Twitter Sentiment Analysis for Presidential Election 2024

Team Members:
Srushti Shinde (ss17454)
Huilin Zhang (hz3455)

1. Introduction

a. In the lead-up to the 2024 U.S. Presidential Election, social media platforms like Twitter play a significant role in shaping public opinion and political discourse.
b. This project focuses on Twitter sentiment analysis to evaluate the public's views on key candidates, specifically Donald Trump and Kamala Harris.
c. By leveraging machine learning models and sentiment analysis tools, the project classifies tweets as positive, negative, or neutral, and further identifies whether a tweet is biased towards Trump, Harris, or neutral.
d. Additionally, the project explores how these sentiments correlate with financial markets and prediction platforms, such as Polymarket.

2. Data Collection

Twitter Data:

a. ***dataset1_uncleaned_tweets_data_with_dates.csv***: Contains raw, unprocessed tweets with associated dates. This dataset includes the uncleaned text of the tweets and is used as the basis for sentiment analysis.
b. ***dataset2_uncleaned_tweets_data_with_dates.csv***: Similar to the first dataset, this file contains uncleaned tweet text but for a different set of data.
c. ***dataset1_preprocessed_tweets_with_dates.csv***: The preprocessed version of the first dataset, where emojis, stopwords, and unnecessary characters have been removed, ready for sentiment analysis.
d. ***dataset2_preprocessed_tweets_with_dates.csv***: The cleaned version of the second unprocessed tweet dataset, prepared for further analysis.
e. ***combined_tweets_with_sentiment.csv***: This dataset combines the preprocessed tweets with their sentiment scores (positive, negative, or neutral) and identifies if the tweet is biased towards Trump, Harris, or neutral.

f. ***tweets_with_contextual_sentiment.csv***: This dataset further analyzes each tweet by including the contextual sentiment score for Trump and Harris individually, allowing for a deeper understanding of bias in each tweet.

Financial Market Data:

a. ***gold_2024_adj_close.csv***: Contains the daily adjusted close prices for gold from January 2024 to October 2024, used to analyze gold's performance in relation to the election sentiment.
b. ***sp500_2024_adj_close.csv***: This file tracks the adjusted close prices of the S&P 500 index over the same period, providing insights into how the stock market reacts to public sentiment regarding the election.
c. ***russell_2000_adj_close_2024.csv***: Records the adjusted close prices of the Russell 2000 index, which focuses on smaller U.S. companies, offering a different perspective on the election's market impact.
d. ***treasury_yield_2024_adj_close.csv***: Contains data for the U.S. 10-year Treasury Yield, often used as a benchmark for investor sentiment towards government debt and long-term economic outlook.
e. ***polymarket_daily_election.csv***: Tracks the prediction prices for the 2024 U.S. Presidential Election on Polymarket, a decentralized prediction platform. This dataset is used to correlate market prediction sentiment with public opinion on social media.

3. Data Processing

a. ***combine_tweet_data.py***: Merges multiple tweet datasets into a single file for further sentiment analysis.
b. ***data_preprocessing.py***: Cleans the tweet data by removing emojis, stopwords, and unnecessary characters, preparing it for sentiment analysis.
c. ***dataset_distribution.py***: Analyzes the distribution of sentiment classification categories (Pro-Trump, Pro-Harris, Neutral) in the dataset.
d. ***overall_sentiment_analysis_using_vader.py***: Performs overall sentiment analysis of tweets using the VADER sentiment analysis tool, assigning sentiment scores.
e. ***trump_harris_sentiment_analysis_using_vader.py***: Conducts specific sentiment analysis of tweets related to **Trump** and **Harris**, identifying bias in favor of each candidate.

4. Financial Indicators Overview

a. **S&P 500**: Tracks the performance of 500 large U.S. companies, used to observe how stock market trends correlate with election-related sentiment.
b. **Gold**: Represents the adjusted close price of gold, acting as a safe-haven asset and reflecting market sentiment during periods of political uncertainty.
c. **10-Year Treasury Yield**: Measures the yield on U.S. government bonds, reflecting investor sentiment about long-term economic stability in relation to the election.
d. **Russell 2000**: Monitors the performance of 2,000 smaller U.S. companies, providing insight into how sentiment around the election impacts smaller businesses.
e. **Polymarket Prediction Prices**: Tracks prediction market data, showing how public opinion on platforms like Polymarket aligns with sentiment analysis from social media regarding election outcomes.

5. Chi square Test

a. The **Chi-Square test** is used in this project to identify the most significant features (words or terms) that have a strong association with the target variables (sentiment classification).
b. By selecting the best features, the Chi-Square test helps to reduce the dimensionality of the data and remove irrelevant or redundant information.
c. Before applying the **Naive Bayes classifier**, using the Chi-Square test ensures that the model focuses on the most meaningful features, improving its accuracy and efficiency by reducing noise in the dataset.
d. This step is crucial for enhancing the performance of Naive Bayes when classifying tweets as pro-Trump, pro-Harris, or neutral.

6. Naive Bayes Classification

a. The Naive Bayes algorithm in this code is implemented to classify tweets into categories such as pro-Trump, pro-Harris, or neutral, based on their features (words).
b. It uses a probabilistic approach, where each word contributes to the likelihood of a tweet belonging to a specific sentiment class, assuming that all features are independent of each other.

c. We obtain a 60% accuracy on the dataset which we created using the tweets obtained from twitter related to the elections.

7. Analysis & Findings

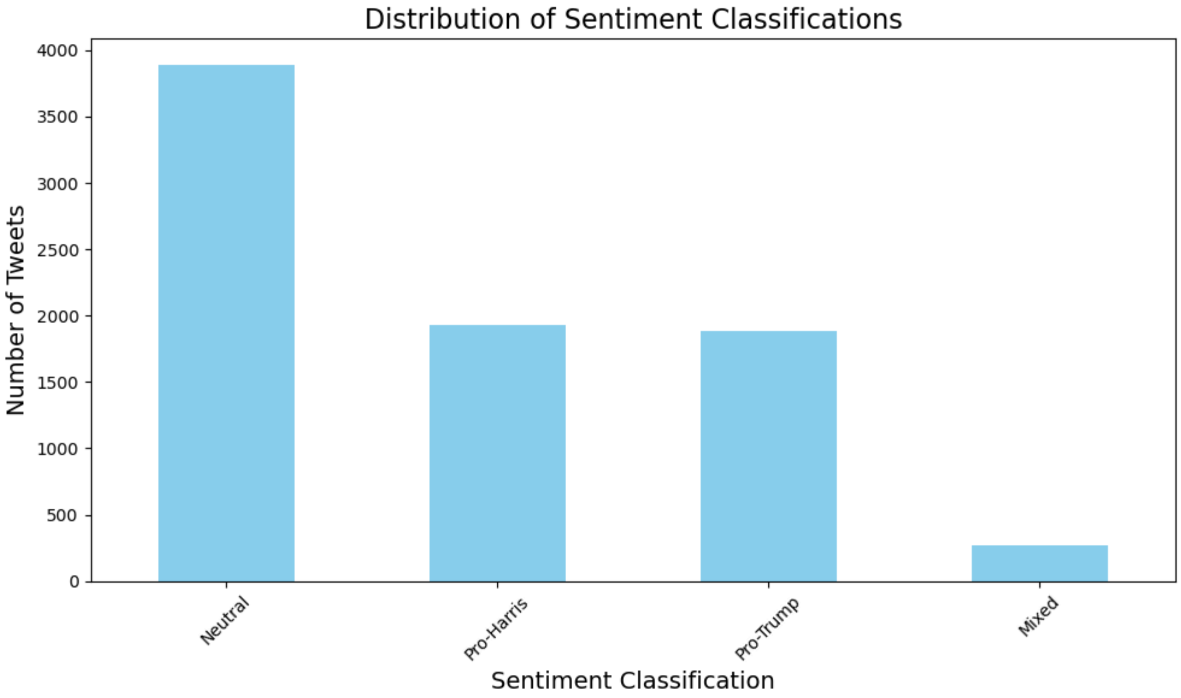

Figure 1: **Distribution of Sentiment Classification**
This chart presents the overall distribution of sentiment classifications (positive, negative, neutral) for tweets related to both Trump and Harris. It gives a snapshot of how public opinion is divided between the two candidates over the analyzed period.

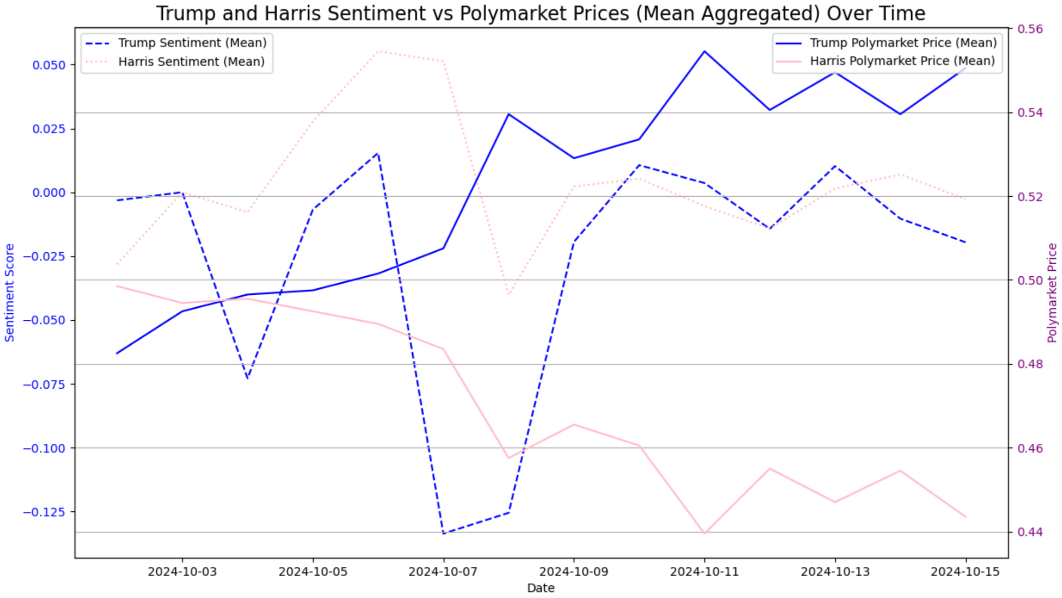

Figure 2: **Sentiment vs. Polymarket Chart**
Shows the relationship between the public sentiment (positive, negative, neutral) toward Trump and Harris, plotted over time, against Polymarket's probability predictions for each candidate. Highlights how shifts in sentiment correlate with the perceived chances of election success. Positive sentiment correlates with an increase in Trump's election probability on Polymarket. Conversely, when negative sentiment rises, his chances tend to drop slightly.

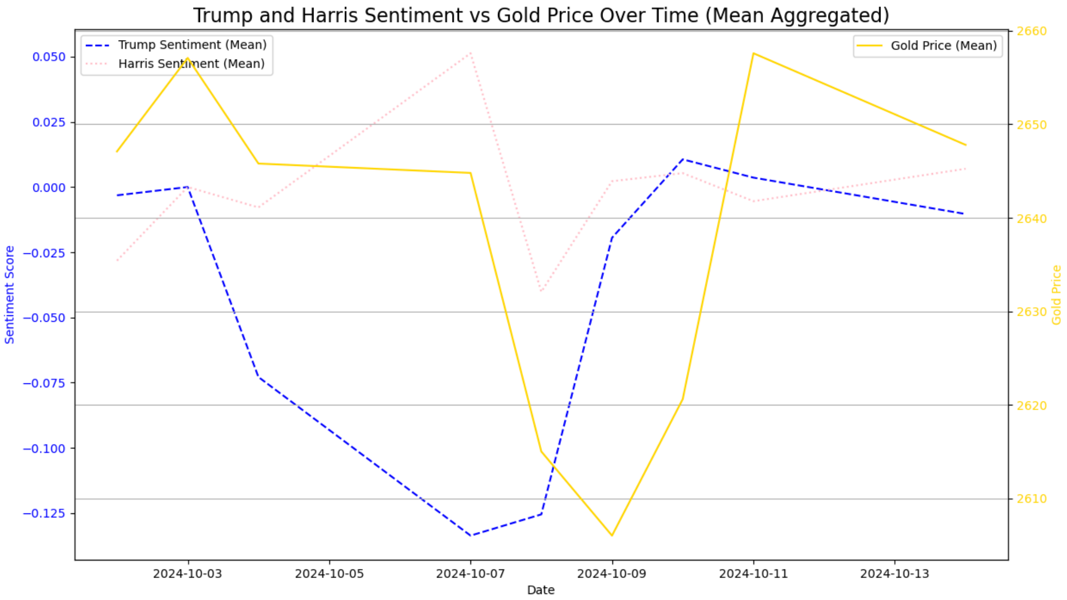

Figure 3: **Sentiment vs. Gold Prices**
This plot visualizes how the sentiment surrounding each candidate correlates with fluctuations in gold prices. Negative sentiment about Trump correlates with a rise in gold prices, suggesting that investors seek safety in gold during uncertain times for Trump.

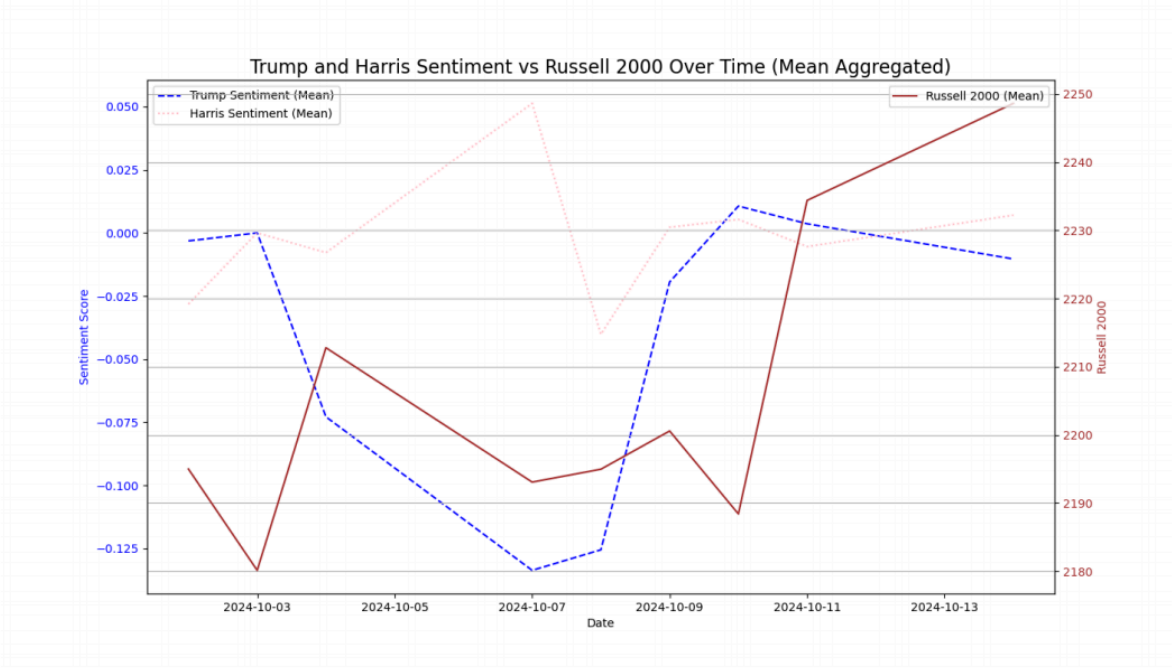

Figure 4: **Sentiment vs. Russell 2000**
The sentiment scores for both candidates are plotted against the Russell 2000 index, which focuses on smaller U.S. companies. The aim is to assess whether sentiment impacts this segment of the market, often more sensitive to political and economic shifts. It shows a modest upward trend when sentiment about Trump is positive, indicating small-cap stocks react favorably to positive news about Trump's policies. While less sensitive to Harris's sentiment.

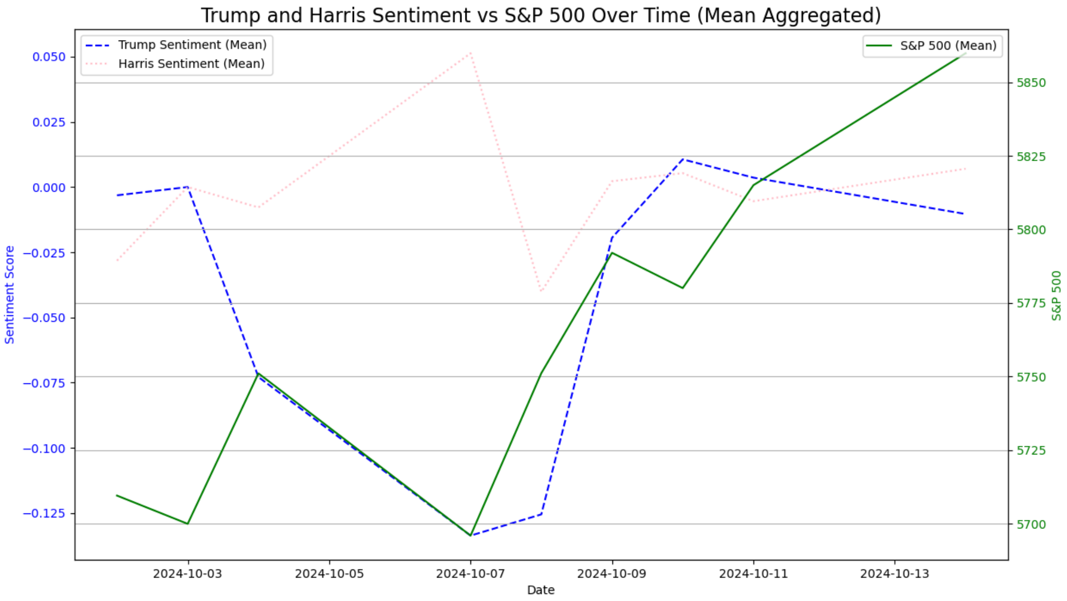

Figure 5: **Sentiment vs. S&P 500**
The sentiment scores for Trump and Harris are compared to the performance of the S&P 500 index. The chart aims to reveal whether changes in political sentiment have any noticeable impact on the broader stock market. There is a slight positive correlation between Trump's sentiment and the S&P 500. Positive sentiment tends to boost market confidence, pushing the index slightly upward, but little reaction to Harris's sentiment, reflecting a weaker market response to shifts in her perceived favorability.

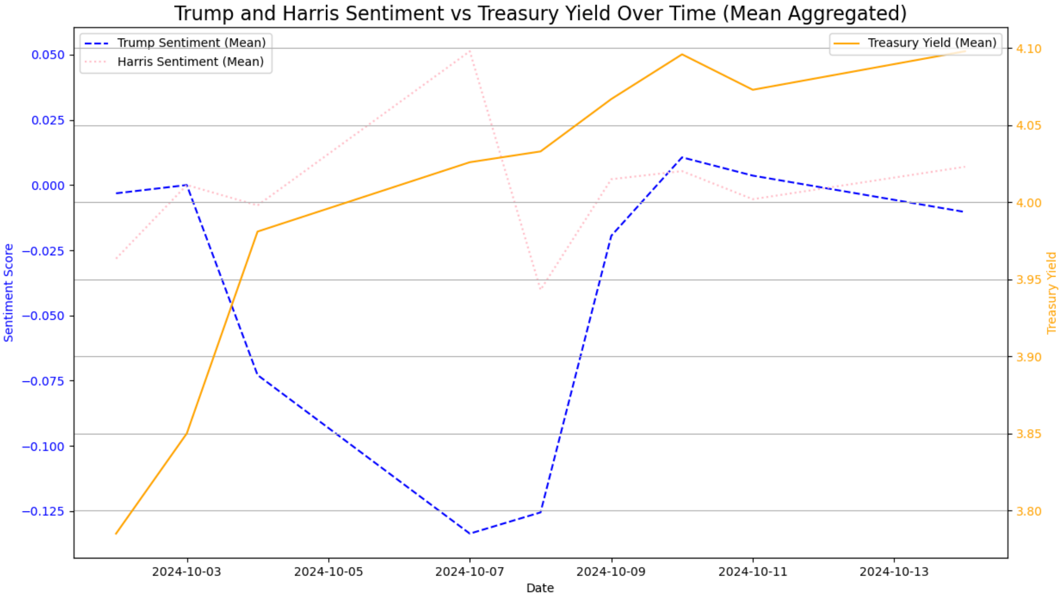

Figure 6: **Sentiment vs. 10-Year Treasury Yield Chart**:
This graph compares the sentiment trends for Trump and Harris with the 10-year Treasury yield, a key indicator of investor confidence. It attempts to show how election-related sentiment might influence long-term economic outlooks. Harris's sentiment doesn't appear to strongly impact the 10-year Treasury yield, showing only minor fluctuations that are less consistent. And Trump is associated with a drop in the 10-year Treasury yield, reflecting investor caution and lower confidence in long-term economic growth.

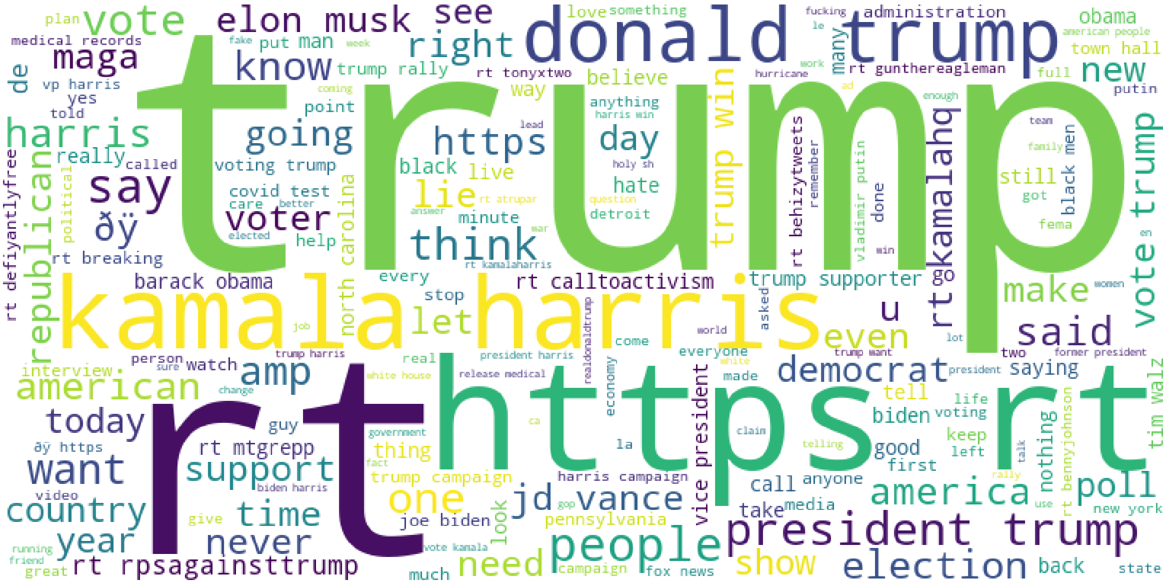

Figure 7: **Word Cloud**
The size of each word represents its frequency in the dataset: the larger the word, the more frequently it appears in the tweets. Prominent words like "Trump," "Harris,"

"RT" (retweet), "vote," and "https" indicate common topics of discussion, showing the central themes of the conversation surrounding the candidates.

8. Challenges & Limitations

a. **Ensuring the quality of the tweet data**. While the preprocessing steps helped clean the datasets, certain nuances in language, such as sarcasm or slang, may have affected the accuracy of sentiment classification. This was especially evident in tweets where sentiment was ambiguous or mixed, which made it difficult for both the Naive Bayes classifier and VADER to correctly assign polarity scores.

b. The **Naive Bayes algorithm** showed limitations, with an accuracy rate of 60%, which, while reasonable, leaves room for improvement. The algorithm's assumption that all features are independent may not always hold true in real-world data, particularly in complex social and political discussions.

c. While we explored correlations between tweet sentiment and financial market indicators, establishing strong, consistent relationships proved challenging due to the **short time span of the data** and the many external factors that influence markets, especially during election periods.

9. Conclusion

a. In this project, we implemented sentiment analysis to **track public opinion on Donald Trump and Kamala Harris** during the 2024 U.S. Presidential Election. By leveraging machine learning models like Naive Bayes and sentiment analysis tools like VADER, we were able to categorize tweets and evaluate how these sentiments aligned with financial market trends and prediction platforms like Polymarket.

b. While we identified some patterns between sentiment and financial variables such as the Russell 2000 and S&P 500, the correlations were not conclusive, likely due to the complexity of both market movements and the unavailability of proper dataset.

c. The project highlights the potential of sentiment analysis for understanding political sentiment but also underscores the need for more **sophisticated models** and **larger datasets** to draw more definitive conclusions.