Twitter Sentiment Analysis:
Understanding Political Trends and Voter Sentiment

Sai Keerthi Reddy Mora

Inamullah Mohammad

Sai Uday Reddy Bhumireddy

Supervised by Dr. Theodore Trafalis

DSA 5900, Fall 2024 (4 Credit Hours)

Dated. 11 October, 2024

Contents

1 Introduction **3**

2 Objective **3**

3 Data **4**

3.1 Data Ingestion 4

3.1.1 Data Source 4

3.1.2 Data Collection 4

3.2 Data Exploration 5

3.2.1 Distribution of Tweets Over Swing States 5

3.2.2 Distribution of Tweets Over Time 6

3.2.3 Distribution of Tweets by State and Party 6

3.2.4 Engagement Metrics Over Time 8

3.2.5 Average Engagement Metrics by Party 9

3.2.6 Tweets Across User Handles or Hashtags 10

3.2.7 Word Cloud Analysis of Tweets Data 11

3.3 Data Preparation 12

4 Methodology **12**

4.1 Techniques 12

4.1.1 Transformer-based Sentiment Analysis and Large Language Models (LLMs) . 12

4.1.2 BERTweet 13

4.1.3 RoBERTa (Robustly Optimized BERT Approach) 13

4.1.4 DistilBERT 13

4.1.5 GPT (Generative Pre-trained Transformer) 13

4.1.6 Fine-tuning Techniques 14

4.2 Process Validation 14

4.2.1 Sentiment Trend Analysis 14

4.2.2 Manual Annotation for a Subset 14
4.2.3 Election Results as a Validation Metric 14

1 Introduction

Social media has revolutionized the way people communicate, especially in the context of political discussions. Platforms like Twitter provide a space where individuals freely express their opinions on current events, including political candidates and elections. With millions of tweets generated daily, Twitter offers an immense amount of data that can be analyzed to capture the pulse of public opinion.

This project focuses on using sentiment analysis to interpret public sentiment related to the upcoming presidential election. By employing advanced machine learning models, such as Large Language Models (LLMs) and VADER, we aim to classify the sentiment in tweets as positive, negative, or neutral. Transformer-based models like GPT-4, BERT, RoBERTa, and BERTweet, pre-trained on vast corpora of text, are utilized to handle the complexity of Twitter data, which often includes informal language, abbreviations, and unique formatting.

To tackle the challenge of unlabelled data, we implement fine-tuning techniques, including zero-shot and few-shot learning, which enable these models to generalize and perform sentiment analysis with minimal manual intervention. Our approach is designed to capture the nuances of political sentiment on Twitter, offering a detailed analysis of how various candidates are perceived in the online public discourse.

2 Objective

In light of the 2024 U.S. presidential election, this project aims to conduct a comprehensive sentiment analysis of tweets related to key political figures, specifically Donald Trump, Joe Biden, and Kamala Harris. Leveraging advanced Natural Language Processing (NLP) techniques and Large Language Models (LLMs), the project seeks to capture and analyze shifts in public sentiment toward these candidates over time, especially in response to significant political events.

The data collection process focuses on tweets from crucial swing states during the months leading up to the election, utilizing relevant hashtags, mentions, and regional filters to ensure the dataset is representative of key electoral regions. After gathering this data, it undergoes a detailed cleaning and preprocessing phase, transforming raw Twitter content into a structured format suitable for analysis.

Using sentiment classification models, the tweets are categorized as positive, negative, or neutral, providing insights into how public perception fluctuates. Additionally, the project aims to build predictive models that forecast changes in sentiment and potential political outcomes. By tracking sentiment trends and voter mood dynamics, the analysis provides a nuanced, data-driven perspective on voter behavior during this pivotal election period.

The project also presents an opportunity to apply modern data science techniques, enhancing skills in data collection, NLP, and predictive analytics in a real-world context.

3 Data

3.1 Data Ingestion

3.1.1 Data Source

For our project, we needed to collect a substantial number of tweets based on specific filters, such as location, date, user handle, and hashtags. While the Twitter API offers an official method for retrieving tweet data, we encountered limitations in terms of the available filtering options and access levels. The basic API did not provide the flexibility we required to apply all the necessary filters simultaneously. Moreover, the API's tiered access models posed a significant challenge. Given the volume of tweets we aimed to collect, none of the available tiers were suitable for our data collection needs, and the only tier that met our requirements came at a prohibitively high cost.

To overcome these limitations, we explored alternative approaches to gather the data. One potential solution was using Python-based Twitter scrapers. We tried various scrapers, each offering a different range of functionalities. However, every tool we tested had some limitations in terms of filter criteria, either lacking the ability to filter by location, date, or user handle, or being inefficient for the large-scale data collection we intended.

After careful consideration, we decided to use an online Twitter scraping tool provided by Apify. This tool allowed us to overcome the restrictions imposed by the traditional API and Python-based scrapers. The Apify scraper offered more advanced filtering options, enabling us to extract tweets based on the specific criteria we needed. Additionally, it supported the large-scale data collection necessary for our analysis, making it the most viable option for this project.

3.1.2 Data Collection

For our data collection, we leveraged a news article[1] from U.S. News to identify seven key swing states: Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin. These states were considered crucial in the context of the upcoming U.S. presidential election. The data collection spanned a six-month period, from March 15, 2024, to September 15, 2024, to capture sentiment and discussion leading up to the election.

We focused on collecting tweets that mentioned specific user handles related to the primary political figures in the election, including @JoeBiden, @KamalaHarris, and @realDonaldTrump. In addition to user handles, we gathered tweets containing popular election-related hashtags. These included hashtags explicitly supporting candidates, such as #DonaldTrump, #Trump2024, #JoeBiden, #TrumpForPresident, #KamalaHarris, and #Harris2024, as well as more neutral or broadly discussed hashtags like #USElection, #TrumpVsBiden, and #TrumpHarrisDebate.

To process the collected data, the results were provided in the form of multiple JSON files. For each combination of location, date, and user handle or hashtag, a separate JSON file was created, each containing multiple tweets. Every tweet in these files included over 100 fields with various

details, ranging from metadata to engagement metrics. However, not all of this information was necessary for our analysis.

For the purposes of this project, we focused on a subset of relevant columns. We combined all the JSON files into a single dataset, extracting only the fields essential for sentiment analysis and further exploration. These fields included:

1. **id** : A unique identifier assigned to each tweet.
2. **tweet_text**: The content of the tweet, representing the textual data posted by the user.
3. **created_at**: The exact timestamp when the tweet was created and posted on Twitter.
4. **retweet_count**: The total number of times the tweet has been retweeted by other users.
5. **reply_count** : The total number of replies that the tweet has received from other users.
6. **like_count** : The total number of likes (or "favorites") the tweet has garnered from users.
7. **view_count** : The total number of views the tweet has accumulated, indicating its reach.
8. **state** : The U.S. state associated with the tweet, reflecting the geographical location from which it was fetched.
9. **party** : The political party associated with the user handle or hashtag mentioned in the tweet (e.g., Democratic or Republican).
10. **handle_or_hash** : Specifies whether the tweet pertains to a user handle or a hashtag related to the analysis.

By streamlining the dataset to these essential columns, we ensured a more focused and efficient analysis while preserving the key information required for understanding voter sentiment and engagement across the selected swing states.

3.2 Data Exploration

We collected a total of 110,077 records from seven key swing states over a six-month period. Since Twitter stores user locations as approximate bounding boxes [2] rather than precise coordinates, there is a possibility that some tweets may originate from neighboring states, leading to a slight overlap in the geographical distribution of the data. This potential overlap is a known limitation of geolocation data in social media platforms.

3.2.1 Distribution of Tweets Over Swing States

The distribution of tweets over swing states can be seen in Fig. 1. Pennsylvania leads with the highest tweet count at 24,183, followed by Michigan with 19,689 tweets and Georgia with 18,553 tweets. Nevada and Wisconsin also have significant engagement, with 16,493 and 14,940 tweets, respectively. North Carolina and Arizona have comparatively lower tweet counts, at 9,668 and 6,551. The variation in tweet counts across these states highlights differing levels of social media engagement among the swing states.

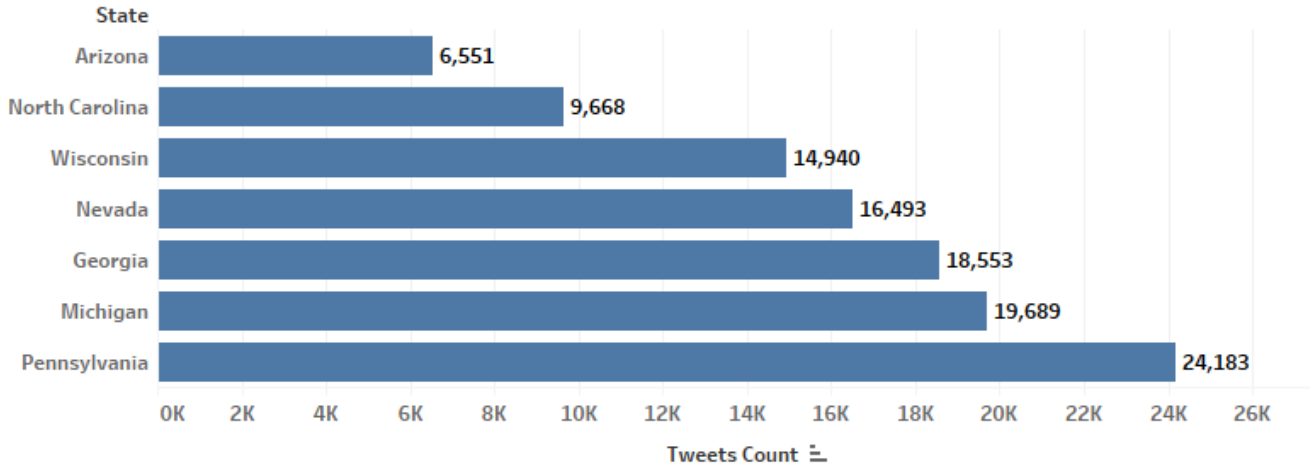

Figure 1: Distribution of Tweets Over Swing States, displaying the tweet counts for each state.

These differences may reflect varying levels of public interest or political activity across the states. States like Pennsylvania, Michigan, and Georgia generated considerably more Twitter activity, possibly due to higher population sizes, key political events, or active online communities, while Arizona and North Carolina saw lower engagement.

3.2.2 Distribution of Tweets Over Time

The graph, Fig. 2, displays the distribution of tweets over time, segmented by political affiliation: Democratic, Republican, and both. Throughout the observed period, tweet activity maintains a relatively steady flow, punctuated by significant surges during key political events. These surges illustrate the public's tendency to engage more actively on social media during critical moments. For instance, around July 11, a sharp increase in tweets can be seen following the Trump assassination attempt, demonstrating how events of great political consequence can trigger heightened discourse and engagement on platforms like Twitter. Such events not only generate immediate responses but also often spark extended conversations that influence online activity for days or even weeks.

In addition to this event, other significant political occurrences also triggered spikes in tweet activity. Key moments, such as presidential debates, Kamala Harris's entry into the presidential race, and the Trump-Musk interview, show noticeable increases in tweets shortly after they occurred. These patterns demonstrate how pivotal political events influence online discussions and engagement across party affiliations.

3.2.3 Distribution of Tweets by State and Party

The figure, Fig. 3, shows the distribution of tweet counts by political affiliation (Democratic, Republican, and Both) across key swing states. Across these states, Republican candidates, particularly Donald Trump, dominate the conversation, with the highest tweet counts for the Republican

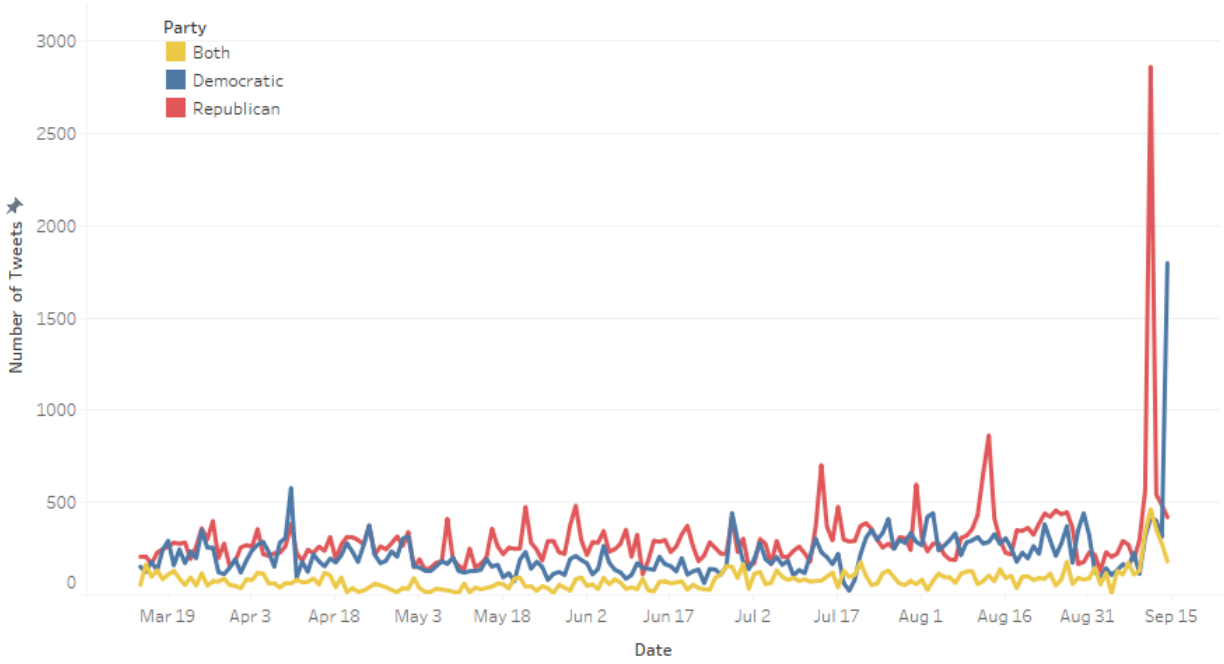

Figure 2: Distribution of Tweets Over Time, displaying the spikes in tweet activity around notable political events.

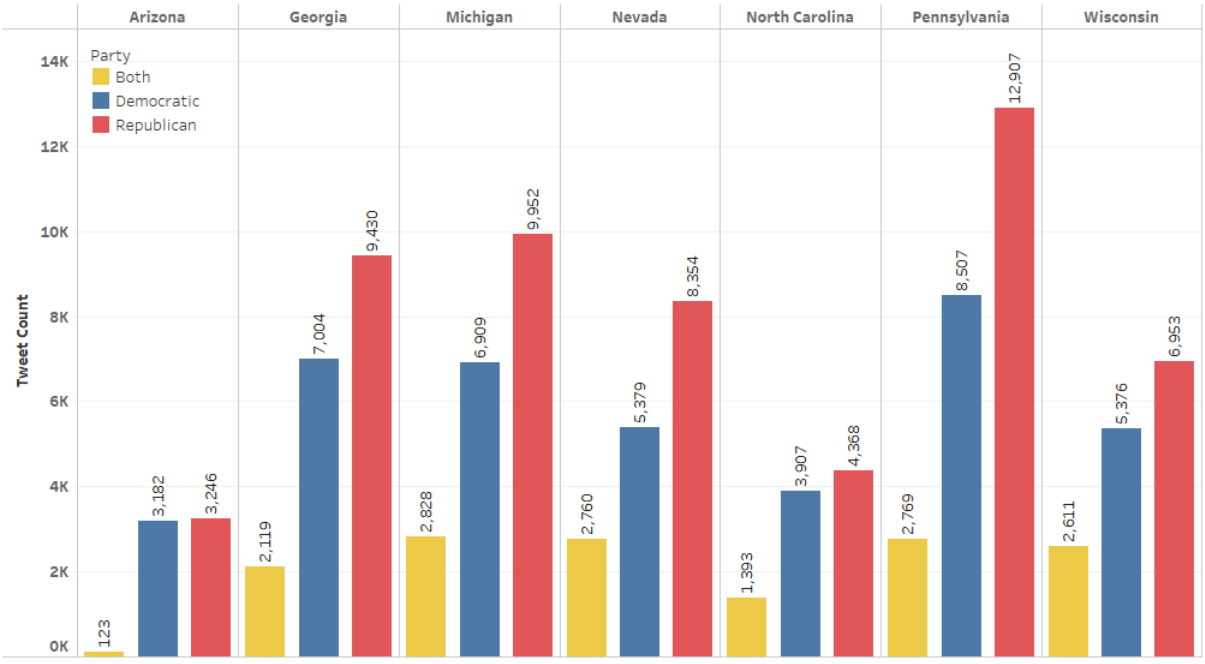

Figure 3: Distribution of Tweets by State and Party, showing how often each party is mentioned.

party in several regions.

Democratic engagement is also substantial, particularly in states like Pennsylvania and Georgia, though generally lower than Republican tweet counts across most regions. The Both category, representing tweets mentioning multiple candidates, has the lowest engagement overall.

While this figure reflects the volume of tweets, it is important to note that the number of tweets does not necessarily indicate positive sentiment. The tweets may include both supportive and critical discourse, as social media activity around political figures often represents a broad spectrum of opinions. Therefore, high tweet counts for a particular candidate or party may reflect a mix of both praise and criticism.

3.2.4 Engagement Metrics Over Time

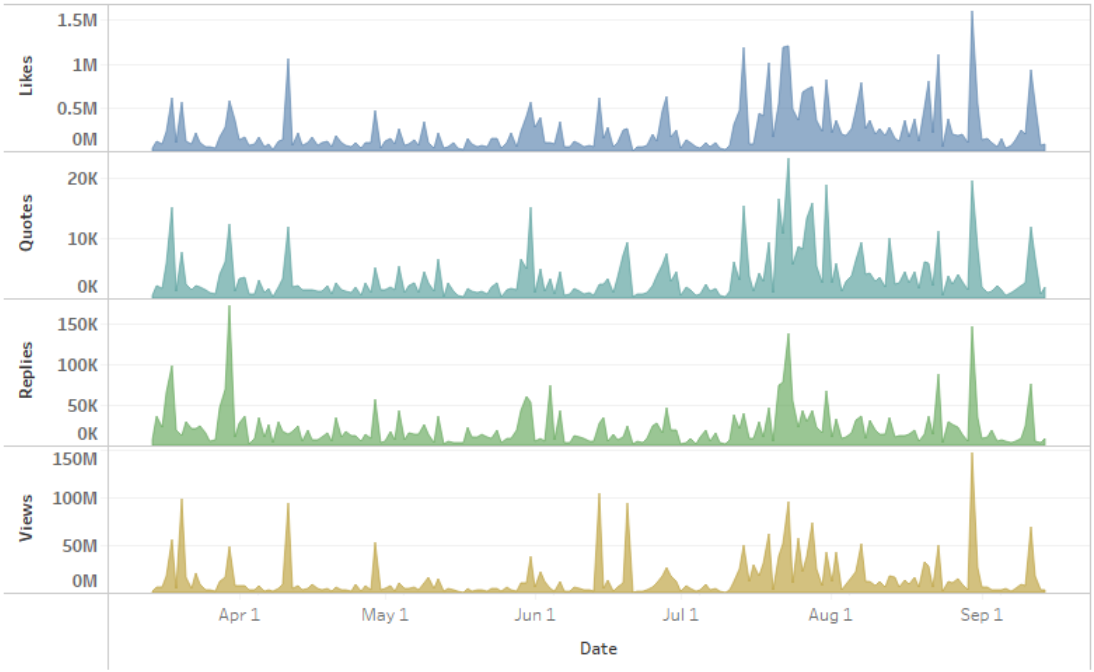

Figure 4: Distribution of Engagement Metrics over Time, showing how user interactions evolve through likes, quotes, replies, and views.

The figure, Fig.4, presents the evolving patterns of Twitter engagement over time, showcasing the dynamic interaction between users and content through various metrics. Engagement levels rise and fall, with certain moments marked by distinct peaks, reflecting periods of heightened public interest and interaction. These surges in activity often align with key political or social events, sparking increased visibility and user responses.

As engagement intensifies, users contribute through a variety of actions, including liking, sharing, and commenting on tweets. The data reveals moments when content resonates more broadly, drawing widespread attention and sparking conversations. During these peaks, users not only view

the content but are also compelled to share their own perspectives, adding to the discourse and amplifying the reach of specific tweets.

The recurring spikes in activity suggest that social media engagement is closely tied to external events, with moments of high interaction providing a glimpse into the public's collective reaction. These patterns demonstrate how Twitter serves as a platform for real-time conversations, where significant moments generate a ripple effect of engagement, driving likes, shares, and discussions across the platform.

3.2.5 Average Engagement Metrics by Party

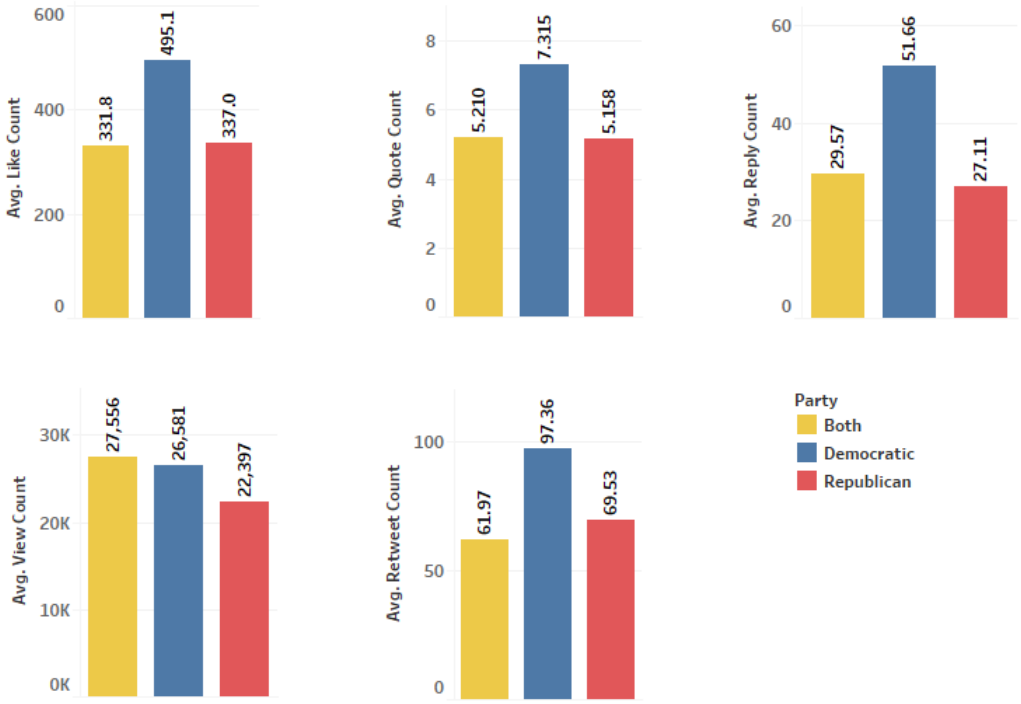

Figure 5: Average Engagement Metrics by Party, illustrating the variation in average engagement for each party.

Fig. 5 provides interesting observations about how user interactions vary between Democratic, Republican, and multi-party (Both) tweets. Notably, Democratic-affiliated tweets tend to see the highest engagement in multiple metrics, such as retweets, replies, and quotes. This indicates that Democratic content is not only widely viewed but also actively shared and discussed by users. The higher quote counts suggest that users often feel compelled to add their opinions or commentary when sharing Democratic-affiliated tweets, driving deeper conversations.

On the other hand, while Republican-affiliated tweets show slightly lower average engagement in several categories, they still maintain significant interaction levels, particularly in likes and views. This suggests that although users may engage less frequently through quotes and replies, they still express their opinions through simpler interactions, such as likes. This might reflect a different mode of engagement among the audience, where Republican content is consumed more passively, with fewer instances of users adding their commentary or entering into discussions.

Tweets mentioning both parties show a balanced level of engagement, with metrics like views and replies falling between those of Democratic and Republican tweets. This suggests that such tweets attract a broader audience, encouraging interaction from users with different affiliations. These tweets may spark cross-party discussions, leading to a more evenly distributed engagement pattern. Overall, the figure highlights how content and political alignment influence user interaction on social media.

3.2.6 Tweets Across User Handles or Hashtags

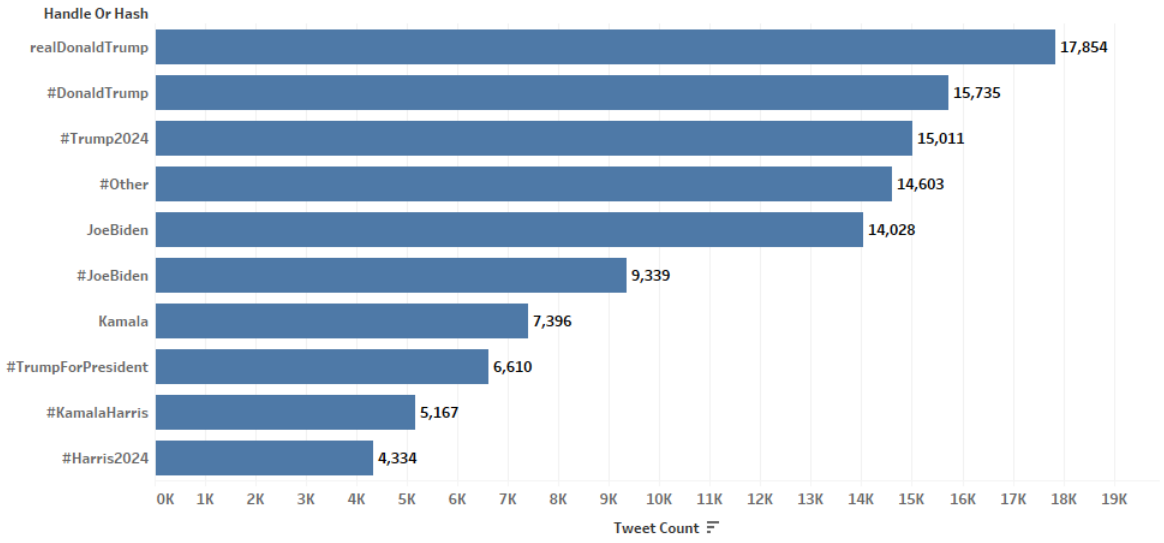

Figure 6: Breakdown of Tweets by User Handles and Hashtags, highlighting the most discussed figures and topics on Twitter.

The plot, Fig. 6, provides a breakdown of tweet counts by user handles and hashtags, revealing the most frequently used handles and hashtags in discussions. realDonaldTrump leads with the highest mentions at 17,854 tweets, followed by #DonaldTrump with 15,735 tweets. Other Trump-related hashtags, such as #Trump2024 and #TrumpForPresident, also rank prominently, reflecting strong engagement surrounding his political activities.

Mentions of Joe Biden and #JoeBiden are also significant, though they trail behind Trump-related mentions, with 9,339 and 7,396 tweets, respectively. Kamala Harris and associated hashtags, like #KamalaHarris and #Harris2024, show fewer mentions, with 5,167 and 4,334 tweets, indicating a

relatively lower level of Twitter engagement.

It is worth noting that the #Other category, which accounts for 14,603 tweets, includes important election-related hashtags such as #USElection, #TrumpVsBiden, and #TrumpHarrisDebate, capturing broader discussions around the 2024 election. This breakdown emphasizes that Trump-related content dominates the social media landscape, followed by Biden, with Harris receiving less attention during this period.

3.2.7 Word Cloud Analysis of Tweets Data

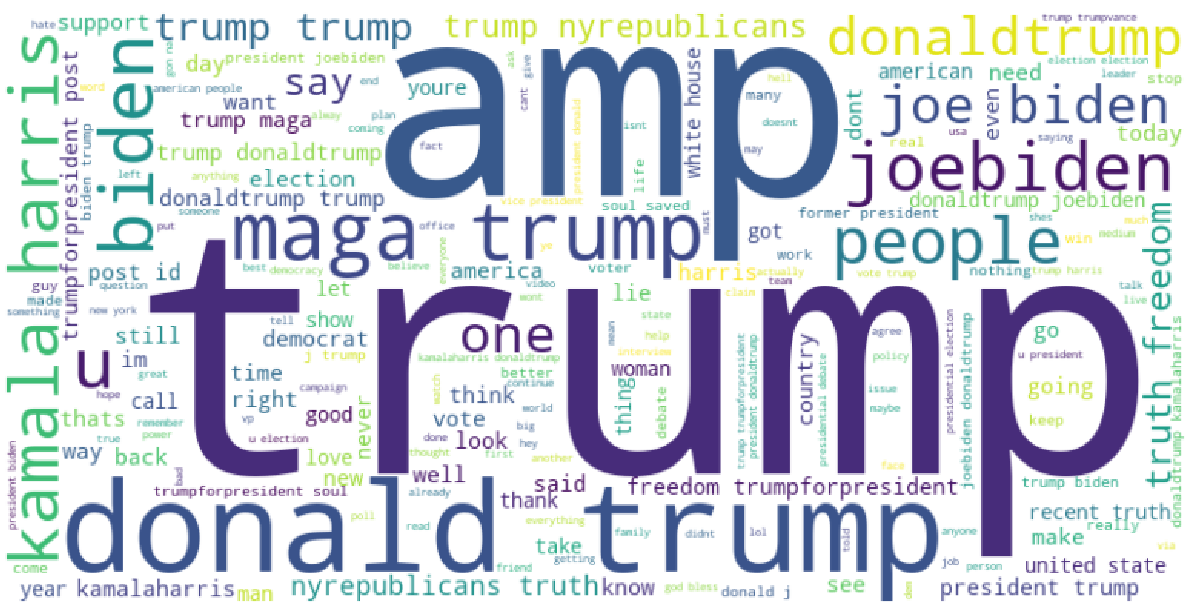

Figure 7: Word Cloud of Tweets Data, highlighting the most frequently used words in political discussions on Twitter

The word cloud visualizes the most commonly used words in tweets related to political discussions, with the size of each word representing its frequency of occurrence. "Trump" appears as the most prominent word, reflecting a strong focus on discussions surrounding Donald Trump. Other frequent terms include "amp", "maga", and "donaldtrump", which are associated with Trump's political campaign and his supporters.

In addition to Trump-related words, the names "Biden", "joebiden", and "Kamala Harris" are also featured, though at smaller sizes, indicating their presence in the conversation but with less frequency compared to Trump. Words like "freedom", "truth", and "people" suggest themes related to political values and ideologies being discussed in these tweets.

Overall, the word cloud highlights the dominant political figures and key terms in the social media discourse, showing that Trump-related topics continue to be a central focus, followed by discussions involving Joe Biden and Kamala Harris.

3.3 Data Preparation

The text data underwent comprehensive preprocessing to prepare it for sentiment analysis. Initially, all tweets were converted to lowercase to ensure uniformity in the dataset and prevent case sensitivity from affecting the analysis. URLs were systematically removed using regular expressions, as they often add noise and do not contribute meaningfully to sentiment detection. Similarly, user mentions (e.g., username) were stripped from the tweets to eliminate unnecessary variability, as these elements typically do not impact the sentiment of a tweet. Hashtags, which often carry significant meaning, were processed by removing the # symbol while retaining the accompanying text. This allowed terms like "#Election2024" to remain useful for sentiment evaluation without the clutter of the hashtag symbol.

Further refining the text, non-alphabetic characters, such as punctuation and numbers, were removed to maintain a focus on alphabetic content that holds semantic value. Once the text was cleaned, tokenization was applied, breaking the text into individual words or tokens. This enabled more granular processing, allowing the next steps—stopword removal and lemmatization—to be applied more effectively. Stopwords, such as common English terms like "and," "the," and "is," were filtered out using NLTK's predefined stopword list. These words contribute little to sentiment analysis and were excluded to reduce noise in the data.

Lemmatization, an important step in the process, was applied to the remaining tokens. By converting words to their root forms (e.g., "running" to "run"), lemmatization ensured that different variations of the same word were standardized, which is crucial for improving the generalizability of the sentiment model. This process helped streamline the dataset by collapsing multiple forms of a word into a single, consistent representation, enhancing the efficiency and accuracy of the subsequent analysis.

Finally, the cleaned and lemmatized tokens were rejoined into cohesive strings of text. At this stage, the data was well-structured, having passed through multiple layers of cleaning and standardization.

4 Methodology

For this project, we aim to predict which candidate has the highest chance of winning the upcoming presidential elections based on sentiment analysis of Twitter data. We use Large Language Models (LLMs) and VADER, a sentiment analysis tool, to classify the sentiment in tweets, providing insight into public opinion toward various candidates. Since our data is unlabelled, we rely on LLMs and fine-tuning techniques like zero-shot and Few-shot learning to enhance the model's performance [3].

4.1 Techniques

4.1.1 Transformer-based Sentiment Analysis and Large Language Models (LLMs)

For this project, we use transformer-based models from the Hugging Face library and state-of-the-art Large Language Models (LLMs) such as GPT-4, DistilBERT, RoBERTa and BERTweet to

perform sentiment analysis on Twitter data related to the upcoming presidential election. These models, pre-trained on large text corpora, are fine-tuned to classify the sentiment of tweets as positive, negative, or neutral, helping us understand public opinion on various political candidates[4].

Our analysis starts by preprocessing the tweet text, including handling unique Twitter elements like usernames and links. We tokenize the text using a pre-trained tokenizer suited for the transformer models. This tokenized text is then fed into the model to generate sentiment scores. To convert the model's output into a more interpretable form, we apply the softmax function, which provides probability distributions over sentiment labels. This ensures that each tweet is categorized based on the model's understanding of the text's sentiment.

4.1.2 BERTweet

BERTweet is fine-tuned specifically for Twitter data and excels at handling the informal and abbreviated language typical of tweets. This makes it indispensable for our sentiment classification task, as it accurately captures public sentiment expressed in short-form texts, including the slang and abbreviations commonly used in political discussions on Twitter [5].

4.1.3 RoBERTa (Robustly Optimized BERT Approach)

RoBERTa is a highly effective model for sentiment classification tasks due to its robust pre-training on large datasets. It captures sentiment nuances by leveraging context from both directions in a sentence. For our political data analysis, RoBERTa helps break down tweets into positive, negative, or neutral sentiment, giving us clear insights into public opinion about candidates and policies[6].

4.1.4 DistilBERT

DistilBERT is a smaller and faster version of BERT, optimized for efficiency without sacrificing much performance. Given the large volume of Twitter data, DistilBERT allows us to handle sentiment classification in a computationally efficient manner. It is particularly useful in processing the sentiment of a large number of tweets in real time while maintaining accuracy.

4.1.5 GPT (Generative Pre-trained Transformer)

While GPT is primarily designed for text generation, it can be adapted for sentiment classification through fine-tuning or by using zero-shot learning approaches. In this project, we utilize GPT to classify sentiment by framing the task as a question-and-answer problem, such as "Is this text positive, neutral, or negative?" This allows GPT to infer sentiment without requiring task-specific training, making it a versatile model for political sentiment analysis. By employing Prompt Engineering Fine Tuning (PEFT), we can leverage GPT's ability to perform sentiment analysis based on the context of the political discourse on Twitter. This model is particularly useful for analyzing broader discourse trends and interpreting more complex or subtle sentiments that arise in political discussions.

4.1.6 Fine-tuning Techniques

Given the lack of labeled data in this project, we explore zero-shot learning to perform sentiment analysis without task-specific training. This allows the models to generalize their pre-trained knowledge to new, unseen data. In addition, we use one-shot and few-shot learning, manually labeling a small portion of the tweets to fine-tune the models further [7]. These fine-tuning techniques ensure that the models are well-adapted to the specific context of political sentiment analysis.

4.2 Process Validation

Since our dataset lacks labeled data, traditional evaluation metrics such as accuracy, precision, and recall cannot be directly applied. Therefore, we adopt alternative evaluation approaches to assess the performance of our sentiment analysis.

4.2.1 Sentiment Trend Analysis

We track the sentiment trends across time or specific political events, such as candidate debates or public announcements. By analyzing these trends, we can observe whether the predicted sentiments align with well-known political developments or shifts in public opinion. A close alignment between model predictions and real-world events can serve as a qualitative validation of the model's effectiveness in capturing the general sentiment of the electorate.

4.2.2 Manual Annotation for a Subset

To gain more direct insight into the model's performance, we manually label a random subset of the tweets. The model's predictions are then compared to these human-labeled sentiments. This allows us to compute metrics like precision, recall, and F1-score for the labeled subset, providing a more concrete evaluation of the model's accuracy on a small sample of the data.

4.2.3 Election Results as a Validation Metric

Ultimately, the results of the upcoming presidential election will serve as an additional validation metric. The predicted sentiment trends for each candidate will be compared to the actual election outcome. While this is a retrospective validation, if the model's sentiment predictions correspond to the election results, it will further indicate the model's ability to capture public opinion accurately.

By combining these methods, we ensure a well-rounded evaluation of our sentiment analysis, even in the absence of a fully labeled dataset.

References

[1] E. D. Jr., "7 states that could sway the 2024 presidential election," *U.S. News & World Report*, Oct. 2, 2024. [Online]. Available: https://www.usnews.com/news/elections/articles/7-swing-states-that-could-decide-the-2024-presidential-election.

[2] *Place object model*, Accessed: 2024-10-08. [Online]. Available: https://developer.x.com/en/docs/x-api/data-dictionary/object-model/place.

[3] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," in *Findings of the Association for Computational Linguistics: NAACL 2024*, Association for Computational Linguistics, June 16-21, 2024, DAMO Academy, Singapore and Alibaba Group, Hangzhou, China, 2024, pp. 3881–3906.

[4] A. Vaswani *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[5] *Twitter roberta base for sentiment analysis*, https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment, Accessed: 2024-10-09.

[6] *Twitter roberta base for sentiment analysis (latest version)*, https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest, Accessed: 2024-10-09.

[7] S. Wu *et al.*, "Bloomberggpt: A large language model for finance," in *Proceedings of the Bloomberg Research Conference*, Bloomberg, New York, NY, USA, 2024.