scores concerning the tweets, indicating that the tweets contain more polarizing sentiment than the user description, which is expected. The relationship between tweet sentiment and user description can be explored further.

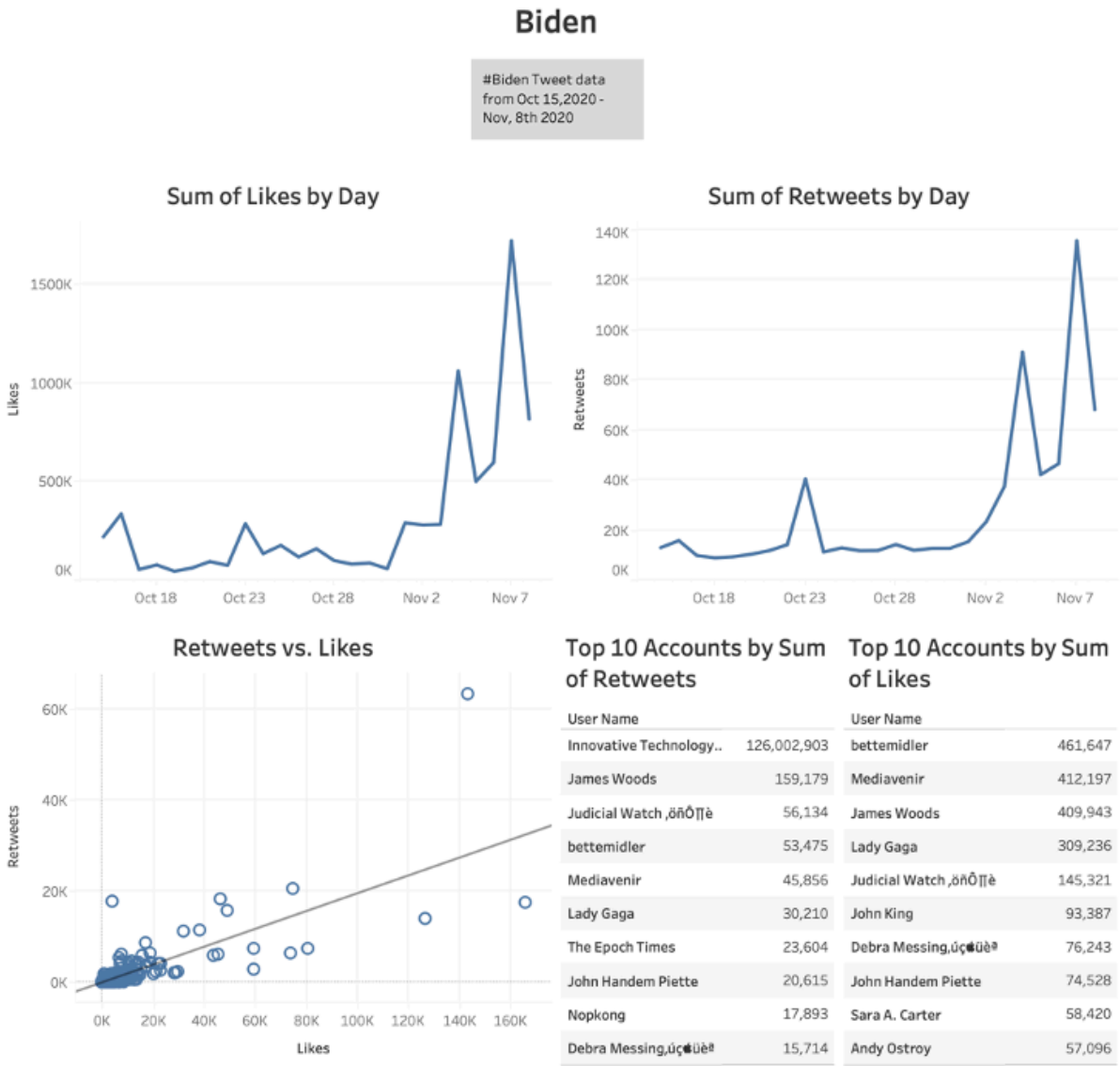
Given the reasonably varied sentiment distribution, analyzing the number of tweets over time using the Top 10 sources of tweets is important. By analyzing the number of tweets over time, a holistic view of the trends in public sentiment develops over specific intervals of time. For example, at first glance, tweets about Trump appear to overlap and peak in mid-October to November, while tweets about Trump go down in late November and tweets about Biden peak. These patterns could indicate a sentiment shift closer to when the actual election takes place.



A.



B.



These graphs summarize likes and retweet numbers leading up to and after the election on November 3rd, 2020. The two graphs are similar and share peaks between October 23rd and the five days following November 3rd. The election occurred on November 3rd, and October 23rd was the day of the final presidential debate.

Some accounts show up with the largest number of retweets and likes. Innovative Technology has ten times the number of retweets as the following closest account but does not make the top ten for likes.

The following values are the descriptive statistics for the Retweets vs. Likes scatterplot. The P-value suggests that there is a high likelihood that the two are correlated. Based on the equation, retweets increase by 0.195388 for every like. The y-intercept of 8.39 signifies that if there are 0 likes, there will still be an average of 8.39 retweets.

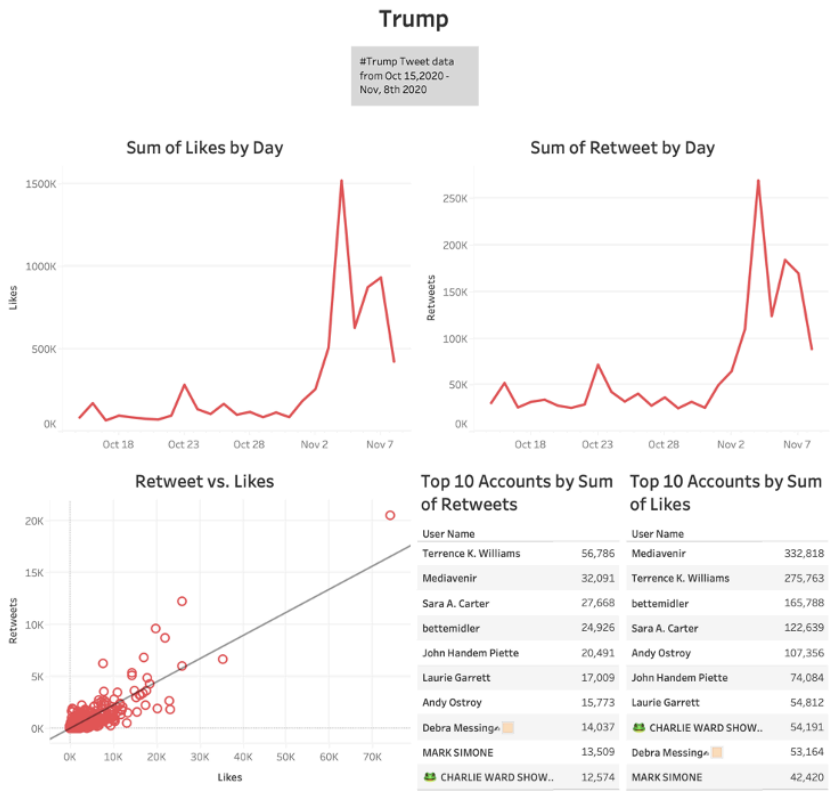
P-value: < 0.0001

Equation: Retweet Count = 0.195388*Likes + 8.39028

Coefficients

Term	Value	StdErr	t-value	p-value
Likes	0.195388	0.0014292	136.708	< 0.0001
intercept	8.39028	4.63542	1.81004	0.0703169

C.



The Trump data follows peaks in the same areas as Biden, but the second peak following Nov 3rd is significantly lower, attributed to his election loss. The p-value is also very low, showing a strong correlation with the likes coefficient similar to Biden's. The y-intercept is lower than Biden's at 2.05. The top 2 users in retweets and likes switch between 1 and 2 when comparing the two charts; otherwise, the top 10 charts show similar usernames differently

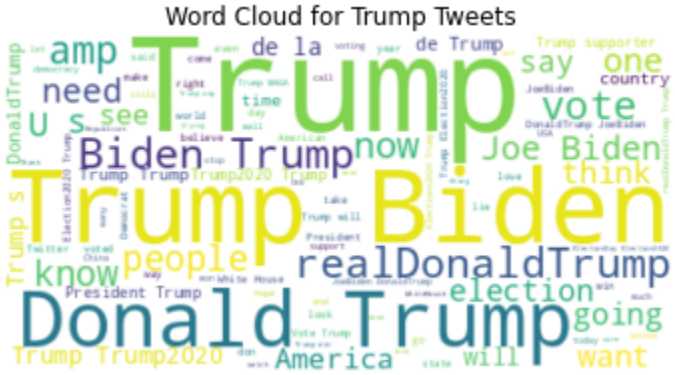
P-value: < 0.0001

Equation: Retweet Count = 0.222725*Likes + 2.05031

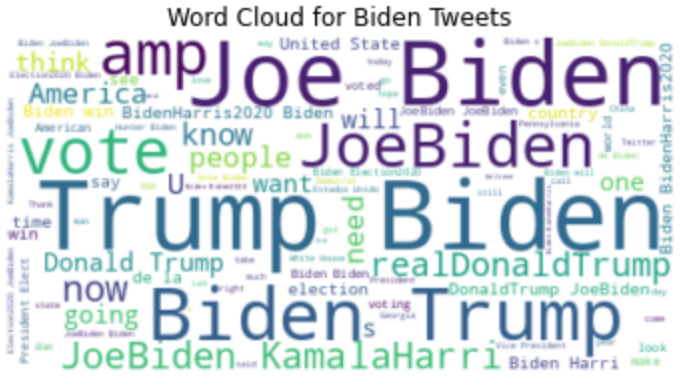
Coefficients

Term	Value	StdErr	t-value	p-value
Likes	0.222725	0.0011377	195.766	< 0.0001
intercept	2.05031	1.61146	1.27233	0.20328

D. TEXT ANALYSIS



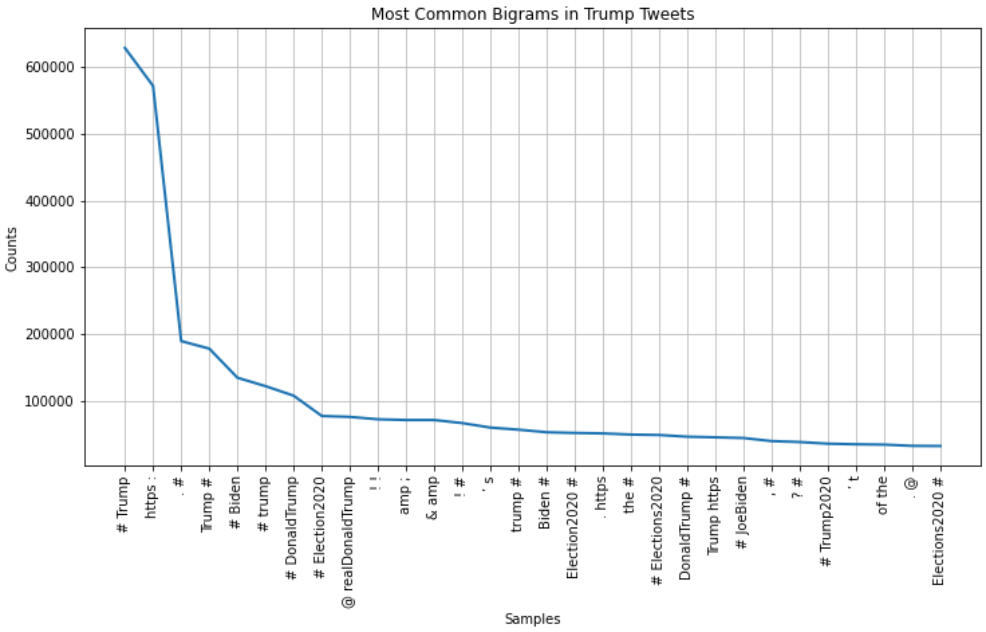
The most prominent words include "Trump," "Biden," "https," "Donald," "realDonaldTrump," "President," "election," and "America". This shows that these tweets frequently mention Trump and Biden and topics related to the election and America.



Similar to Trump's word cloud, prominent words include "Biden," "Trump," "https," "JoeBiden," "Vote," "Kamala Harris," and "America." This shows that the focus is also on the election, with significant mentions of Biden, voting, and related topics.

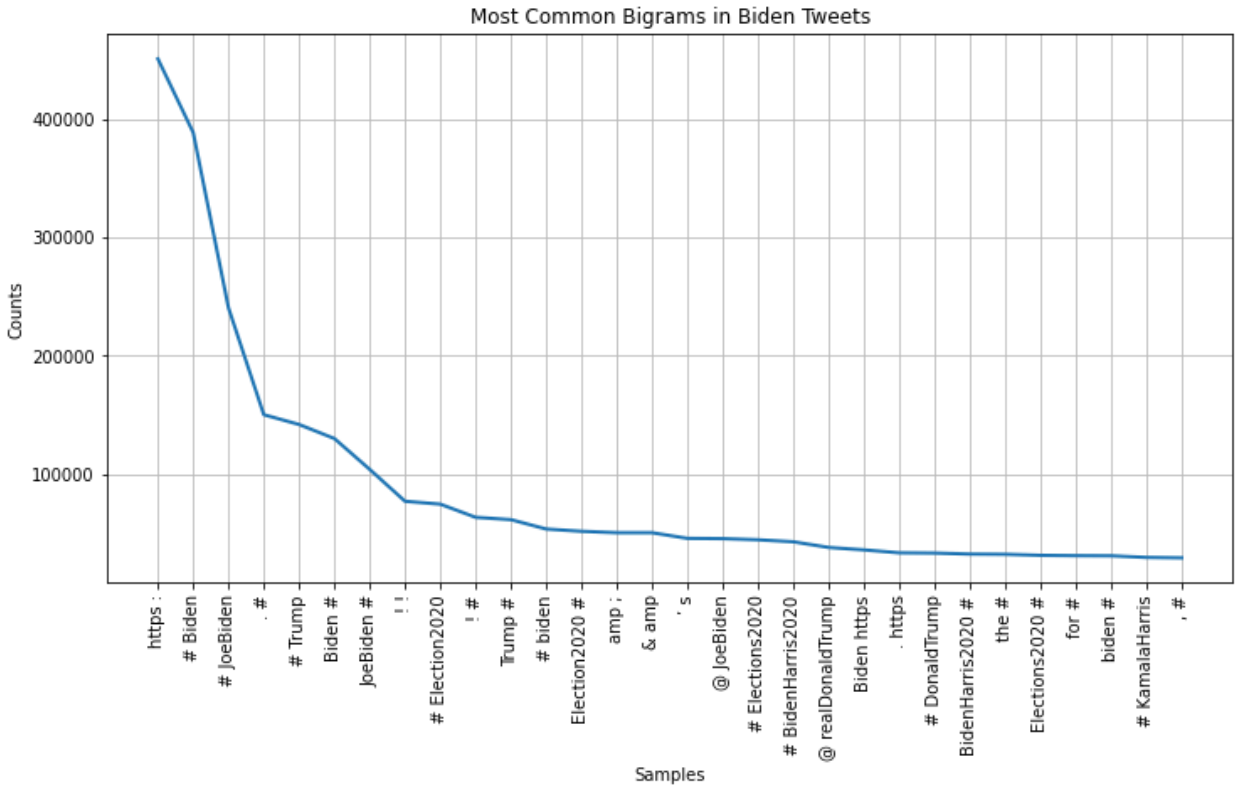
Both word clouds indicate a significant focus on the candidates themselves (Trump and Biden) and the election context.

E.



The follower count of users tweeting about Trump shows a sharp drop from the top followed users to the rest, showing that fewer users with very high follower counts contribute significantly

to the tweets. The distribution flattens out quickly, suggesting that the majority of users have lower follower counts.



Similar to the distribution of Trump's tweet followers, Biden's tweet followers also show a sharp initial drop. This means that very few influential users are followed by a large number of users with fewer followers.

For both Trump and Biden, the drop-off in follower count means a concentration of influence among a few high-follower accounts. These include official accounts, celebrities, government officials, and popular influencers.

Flattening the distribution curve means that a more extensive base of Twitter users with lower follower counts who are also actively tweeting about the candidates are ‘normal’ everyday people: people who live in America and people watching from other countries.

F. This analysis shows that many top influencers are verified accounts belonging to big media organizations and Lady Gaga.

```
Top Trump Influencers:
            user_name  user_followers_count  likes  retweet_c
ount  \
155994  CNN en Español            19115332.0  182.0
56.0
151931  CNN en Español            19115332.0  228.0
54.0
158668  CNN en Español            19115330.0  397.0          1
35.0
12536   CNN en Español            19108604.0  217.0
66.0
18856   CNN en Español            19108602.0  228.0
49.0
913945        detikcom            16296938.0   32.0
10.0
917283        detikcom            16296937.0   67.0
13.0
739330        detikcom            16272066.0   28.0
8.0
737139        detikcom            16272066.0   32.0
4.0
741108        detikcom            16272065.0   22.0
0.0

                                                    tweet
155994  A menos de 2 semanas de las elecciones, #JoeBi...
151931  A menos de 2 semanas de las elecciones, #JoeBi...
158668  A menos de 2 semanas de las elecciones, #JoeBi...
12536   .@CamiloCNN conversa con el actor, productor y...
18856   .@CamiloCNN conversa con el actor, productor y...
913945  Dalam pernyataan terbaru via Twitter, Trump ke...
917283  Pendukung Trump belum bisa menerima kenyataan ...
739330  Presiden Amerika Serikat (AS), Donald Trump, m...
737139  Hakim negara bagian Michigan dan Georgia menol...
741108  Kritikan terhadap Presiden Amerika Serikat (AS...
```

The top influencers for Trump tweets have accounts like “CNN en Espanol” and “detikcom.” These accounts have significant follower counts. The likes and retweet counts for these tweets are modest compared to their follower counts. This means that even though these accounts have a large following, their tweets about Trump might not generate high engagement. The tweets from these accounts focus more on news and updates related to the election and Trump’s activities. This seems to be more informational than opinionated.

```
Top Biden Influencers:
             user_name  user_followers_count     likes  retwe
et_count  \
287484        Lady Gaga            82417099.0   28146.0
2195.0
298231        Lady Gaga            82417077.0  126772.0
14024.0
278950        Lady Gaga            82396325.0   80670.0
7553.0
283057        Lady Gaga            82396310.0   73648.0
6438.0
711705   Alejandro Sanz            19793224.0    5498.0
684.0
1021549  CNN en Español            19157629.0     514.0
34.0
700228   CNN en Español            19154807.0     291.0
53.0
974598   CNN en Español            19154802.0     536.0
54.0
789667   CNN en Español            19154800.0    2191.0
206.0
976958   CNN en Español            19154800.0    2622.0
458.0

                                                    tweet
287484   Vote early! Vote Tuesday! Vote #Biden! #vote h...
298231   Good morning PENNSYLVANIA! I’m so excited to s...
278950   That’s a pic of me in Pennsylvania when I live...
283057   I AM SO EXCITED to be back in Pennsylvania! (P...
711705   #Election2020 #PresidentElectJoe #JoeBiden htt...
1021549  Desde #California a #NuevaYork, muchos están c...
700228   Momentos después de que los medios, incluyendo...
974598   Desde #California a #NuevaYork, muchos están c...
789667   Barack Obama reaccionó a la victoria de Biden ...
976958   Con el canto de "Tonight is going to be a good...
```

The top influencers for Biden's tweets include notable personalities like Lady Gaga and Alejandro Sanz and CNN en Espanol. Lady Gaga, in particular, stands out with a follower count exceeding 82 million. Her tweets show significantly higher engagement, with likes and retweets reaching thousands. This means that she has a strong influence and ability to drive engagement from followers. Lady Gaga's tweets are a mix of personal endorsements and campaign-related messages, and some directly encourage people to vote for Biden.

.

Methodology

RQ1 - How does daily sentiment analysis predict overall national polls, and what are the relationships between the two?

Correlation

To successfully attempt the daily sentiment to predict overall national polls, we needed to revolve around the sentiment analysis of both Biden's and Trump's tweets. We took a tweets dataset from Kaggle that had data shortly before and after the 2020 presidential election and focused on keywords and hashtags associated with the leading candidates. The decision was made to include hashtags because of the sentiment involved in hashtags and data from the bi and trigram analysis. Lots of data would be lost if hashtags were not included. The daily polling required to compare the sentiment scores needed to be scraped and aggregated from an online source.

Before proceeding with our analysis, we needed to prepare our data. This involved removing irrelevant information such as URLs, special characters, and emojis. These steps were crucial to avoid complications when analyzing our data in Jupyter Notebook or Tableau. We also decided to remove certain features, such as user_location, lat, and long, despite their potential usefulness due to insufficient data for efficient analysis. We added the sentiment analysis and tweet length columns to the dataset.

Our models needed a clear, robust, and nuanced sentiment scoring system. Initially, we considered a scale of positive, neutral, and negative, but this did not allow enough nuanced information to be passed to our model. A scoring system that contains a larger scale was more useful for complex analysis. Our scale was -3 to 3, rounded to the hundredths, which will provide enough information to our model for analysis.

Finally, we must conducted a correlation analysis between the average sentiment scores and national poll percentages and explore the relationship between these variables. Due to the complexity of the dataset, it was challenging to build predictive models other than conducting a correlation analysis.

Correlation Analysis:

-Pearson or Spark correlation

- Calculated the correlation between the average daily sentiment scores and national poll percentages. Involved statistical analysis such as Pearson to quantify the strength of relationships between the daily sentiment and polling data

-Visualizations

- Created a series of histograms, scatterplots, line graphs, and word cloud visuals to showcase the relationship between sentiment scores and national polls

-Time-series analysis

- Investigated how changes in sentiment score correspond to the changes in national polls over a period of time (3 weeks).

RQ2 - Can a machine learning model accurately predict the amount of retweets a tweet gets based on sentiment analysis? Which model most accurately predicts this?

Sentiment Analysis

The first step in analyzing RQ2 was to use our created sentiment score from RQ1, which could be used to analyze the most influential discussion on Trump and Biden. We used a Textblob and Vader and various preprocessing steps, such as removing stop words and stemming, to perform the sentiment on Q1. We then compared the two with different models to see which provided

better results. In addition to the general sentiment of the tweets, additional features such as user description sentiments and hashtags improve the potential accuracy of our analysis.

Through the EDA, we discovered various trends and insights that informed our understanding of tweets and their engagement and how they spread across other platforms. The most important steps taken in the EDA include:

Descriptive Statistics: Calculated summary statistics for key numerical fields, including likes, retweets, and sentiment scores. This provided an overview of the data distribution and helped identify outliers or anomalies. We also analyzed the distribution of sentiment scores for tweets mentioning Trump and Biden to understand the overall sentiment trends for each candidate.

Correlation Analysis: Calculated the correlation between sentiment scores and likes and retweets. This helped us understand the relationship between tweet sentiment and user engagement. We also examined the correlation between features like follower count and user activity to identify potential predictors for retweet counts.

Visualization:

- Time-Series Analysis: Visualized the sentiment scores and engagement metrics over time to observe how they fluctuated during the election period.
- Distribution Analysis: Created line diagrams and histograms to visualize the distribution of likes, retweets, and sentiment scores.
- Word Clouds: Generated word clouds to identify the most frequently used words in tweets mentioning Trump and Biden.

Modeling techniques:

To predict the number of retweets a tweet will receive based on sentiment analysis, we will use the following models:

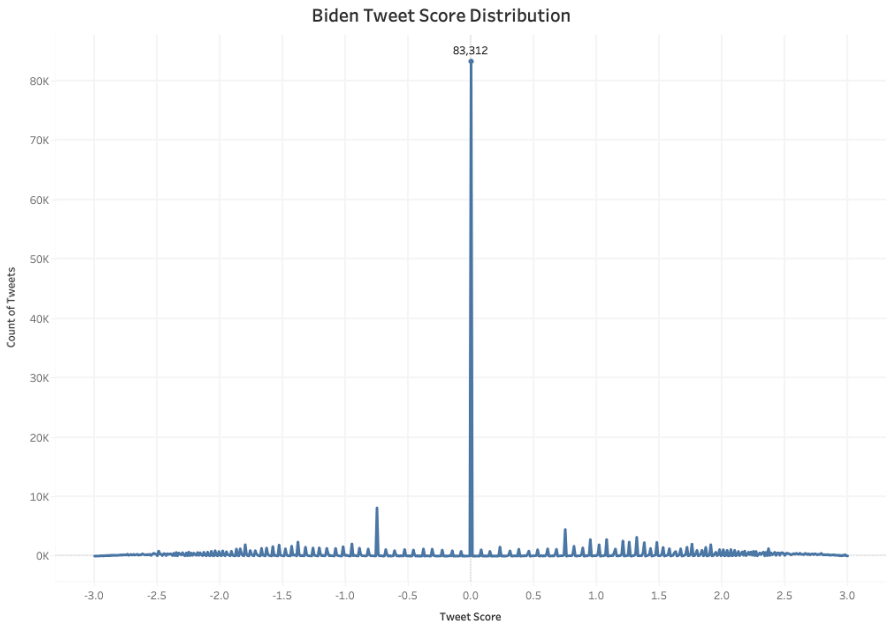
- Linear Regression: This model will help us predict retweet counts based on the features. Features like sentiment scores, follower counts, user activity, and the information in the tweet. The predictions will be compared to the actual retweet count for accuracy.
- Decision Tree: This model will show the non-linear relationship between the features and retweet counts, using the same features as Linear Regression.
- Random Forest: This model will improve prediction accuracy by combining multiple decision trees. We will construct a random forest with several trees, and each will be trained on a random subset of the data and its features.
- Support Vector Machines (SVM): This model will use regression to handle complex relationships in the data. We will use this model to capture non-linear patterns, giving us better prediction accuracy for our data.

Data Visualizations and Analysis

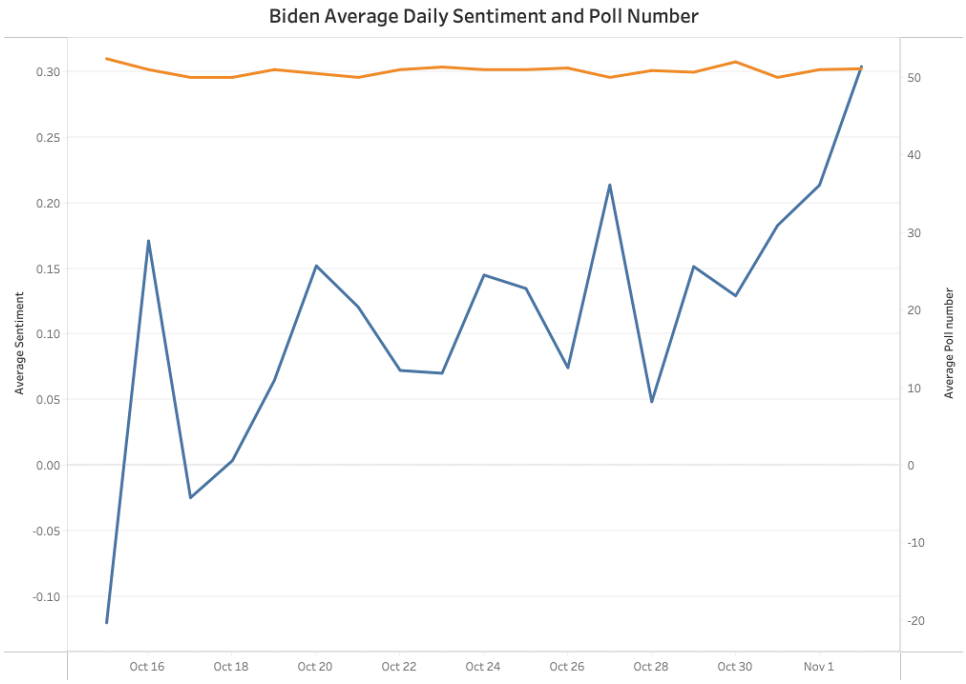
Correlation:

From the 2020 tweet dataset, we used essential features we created, such as sentiment, to test the distribution between Trump and Biden. We analyzed tweets to categorize them as positive, neutral, or negative. The x-axis is scored between -3, 0, and 3 from negative, neutral, and positive readings from left to right. This sentiment analysis was done using a customized VADER Sentiment Intensity Analyzer for each candidate which had added hashtags and phrases scored positively and negatively. Analysis was done on the most common bi and trigrams to see the frequency of usage and the scoring was done accordingly.

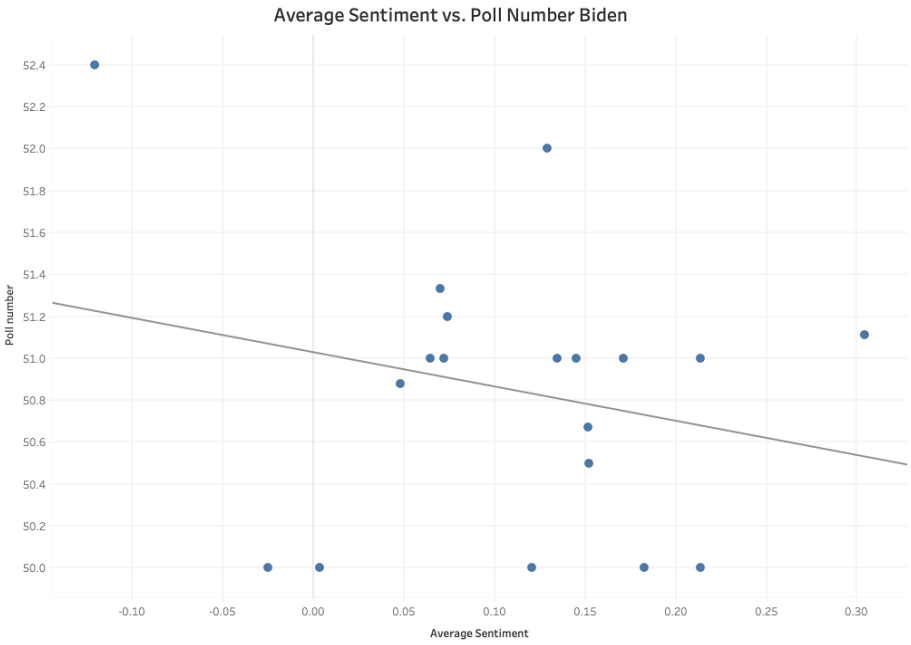
The average score of tweets was calculated. The blue color-coded graph shows a peak centered around 0 on the a-axis, with the most being neutral and having a count of 83,312 neutrally graded sentiment tweets down from 326,261 with just the standard VADER model. However, there are several small peaks around negative and positive sentiment for Biden from 5-10k on each side. A very similar shape to Donald Trump's sentiment.



This graph shows the distribution of tweet scores after customizing the VADER tweet sentiment model to include positive and negative bi and trigrams. Some tweet sentiment analysis projects remove hashtags from scoring models, but our model did not. A large amount of sentiment can be extracted from hashtags, so the decision was to leave them in and score the tweets on a scale of -3 to 3. When these steps were implemented, the Biden neutrals count dropped to 30% of the overall dataset being scored as neutral.



This is a dual-line graph of the days polling data was gathered. A limitation of this project was the few days in which polling data was available. The polling data was scraped from https://www.270towin.com/2020-polls-biden-trump/national/ aggregated daily polls, which were averaged for each day. Our dataset had an upward sentiment trend closer to election day, yet the polls stayed flat throughout.



Poll number = -1.63673*Average Sentiment + 51.0284

StdErr	1.64
t-value	-1
p-value	0.33
SSE	7.6
MSE	0.45
R-Squared	0.06

A linear regression was done on this scatter plot of Average Sentiment for the whole day's tweets and scraped Average Poll Number. The Y-intercept was 51.03, which is the baseline poll number for the equation. The slope of -1.64 indicates a negative correlation between sentiment and poll number. It has a high P value of 0.33, which conveys a weak correlation between the two variables—the standard error of 1.64 and t-value of -1 further support that this relationship is weak. An R^2 value of 0.06 indicates little variability in the poll number, which

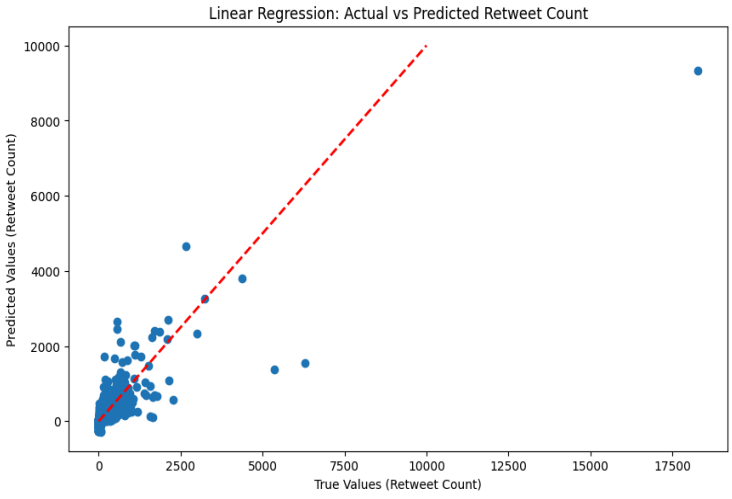
means that sentiment is not a good predictor for average polling. The MSE is relatively low because of a clustering of some of the data, but there are few outliers that make this value larger.

A -0.265 Pearson Correlation was calculated for the Biden data. This aligns with what was seen earlier with the linear regression equation: the relationship is a weak negative relationship.

RQ 2

- **Linear Regression Analysis**

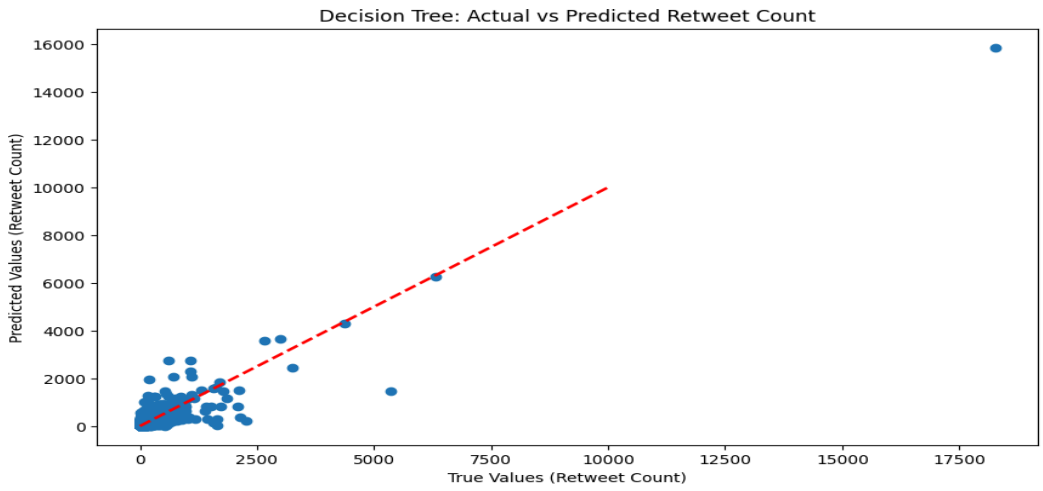
```
Linear Regression:
MSE: 636.6156081220882
MAE: 1.9904303312067035
R2: 0.6625059670905832
```



The graph above is a visual representation of Linear Regression in the form of a scatter plot. There is some dispersion for overall higher retweet counts. Visually, there is a clear trend for higher retweets, which means the model performs well when predicting tweet recounts. Very few outliers are present within the model. In addition, the MSE scoring metric represents that the squared prediction errors are relatively small on average, indicating a good fit for this model. With an R^2 of .662, there is a relatively strong relationship between the predictor on the Y-axis and the predictor on the X-axis. We can start evaluating other potential features with a reliable model to improve prediction accuracy.

- **Decision Tree Analysis**

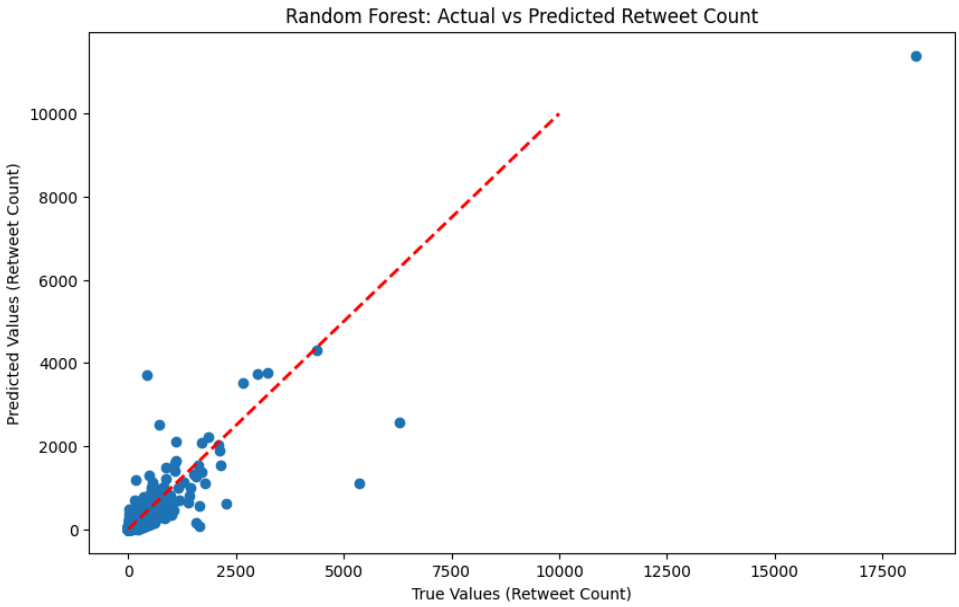
```
Decision Tree:
MSE: 312.40101821709925
MAE: 1.2586969579464697
R2: 0.8343843943221743
```



The decision tree model's MSE is significantly lower than the linear regression model's, meaning it has better predictive performance. The MAE is also lower than that of the linear regression, which means that, on average, the predictions in the decision tree model are closer to the actual retweet counts. The decision tree model outperforms the linear regression model regarding both MSE and MAE. The higher R^2 value means that the decision tree model better captures the relationship between the features and retweet counts.

- **Random Forest Analysis(3 graphs)**

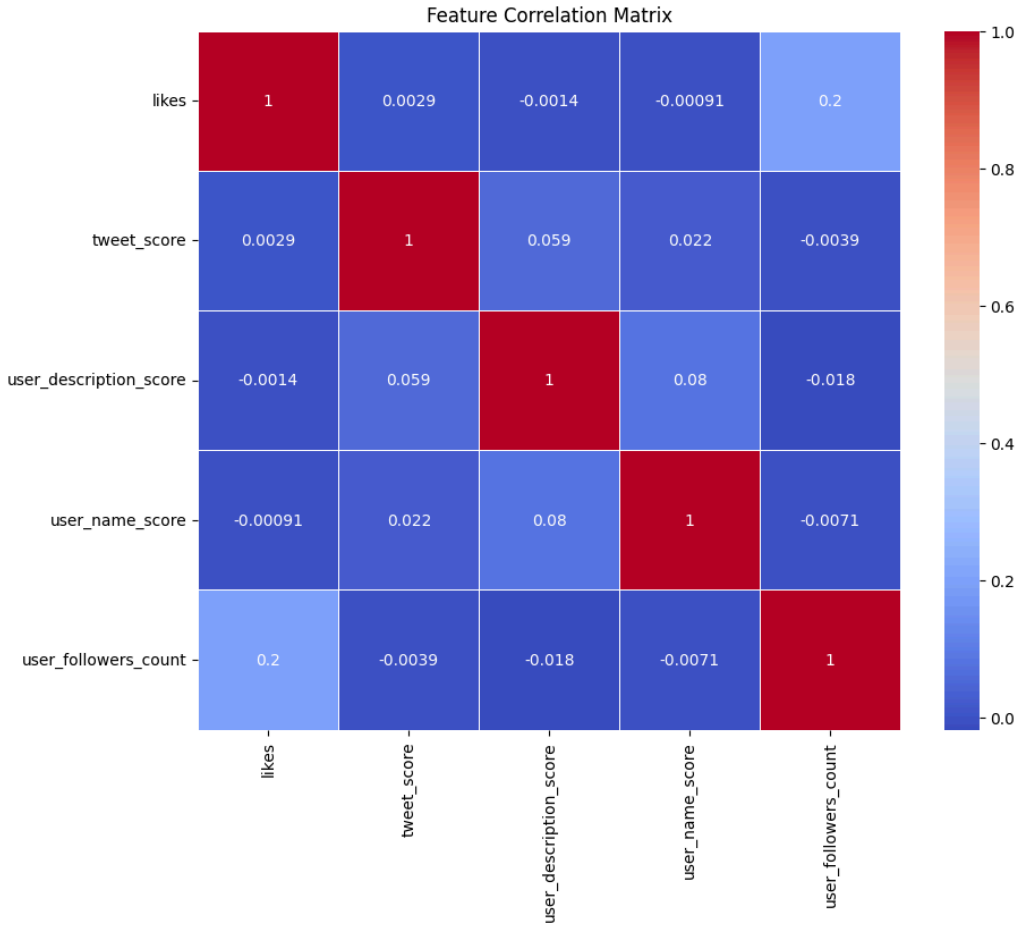
```
Random Forest:
MSE: 415.98154807656766
MAE: 1.1050389981668638
R2: 0.7794724344092119
```



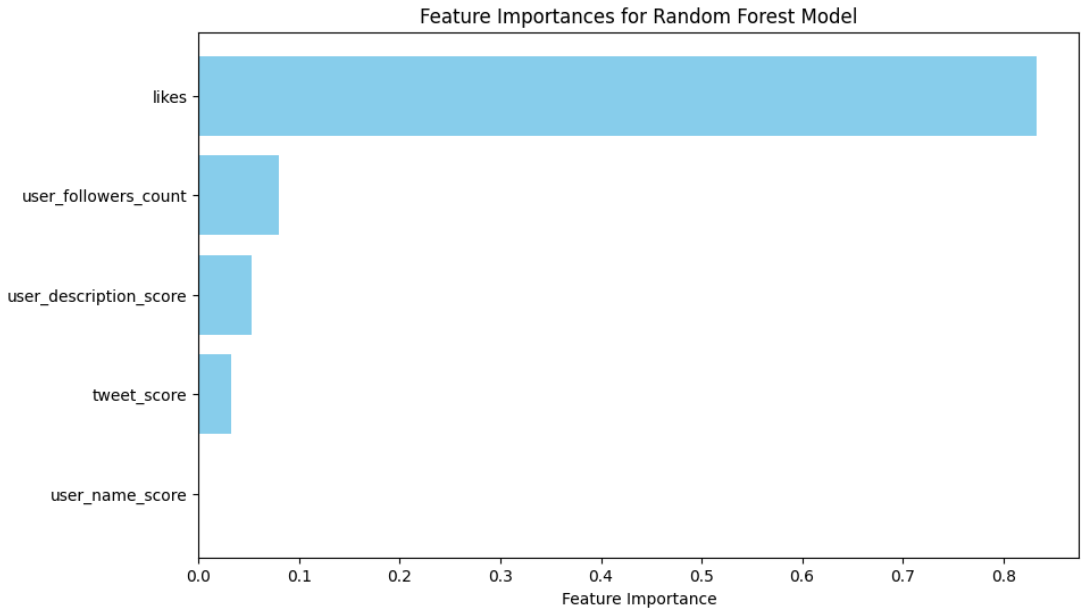
The Random Forest model's MSE is lower than that of the decision tree model (467.32) and the linear regression model (636.62), improving its performance. The MAE of 1.11 is the

lowest among the three models, showing that the Random Forest model provides the most accurate predictions on average. The R^2 value captures about 78% of the variance in retweet counts, which strongly fits the data.

The Random Forest model outperforms the Linear Regression and Decision Tree models in the MSE, MAE, and R^2 values. This shows that the Random Forest model gives us a more accurate and reliable prediction for our dataset.



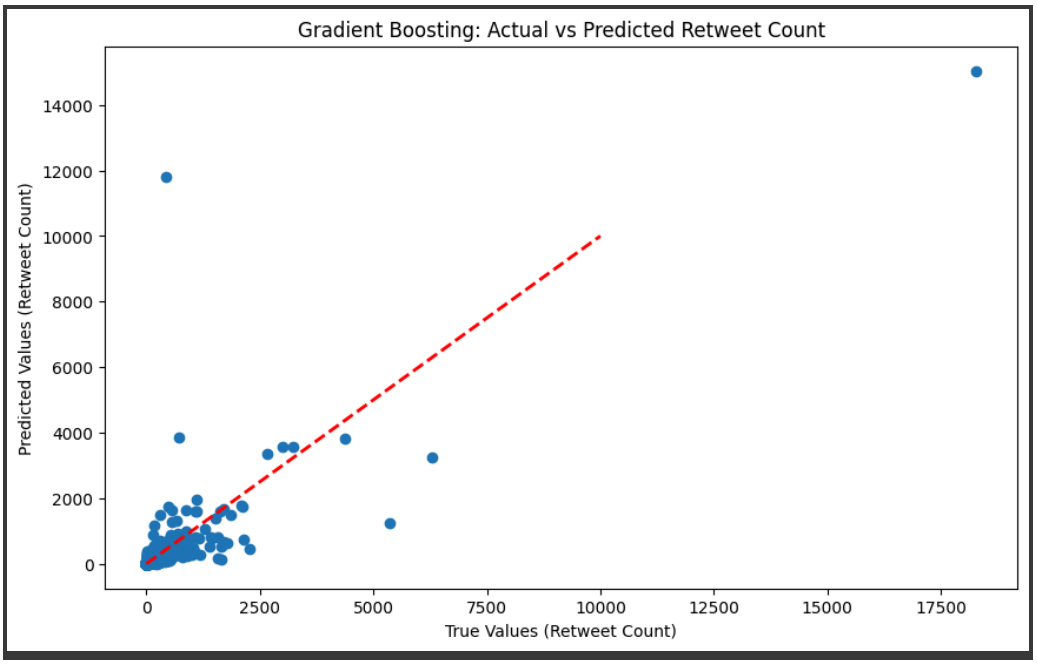
The feature correlation matrix shows that ‘likes’ and ‘use_followers_count’ have some correlation with retweet counts, whereas other features have very low or negligible correlation.



The feature importance plot for the Random Forest model showed that the number of 'likes' was the most influential factor in predicting retweet counts, followed by 'user_followers_count' and 'user_description_score.' Tweet sentiment scores and scores based on user descriptions have a lower impact on retweet counts.

- **Gradient Boosting Analysis:** *(Note: We decided to use Gradient Boosting instead of a Support Vector Machine because, in our regression context, SVM took significantly longer to train and predict, especially with our large dataset.)*

```
Gradient Boosting:
MSE: 730.7111809609236
MAE: 1.398495982659553
R2: 0.6126223419467112
```



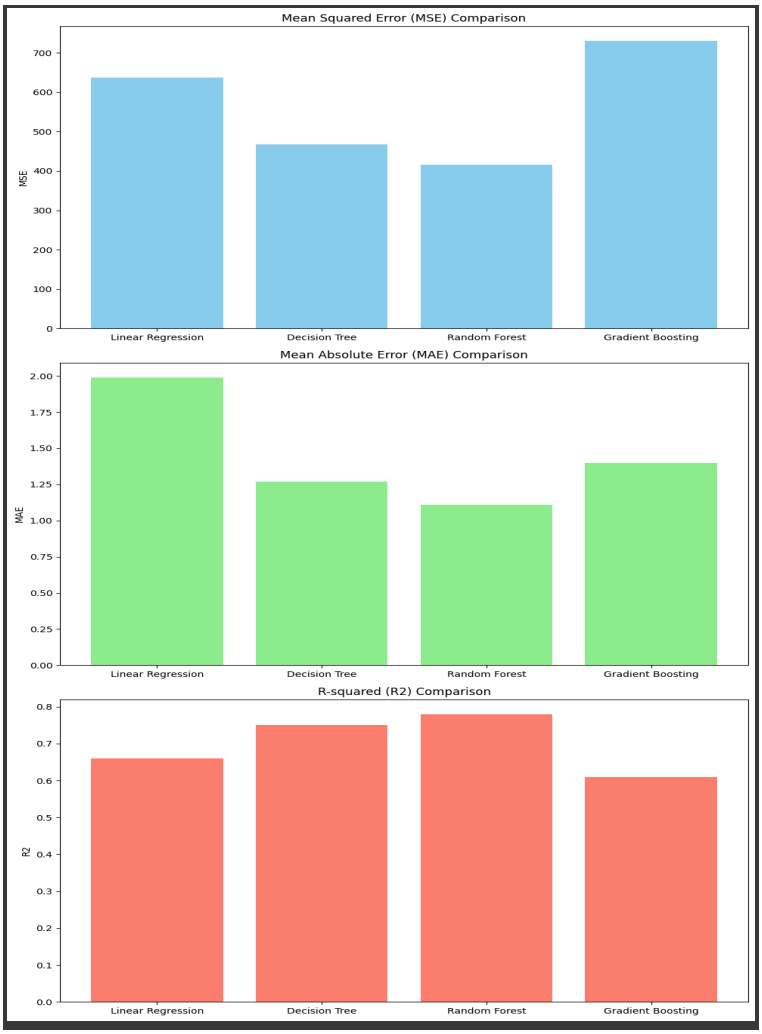
The MSE of the Gradient Boosting model is higher than that of the Linear Regression, Decision Tree, and Random Forest models, which gives us a less accurate predictive performance. The MAE of 1.40 is higher than the Random Forest and Decision Tree models but lower than the Linear Regression model, which means that the Gradient Boosting model is not as accurate in predicting as the Random Forest and Decision Tree models. The R^2 value only captured 61% of the variance in retweet counts, giving us a weaker fit in the dataset.

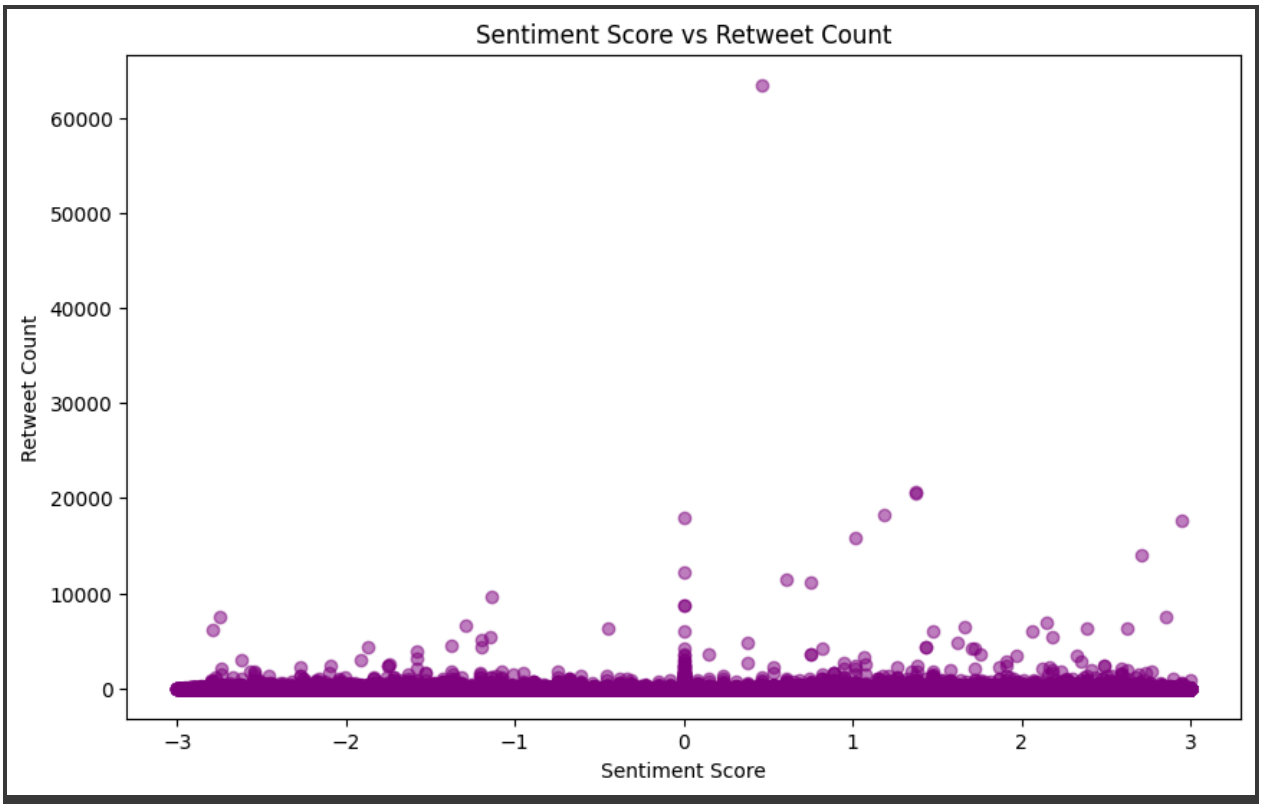
The gradient Boosting model underperforms compared to the Random Forest, Decision Tree, and Linear REgression models, which means that the Random Forest model is still the best performer.

Summary of Model Analysis

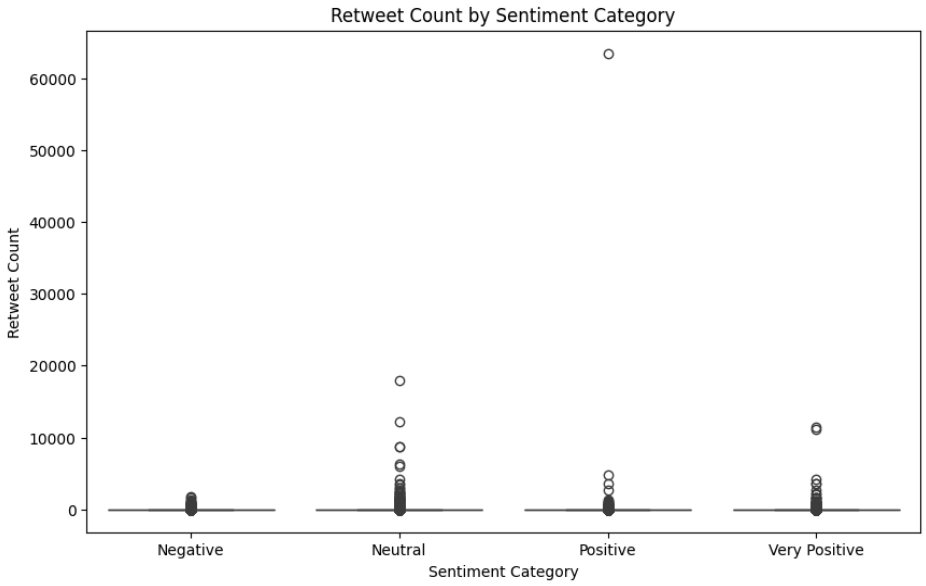
	MSE	MAE	R2
Linear Regression	636.615608	1.990430	0.662506
Decision Tree	312.401018	1.258697	0.834384
Random Forest	415.981548	1.105039	0.779472
Gradient Boosting	730.711181	1.398496	0.612622

- Random Forest is the best-performing model with higher accuracy and better generalization.
- Decision Tree is the second best because it captures non-linear relationships better than linear regression.
- Linear Regression shows a moderate predictive power but leaves significant room for improvement
- Gradient boosting underperformed other models, especially the random forest model.





The scatter plot shows the relationship between tweets' sentiment scores and the number of retweets they receive. There are some outliers, where tweets with high retweet counts have sentiment scores around zero, which means that these tweets might have gone viral for reasons other than sentiment.

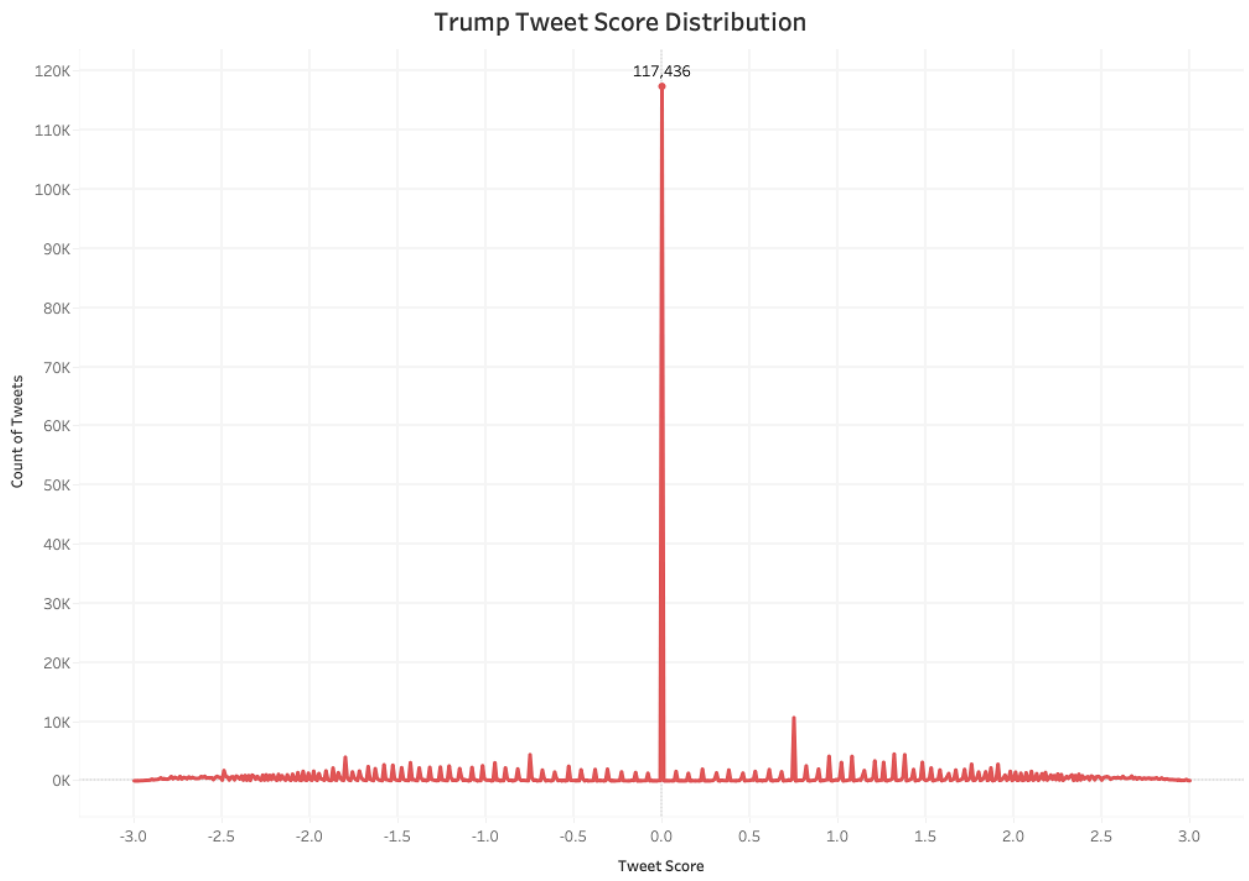


The box plot categorizes the tweets into different sentiment categories and shows the distribution of retweet counts within these categories. The median retweet counts are similar across sentiment categories, suggesting that sentiment alone is not a strong predictor of retweet count.

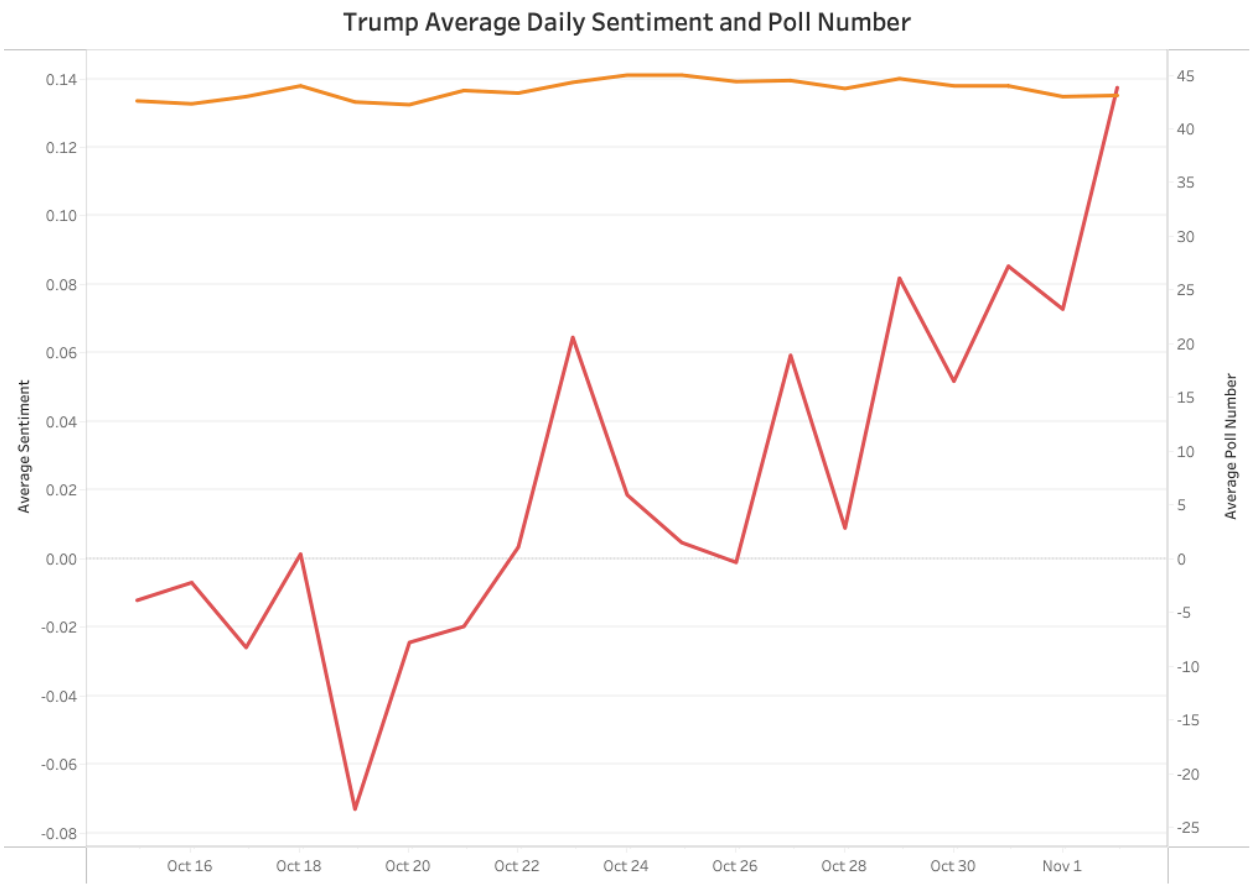
The conclusion from RQ2:

The analysis revealed significant information about tweet engagement and sentiment. The Random Forest model best-predicted retweet counts, showing the complex relationship between tweets and engagement metrics. It also suggests that sentiment plays a significant role in tweet engagement but is not a strong predictor of retweet counts, and the Random Forest model effectively predicts social media engagement.

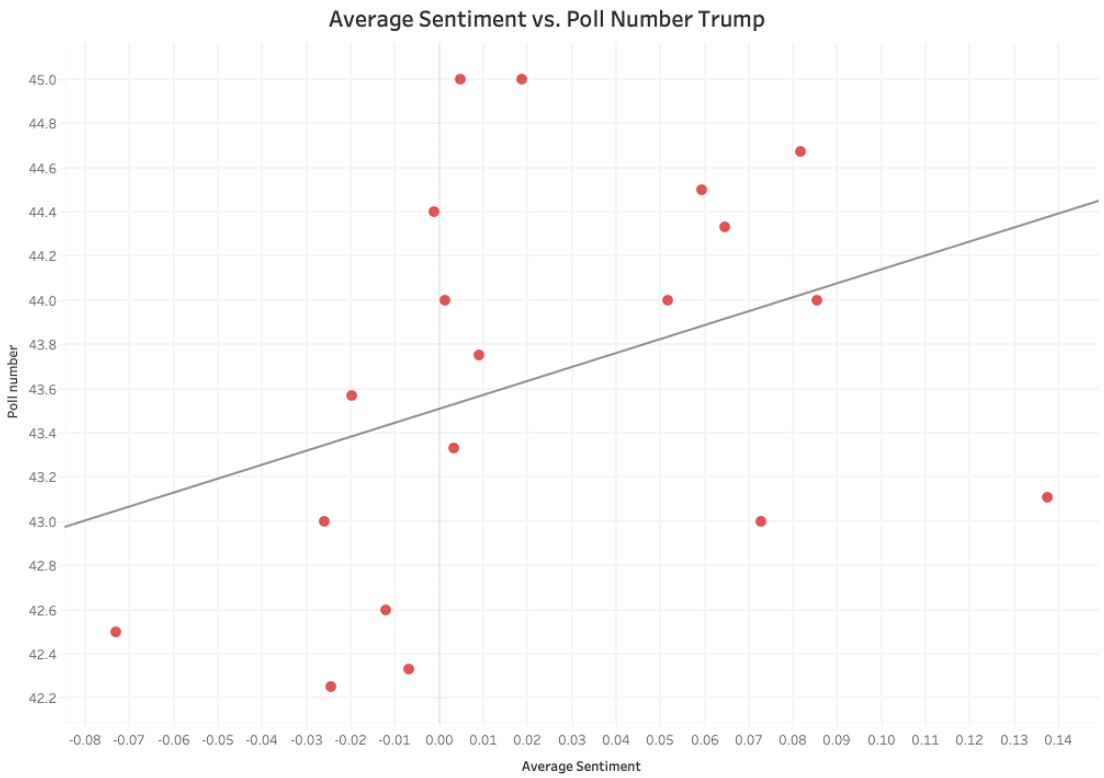
Overall, the final findings also show the importance of tweet content and user influence in driving engagement on social media platforms.



Just as with the Biden set, the Trump set of tweets were analyzed using a custom VADER analysis. After adding relevant bi and trigrams, 28% of the tweets remained neutral, down from about 50% before the custom scoring model.



As with the Biden analysis, the average poll numbers were flat, but averaging lower compared to Biden. The sentiment leading up to the election trended upwards, with the highest average score coming the day before the election. Comparing the two sentiment graphs, there was a significant fluctuation around October 22, 2020, the date of the last presidential debate.



Poll number = 6.31038*Average Sentiment + 43.5075

StdErr	0.85
t-value	1.6
p-value	0.13
SSE	12.33
MSE	0.73
R-Square d	0.13

54

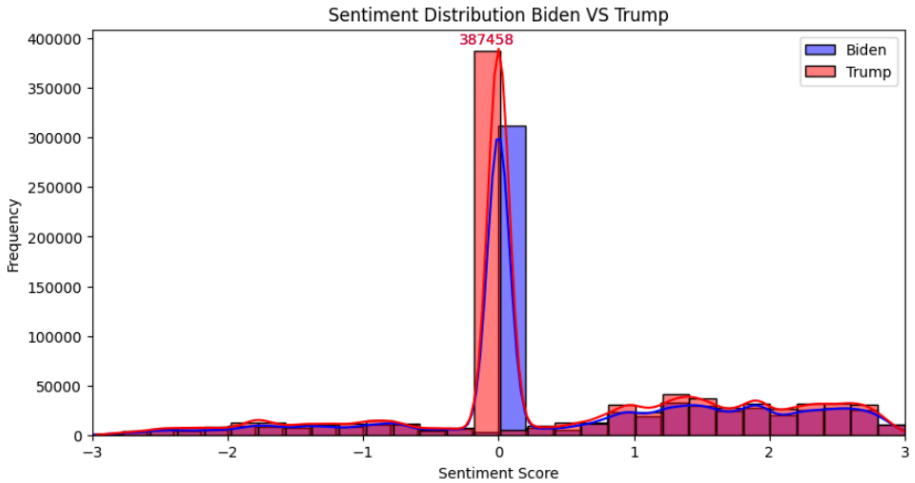
For the Trump tweets, the slope of the linear regression line is positive, and the y-intercept is lower at 43.51. The p-value is lower at 0.13 but still not statistically significant. 0.85 for the StdErr and 1.6 for the t-value indicate a weak positive correlation, and R^2 of 0.13, although better than the Biden model, is still not large enough to say that sentiment is a predictor

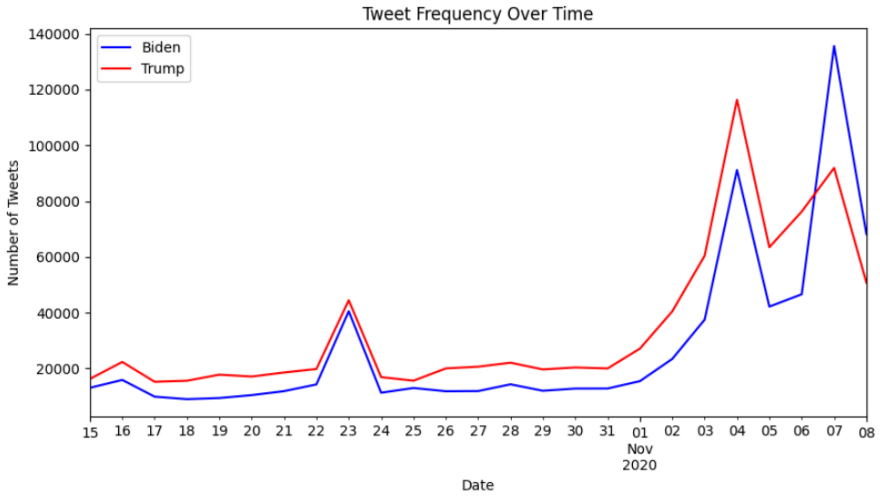
of poll number. The MSE is slightly higher compared to the Biden model, indicating a larger residual difference between the data points.

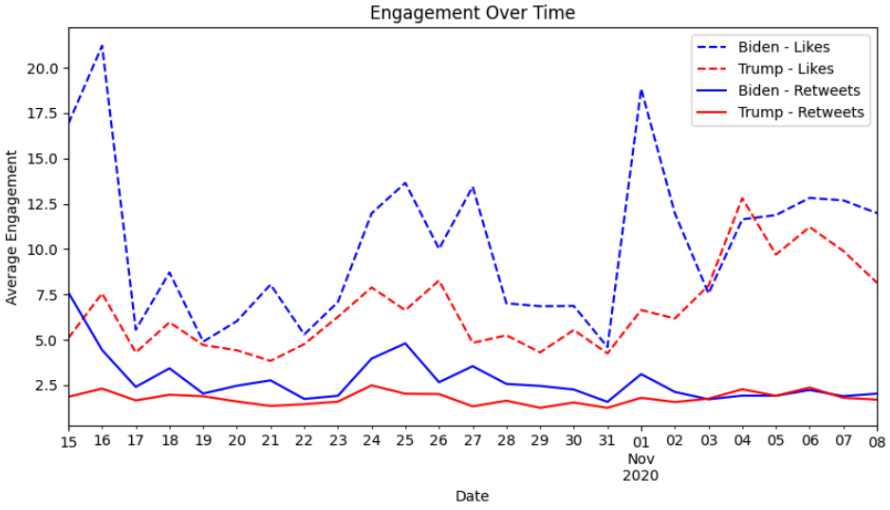
A 0.254 Pearson correlation was calculated for the Trump sentiment and average poll tweets which is about the same as the Biden one just positive. It's a weak positive relationship between average calculated sentiment and poll number.

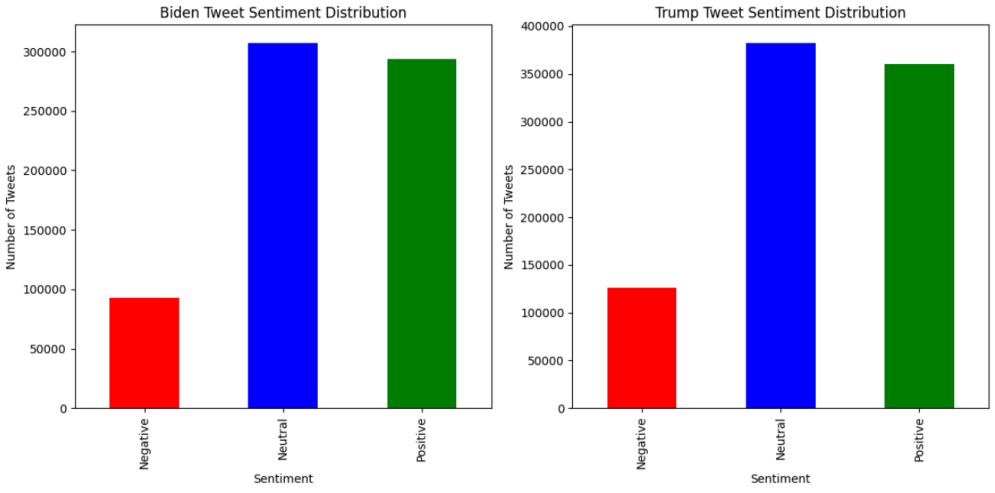
Logically, a higher sentiment score should be coupled with a higher poll number. In the beginning of our project, we aimed to have a location based scoring model and predict the poll numbers based on this. Our data is from all of Twitter and relating overall Twitter sentiment to poll numbers never was going to be a perfect correlation. Users on Twitter have different patterns of posting and some make use of hashtagging more than others which may have affected the sentiment score for the tweets. Furthermore, the scoring model of the VADER sentiment was custom for each candidate and may not have accurately represented the large amount of data that was present.

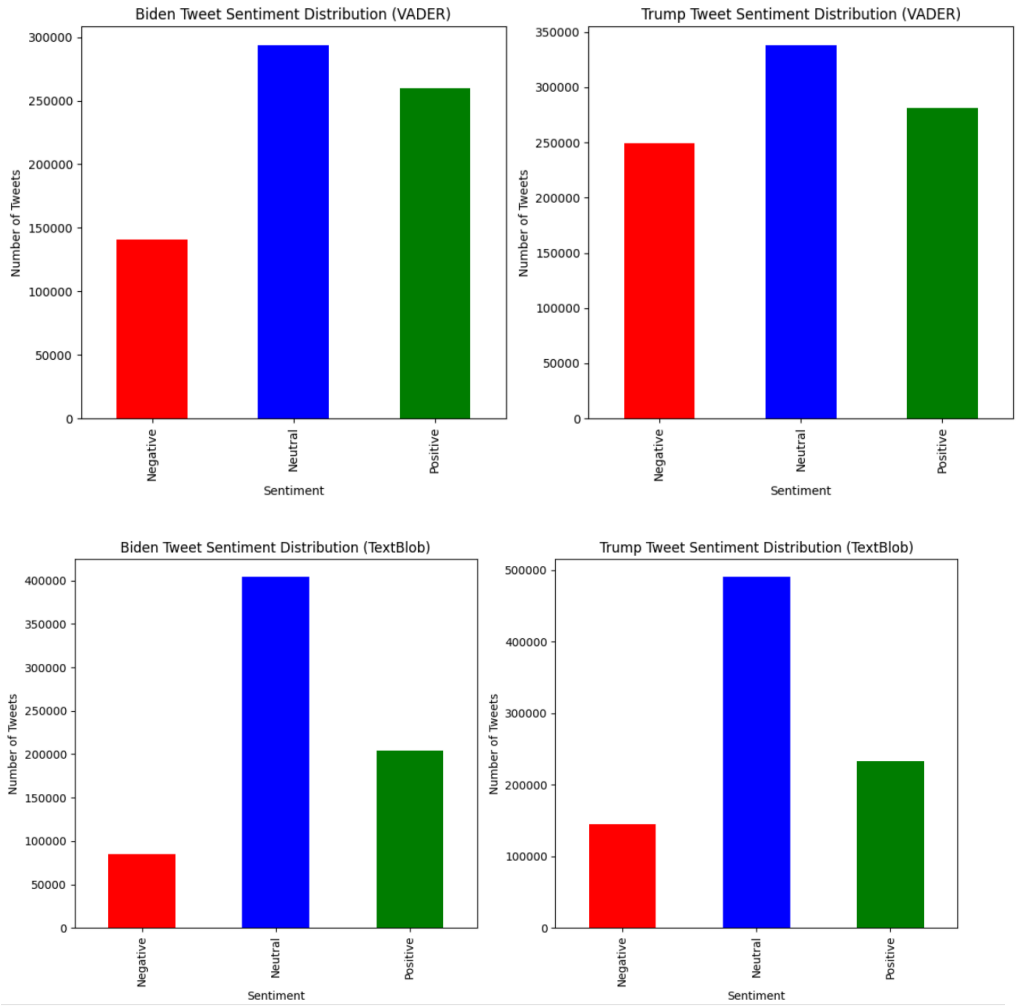
Finally, we made some merged icons for intuitive comparison.




Tweet Frequency Over Time
Biden
Trump
Number of Tweets
140000
120000
100000
80000
60000
40000
20000
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 01 02 03 04 05 06 07 08
Nov
2020
Date


Engagement Over Time
Biden - Likes
Trump - Likes
Biden - Retweets
Trump - Retweets
Average Engagement
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 01 02 03 04 05 06 07 08
Nov
2020
Date


Biden Tweet Sentiment Distribution
Number of Tweets
300000
250000
200000
150000
100000
50000
0
Negative
Neutral
Positive
Sentiment
Trump Tweet Sentiment Distribution
Number of Tweets
400000
350000
300000
250000
200000
150000
100000
50000
0
Negative
Neutral
Positive
Sentiment



The sentiment analysis of tweets related to the 2020 election reveals that most tweets for both Biden and Trump were neutral, with fewer showing positive or negative sentiment. This neutrality suggests a diverse public opinion without strong polarization towards either candidate. The tweet frequency over time highlights significant spikes around key dates, indicating increased public engagement during critical events. The engagement analysis shows that while both candidates' tweets received substantial likes and retweets, engagement depended on factors beyond sentiment, such as tweet content and user influence. These findings underscore the complexity of predicting electoral outcomes based solely on sentiment analysis and highlight the importance of considering multiple variables to understand public opinion and tweet engagement

accurately. Additionally, the data was divided into three parts (negative, neutral, and positive) to understand better the distribution of sentiments and their impact on tweet engagement.

Ethical Recommendations

Delving into sentiment analysis on a specific dataset of tweets from 2020 opens up a world of ethical possibilities, particularly in our predictions of national polls and retweets. Sentiment analysis's inherent vulnerability to manipulation is a risk that cannot be ignored. In tandem with its integration with machine learning, the responsible use of sentiment analysis is paramount in this context.

Regarding RQ1, using the breakdown of sentiment analysis to find overall average sentiment analysis to predict the presidential elections poses its own unique set of ethics. Sentiment analysis using the national poll may exhibit bias due to the nature of the text from the dataset. If too many are neutral, we must remove or re-classify these texts. Sometimes, sentiment analysis does not consider the tweet's sarcasm and context. This ambiguity can, in turn, lead to skewed predictions and misrepresentation of the data itself. It is essential to ensure that these texts within the data are classified correctly, whether negative or positive, to avoid an abnormal amount of neutral sentiment. There also comes into question the term "national." The dataset comprises three weeks' worth of data. However, that might not account for changes within one national polling within the same three-week period. Since we are breaking down the data and aggregating it, finding a well-suited national dataset to correlate with can raise issues with the authenticity of our results. It is essential to maintain consistent data to maintain consistent results. Therefore, regularly assessing the data is crucial to maintain a fair and accurate sentiment analysis. In addition, it is also essential to mention privacy. Kaggle obtained the data, but that does not take away from the fact that it includes geolocation data and personal tweets. While our analysis does not include geolocation data, they rely heavily on these personal tweets,

such as the wording and content- which is where our graded sentiment comes from. It is essential to be transparent about where this data comes from and comply with privacy regulations.

Regarding RQ2, further research is necessary to analyze the relationship between tweets and their behavior or sentiment. However, tweets can perpetuate stereotypes and impact the accuracy of the dataset. The analysis of the influence and sentiment analysis of retweets can also present ethical considerations. Focusing on the most influential users and their content can skew or manipulate the data and our potential models. Remembering that influential users can represent specific views is crucial, as well as the importance of retaining data from a more diverse range of tweets. Analyzing hashtags can also pose issues, revealing biased or "neutral" graded sentiments. Some may argue that focusing purely on influential texts and a base sentiment analysis provides a clear enough picture. However, learning from a diverse range of graded texts and considering other possible factors is also good. Addressing these issues, such as bias within retweets and users, authenticity of sentiment, and transparency, is essential for a consistent graded sentiment and modeling accuracy. By addressing these concerns, we can ensure that sentiment analysis provides a meaningful and influential impact on society by maintaining the integrity of the dataset.

Challenges

During our analysis of tweet sentiments and their influence on polling, there were several challenges that presented themselves. No dataset is ever perfect and a large portion of time is dedicated to preparing the data in order for analysis. Upon assessing the steps that were necessary to clean and prepare the data, the question of whether or not the data was suitable for our research questions was discussed. At this point we changed one of our questions to one that was more suitable for the cleaned data we produced. Getting clear, concise, and measurable questions was one of the project's most important parts and getting this correct was important to the success and direction of the entire project.

At first, we wanted to use the user location data to compare the state's sentiment toward candidates. However, upon further inspection and data wrangling, it was clear that removing data containing nulls in the 'user location' column would delete more than half the data, which would limit our ability to analyze a large enough dataset. An answer to this problem would have been to add more data from other sources, but in order to access this data and combine it with our first dataset seemed unrealistic given the time frame we were working with.

Thus, we devised a question that would drill down into the data set and evaluate the tweets daily within a 2.5 week interval. By comparing average daily sentiment to scraped average national poll data , we were able to make a comparison between the two measures which were analyzed with various statistical methods. Finding daily national polling data was challenging because most polling is done weekly or frequencies other than daily. With our project we needed a daily poll and a website was found that aggregated polling data from various polling organizations. Given the short time period of data, the average polling data for each candidate did not change very much: a point or less each day. The amount of tweets resulted in a

large number of observations, but the polling data measured was fewer than 20 average poll number days.

Using Python and Jupyter Notebook, we added two additional sentiment analysis columns to the data set and removed the 'user_location' column. One sentiment was based on user description, and the other was based on the tweet sentiment. However, another complication arose: determining an appropriate scoring method. VADER scores the text with a positive, negative, neutral, and compound score. The compound score is the sum of these normalized between -1 and 1, 0 being neutral. In order to gain a more nuanced scoring model, we adjusted these scores by a scale of 3 to capture a more holistic representation of the data.

The scoring model for VADER is based on a predefined lexicon that is suitable for social media applications. This is the reason we chose this modeling instead of the other options. In our Twitter analysis, however, much of the sentiment is contained in unique hashtagging of bi and trigrams. For example “SleepyJoeBiden”, “HunterBidensEmails”, “DumpTrump”, and “Maga”. These bi and trigrams aren’t a part of the default VADER lexicon and challenges arose with the analysis that had to be done to find them and score them appropriately for each candidate. In what context is it positive or negative? How should they all be scored? After combing through bi grams and trigrams and their usage in the dataset, a simple score of -1 and 1 was given to these negative and positive combinations.

With a newly refined scale and a cleaned dataset, we created visualizations to interpret the distribution of sentiment scores. Despite this, we faced another challenge: the high prevalence of neutral sentiment scores. The dominance of neutral sentiment was high, which made it a bit challenging to analyze positive and negative sentiment scores. Implementing the custom scoring, however, cut the percentage of neutrals down ~30% for each of the two

datasets. This resulted in datasets that more accurately reflected the sentiment of the tweets. Regarding the visualizations, they were skewed to the middle, which is where our neutral scoring was. As a result, it was challenging to compare the differences between the candidate's polling results since a significant portion of the tweets were classified as neutral. However, unless the neutral scoring surpassed 70-99%, it didn't pose much of an issue to analyze. There were still a large portion of data classified as positive and negative. In addition, if we took out most of the neutral data, it would’ve posed other complications discussed in the ethics section.

Since our tweets were from around the world, the language aspect proved to be one of the unsolved issues we ran across. Each non-english tweet was scored neutral. Attempts to try and write code to detect and translate each foreign language tweet were unsuccessful. The processing power inorder to process these 1M+ tweets as outlined required migration of the dataset into a cloud service platform and usage of cloud based servers. Furthermore, when this was attempted problems occurred with user permissions and just the code in general. More experience and research needs to be done in order to complete these steps and the time frame did not allow for this to happen.

Recommendations and Next Steps

Based on the challenges, results, and effectiveness of the dataset, we will recommend that:

- Given the diversity of tweets in multiple languages, using advanced natural language processing techniques to accurately analyze sentiment across different languages can provide more comprehensive insights.
- Adding tweet length, media presence, and comments can improve the model's predictive accuracy.
- Combining different machine learning models or including deep learning techniques can capture complex patterns and improve prediction performance.
- Choosing a different research question instead of polling number relationships. Interesting as it is, this only really works with a data set that has full and accurate location data .
- A different research question may be comparing different custom VADER models to include hashtagging and not hashtagging or different lexicons and exploring how these can affect the prediction of likes and retweets .

Next Steps:

- Include data from other social media platforms to provide a more holistic view of social media engagement.
- Learning more about cloud computing, permissions, provisioning instances, and machine learning on cloud platforms in order to process the multiple languages that were present in our dataset.

- More research on the different sentiment analysis tools that maybe have been used instead of the ones that were.

References

Ali, R.H., Pinto, G., Lawrie, E. et al. A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election. J Big Data 9, 79 (2022). https://doi.org/10.1186/s40537-022-00633-z

Sulistyo Dwi Sancoko, Saucha Diwandari, Muhammad Fachrie, "Ensemble Learning for Sentiment Analysis on Twitter Data Related to Covid-19 Preventions", 2022 International Conference on Information Technology Research and Innovation (ICITRI), pp.89-94, 2022.

Yaqub, Ussama, et al. "Location-Based Sentiment Analyses and Visualization of Twitter Election Data." *Digital Government (New York, N.Y. Online)*, vol. 1, no. 2, 2020, pp. 1–19, https://doi.org/10.1145/3339909.

Rizk R, Rizk D, Rizk F, Hsu S. 280 characters to the White House: predicting 2020 U.S. presidential elections from twitter data. Comput Math Organ Theory. 2023 Mar 28:1-28. doi: 10.1007/s10588-023-09376-5. Epub ahead of print. PMID: 37360912; PMCID: PMC10042672.

Adedeji, A. (n.d.). Sentiment Analysis for Amazon Products and Services. Sentiment Analysis for Amazon Products and Services. Adedeji https://www.linkedin.com/pulse/sentiment-analysis-amazon-products-services-aderemi-adedeji/

Zhang, Yazhou, et al. "A Quantum-Inspired Sentiment Representation Model for Twitter Sentiment Analysis." *Applied Intelligence (Dordrecht, Netherlands)*, vol. 49, no. 8, 2019, pp. 3093–108, https://doi.org/10.1007/s10489-019-01441-4.

Minty, D. (2023) Sentiment analysis – treat with a lot of caution, Ethics and Insurance. Available at: https://www.ethicsandinsurance.info/sentiment-analysis/ (Accessed: 16 June 2024).

Ali, R.H. et al. (2022) A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election - journal of big data, SpringerOpen. Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00633-z (Accessed: 16 June 2024).

https://www.270towin.com/2020-polls-biden-trump/national/(Accessed: 1 June 2024.)

Appendix

Main notebook

https://colab.research.google.com/drive/1sywyGe-9ytur6ZkudFUpWEebidznIfY-?usp=sharing

RQ2

https://colab.research.google.com/drive/111cQzssSrawGEboOX7jUHxcCeskMQeXL?usp=sharing

ViZ

https://colab.research.google.com/drive/114Ad95Bo8ub6-8b7q2fX2Z5WSw5nQgnB?usp=sharing

https://colab.research.google.com/drive/1HoqZTp1-U4ia5dywq71QqzSJF71xjyGq?usp=sharing

https://colab.research.google.com/drive/1esHxczRTfyln5IHqoOG6LrLV8F-bflff?usp=sharing