#### 

#### **ITCS 6150 - Intelligent Systems**

#### **Final Project Report :Group 17**

#### **Twitter Sentiment Analysis and Presidential Candidate Prediction**

#### **Group Members:**

Kevin Avila

Micheal Rawlings

Namasani Deepak

Ruchitha Kudumula

Saniddhya Dubey

### 

### 

### Abstract:

Any country's leadership and policies are shaped by its elections, which are a pivotal event. Our project's goal is to investigate if Twitter sentiment may reveal information about the results of US presidential elections. We want to identify trends and connections between online conversation and election outcomes by examining tweets, including those from international users.

Cleaning the dataset to eliminate missing values and lessen bias by decreasing overrepresentation from regular users will be the first step in achieving this. After everything is ready, we will use Natural Language Processing (NLP) techniques such as Word Embeddings, TF-IDF, and Bag of Words (BoW) to extract significant features for sentiment analysis. We will use time-series analysis to track changes in public opinion and candidate popularity over time in order to look for relationships with election.Visualizing the information will be an important part of our project since this will be able to give us multiple insights when comparing features against each other. For example, we will be able to visualize just how much of an impact a certain user has if they receive more interactions than another user. Our findings aim to shed light on how digital platforms influence political narratives and outcomes, offering a fresh perspective on the impact of online conversations in shaping election dynamics.

### **Introduction**

Election years are pivotal times in a nation's political landscape, shaping leadership and determining the policies that guide governance. In recent years, social media–particularly Twitter–has become a central platform for political discussions, providing real-time insights into public sentiment and opinion. Our project seeks to explore the relationship and correlation between the sentiment of tweets and the outcomes of presidential elections. Using a dataset of nearly 1 million tweets from Kaggle, we aim to identify correlations between tweet sentiment and predicting election results. Through conducting sentiment analysis and tracking shifts in public opinion, this project aims to offer valuable insights into the role social media plays in shaping political narratives and influencing election outcomes.

### Problem Statement

This project aims to predict the election outcome by leveraging tweet sentiments to forecast the winner of the 2020 U.S. presidential election. To achieve this we first will create a clean, usable dataset through various preprocessing steps. Then we will conduct sentiment analysis to extract the sentiment of each tweet. Finally we will use multiple robust models to accurately predict election results based on the sentiment analysis. Additionally in this study we plan to also track sentiment trends towards each candidate. Through combining sentiment analysis with prediction models, we seek to show the correlation between twitter sentiment and election results.

**Motivation**

Elections are at the heart of any democracy, and in today’s digital world, platforms like Twitter have become essential spaces for people to express their thoughts and engage in political conversations. This project was inspired by the need to better understand how public sentiment on social media reflects or even influences political outcomes, focusing specifically on tweets about Joe Biden and Donald Trump.

By analyzing trends in sentiment, we aim to uncover patterns that reveal voter preferences and highlight the role of social media in shaping public opinion. This research helps connect digital behavior to real-world electoral processes, shedding light on how platforms like Twitter can amplify voices, create narratives, or even distort perceptions.

Ultimately, we hope our findings will add to the growing conversation about the impact of technology on modern democracy, offering insights that could pave the way for future studies at the crossroads of sentiment analysis, social media, and election forecasting.

### Datasets

### [US Election 2020 Tweets Dataset](https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets)

### [Twitter Tweets Sentiment Dataset](https://www.kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset)

### <https://www.kaggle.com/code/mishki/twitter-sentiment-analysis-using-nlp-techniques/input>

**Challenges**

As this project progressed, we ran upon a number of obstacles that needed thorough problem-solving. The vast volume of noisy data, which included duplicates and irrelevant tweets, was one significant obstacle that made analysis challenging. We used strict cleaning and filtering techniques to deal with this and concentrate on relevant material. The dataset's linguistic diversity presented another difficulty because tweets were sent by a worldwide audience. By using pre-trained multilingual models that can effectively analyze text in many languages, we were able to overcome this issue. Furthermore, the study might be skewed by the existence of bots and phony accounts. We reduced their influence by spotting automated behavior patterns. Last but not least, the inherent biases in social media data created a problem since they can overrepresent particular groups or viewpoints.

**Background/Related Work**

In recent years, social media platforms like Twitter have become vital tools for analyzing public opinion, especially during critical events like elections. Many studies have explored the connection between social media sentiment and electoral outcomes. For instance, research by Tumasjan et al. (2010) highlighted how Twitter activity and sentiment could reflect public opinion, offering insights into how political candidates are perceived during campaigns.

Past research has also identified significant challenges in working with social media data, such as dealing with noisy datasets, navigating language diversity, and accounting for the impact of bots and fake accounts. Techniques from Natural Language Processing (NLP), including methods like Bag of Words (BoW), TF-IDF, and Word Embeddings, have been widely applied to extract and analyze textual data. Recent advancements in pre-trained models like BERT and multilingual transformers have further improved the ability to process diverse and complex datasets accurately.

Another crucial area of related work focuses on the role of bots and foreign influence in shaping online discussions during elections. Studies like Ferrara et al. (2016) examined how automated accounts spread misinformation and affect public perception, highlighting the importance of filtering such activity to maintain reliable insights.

Building on these foundations, our project dives deeper into the analysis of sentiment in election-related tweets. We also explore whether patterns in online discussions align with election outcomes, contributing to a broader understanding of how digital platforms influence voter behavior and political processes. This work aims to provide fresh perspectives on the evolving role of social media in modern democracy.

**Existing Systems**

Several systems have been developed to analyze sentiment on social media, particularly during elections, by utilizing Natural Language Processing (NLP) techniques. Popular tools like VADER and TextBlob are commonly used to classify the sentiment of tweets as positive, negative, or neutral. These tools are effective for real-time analysis of public opinion during political events, but they can sometimes struggle with complex sentiments or handling content in multiple languages.

To tackle the challenge of language diversity, more advanced systems rely on pre-trained multilingual models, such as BERT and Hugging Face's transformers. These models allow for sentiment analysis across a wide range of languages, which is essential given Twitter’s global audience. Additionally, systems like Botometer are used to detect and filter out automated accounts, ensuring that the analysis reflects authentic user sentiment rather than being skewed by bots or fake activity.

Visualization tools, such as Tableau and D3.js, are often incorporated into these systems to present the analysis in a clear and engaging way. These platforms allow for the tracking of trends in candidate popularity and sentiment over time, offering valuable insights into how political discussions evolve. While these existing systems are useful, they still face challenges like dealing with biased data and the constant evolution of misinformation tactics. Our project aims to build on these tools by improving data cleaning, multilingual analysis, and the use of time-series visualizations to provide more accurate and comprehensive insights into election-related social media sentiment.

**Drawbacks**

Existing systems face challenges such as difficulty in detecting complex sentiments like sarcasm or mixed emotions, and issues with handling informal language or slang. Bot detection is also problematic, as bots can mimic human behavior. Social media data biases can distort sentiment analysis, and systems struggle to adapt to evolving misinformation. Additionally, many systems lack the ability to effectively track how sentiment changes over time in response to events.

### 

### 

### 

### 

### 

### 

### Our Approach

#### **Data Preparation**

### To prepare the data we began with cleaning the dataset by removing rows with missing or null values and dropping columns that didn’t make sense for our study. Next, we began cleaning the tweets themselves by removing URLs, mentions, extra spaces and excessive punctuation. Then, we converted all the tweets to lowercase. Lastly, we broke our tweets down into tokens, used lemmatization to reduce the words down to their base form, and removed all stop words.

#### 

#### **Sentiment Analysis**

### For our Sentiment Analysis we used Vader Sentiment Analysis through the NLTK library. Vader was our choice because it is a sentiment analysis model built to handle social media datasets. Through sentiment analysis we were able to get a sentiment score for all tweets and begin candidate prediction.

#### 

#### **Time-Series Analysis**

### After running sentiment analysis on our dataset we could begin to leverage the sentiment scores produced to track the sentiment over time. Using pandas and matplotlib, we accurately tracked the average sentiment everyday for each candidate.

#### 

#### **Prediction Models**

### The first model we used was Logistic Regression. This was a simple choice since we can break the problem into a binary classification. The second model we chose was RandomForest because it is a powerful prediction model. Lastly, we used XGBoost since it is a state of the art model.

#### 

#### **Visualization**

### We used many visualization to see how public opinion was looking for both candidates. We visualized the sentiment scores as a whole for both candidates, the overall sentiment day to day for both candidates and the results to all algorithms.

### Data Preprocessing

1. **Case Normalization**: Convert all tweets to lowercase for consistency.
2. **Initial Exploration**: Check data shapes and column names. Display first and last five rows for preview. Analyze column data types.
3. **Tweet Cleaning**: Remove URLs, mentions, extra spaces, and normalize excessive punctuation using the clean\_tweet function.
4. **Tokenization and Lemmatization**: Tokenize words in tweets using word\_tokenize. Remove stopwords and punctuation while lemmatizing tokens for meaningful analysis.

**Sentiment Scoring**

We used the VADER SentimentIntensityAnalyzer to compute sentiment scores for each tweet. Store compound sentiment scores in a new column. Visualization: Plot histograms to compare sentiment distributions between Biden and Trump tweets. Highlight neutrality with a vertical line. Sentiment Trends: Convert created\_at to datetime format and calculate daily average sentiment for both candidates. Plot time series to visualize sentiment trends over time. Average Sentiment: Calculate and print average sentiment for Biden and Trump.

### Data Integration for Modeling

**T**arget Labeling: Assign 1 to Biden tweets and 0 to Trump tweets. Merge both datasets into a single DataFrame. Feature-Target Split:Separate features (sentiment) and target (target). Split data into training and testing sets.

### Machine Learning Models

Logistic Regression: Train the model and evaluate using accuracy and classification reports. Random Forest Classifier: Train and evaluate with the same metrics. XGBoost Classifier: Train using a DMatrix format and evaluate with classification metrics.

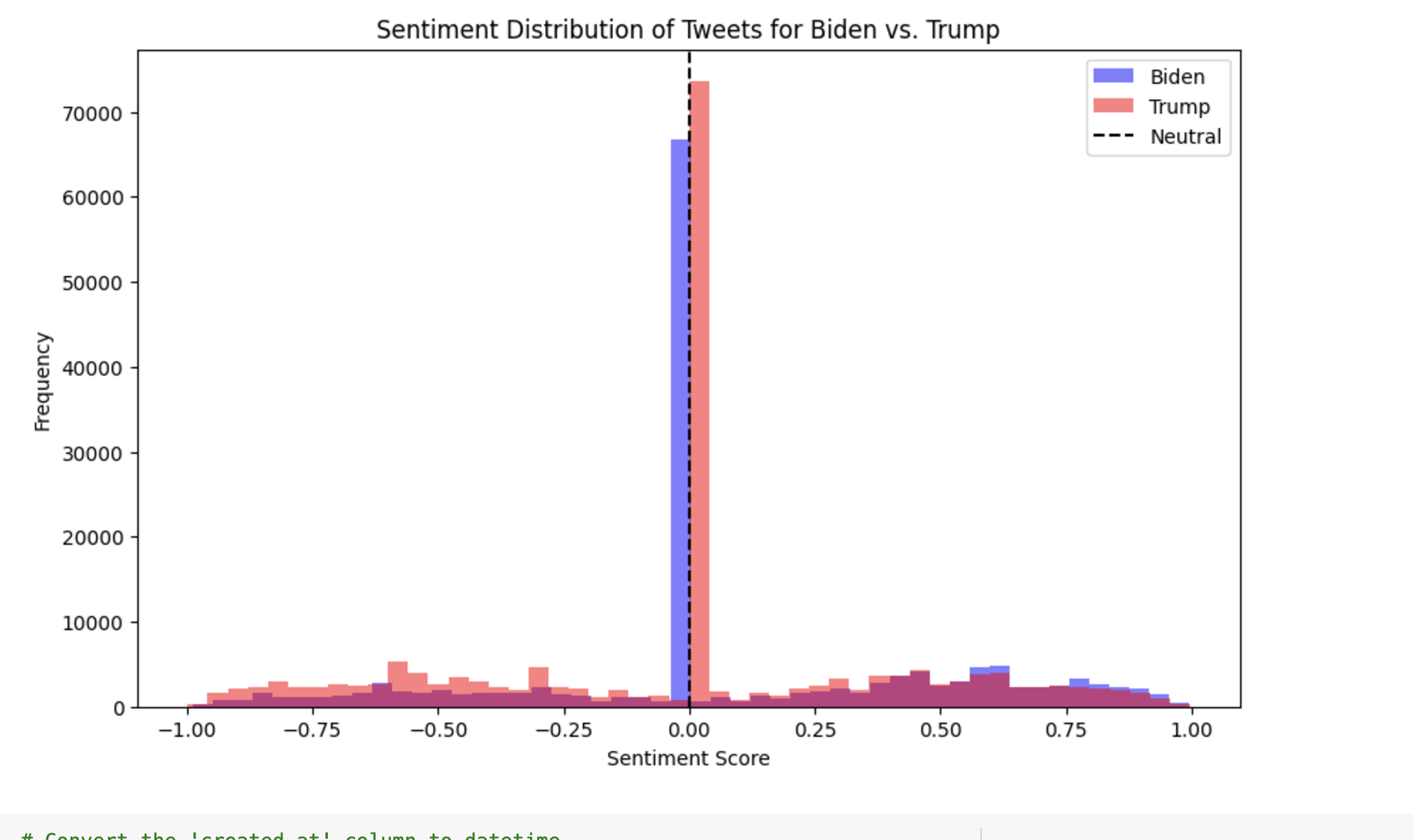
### Model Evaluation

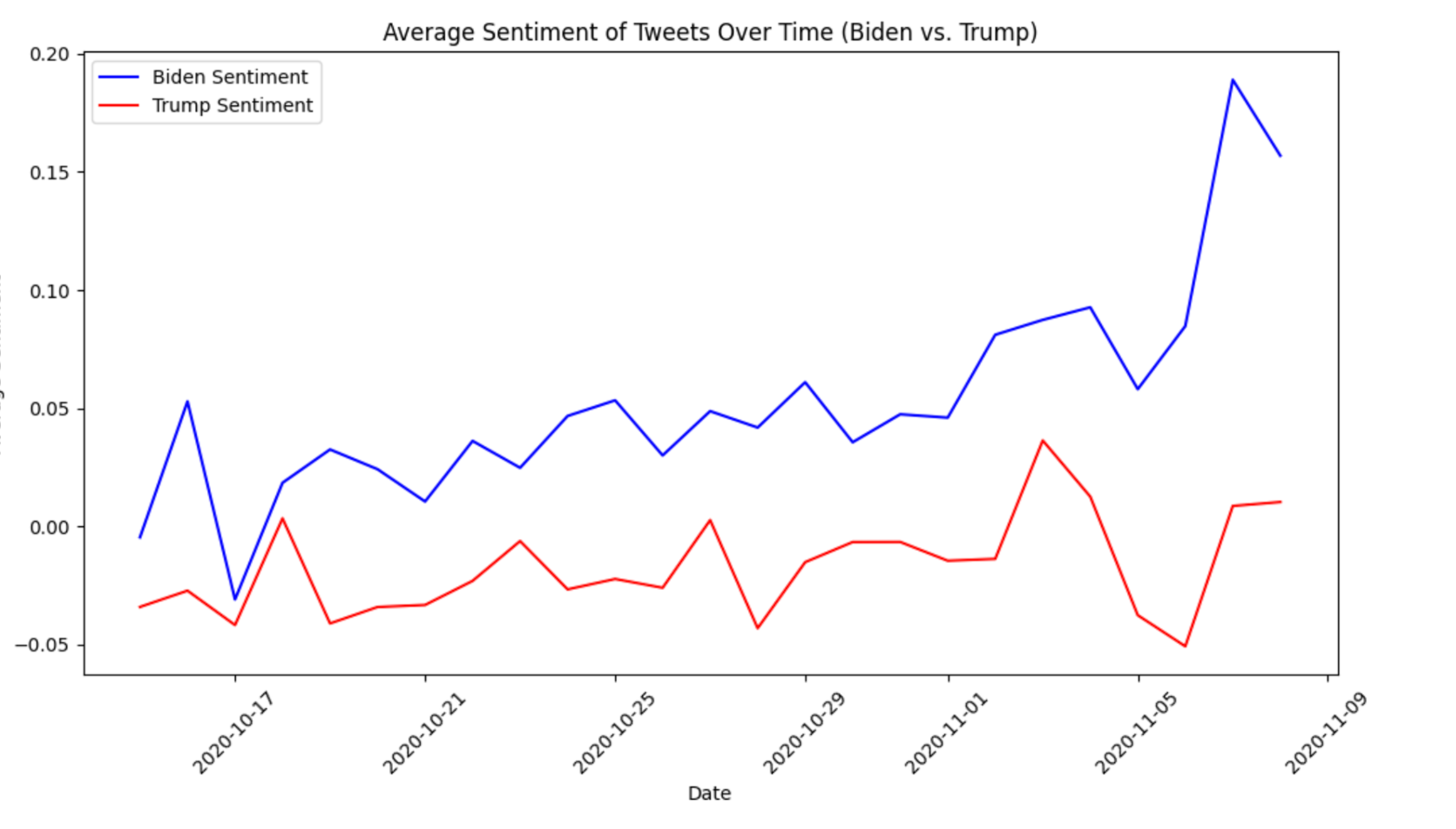
Accuracy Comparison: Compare accuracy scores across models using a bar plot. Annotate the plot with accuracy values. Detailed Metrics: Extract precision, recall, and F1-scores for each class and model. Visualize metrics using a grouped bar plot.

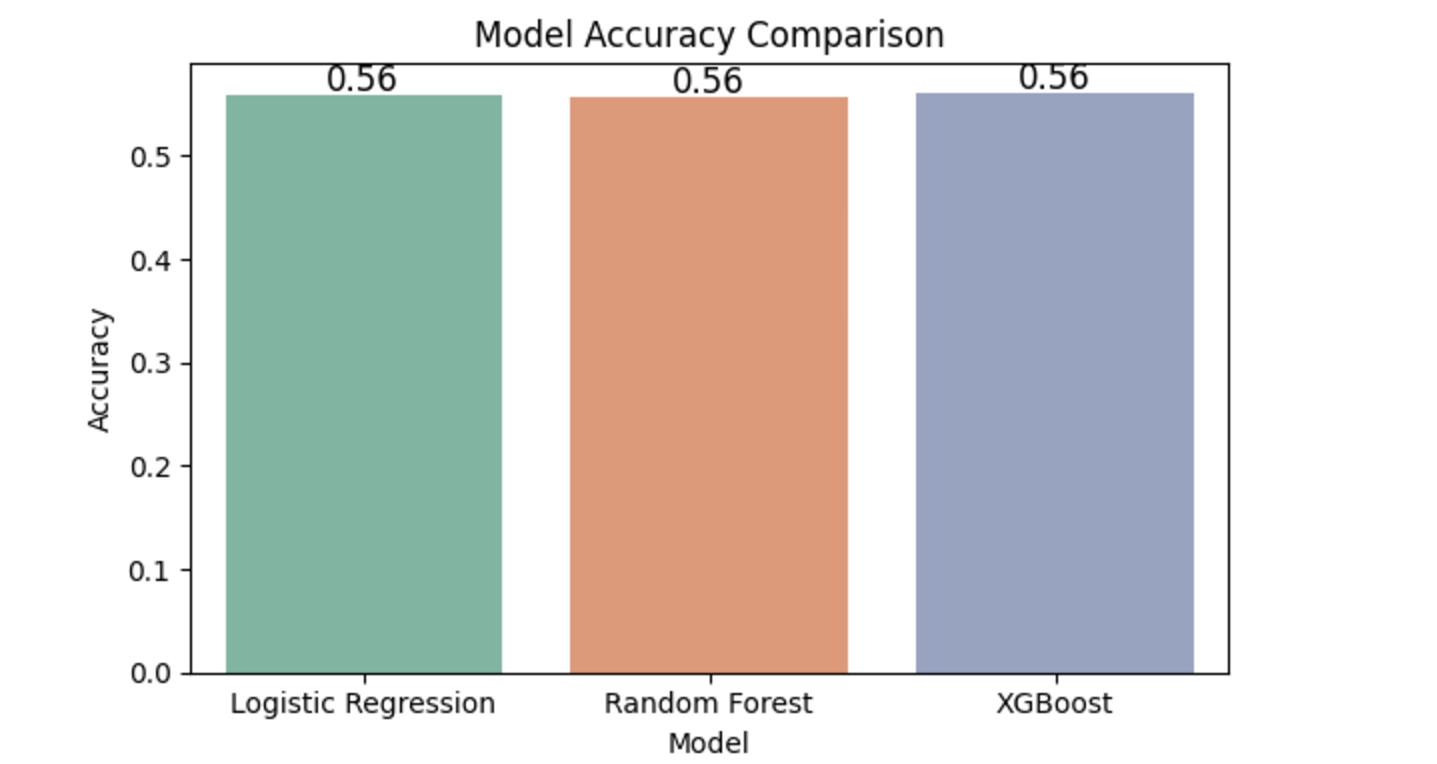
### **Anticipated Results**

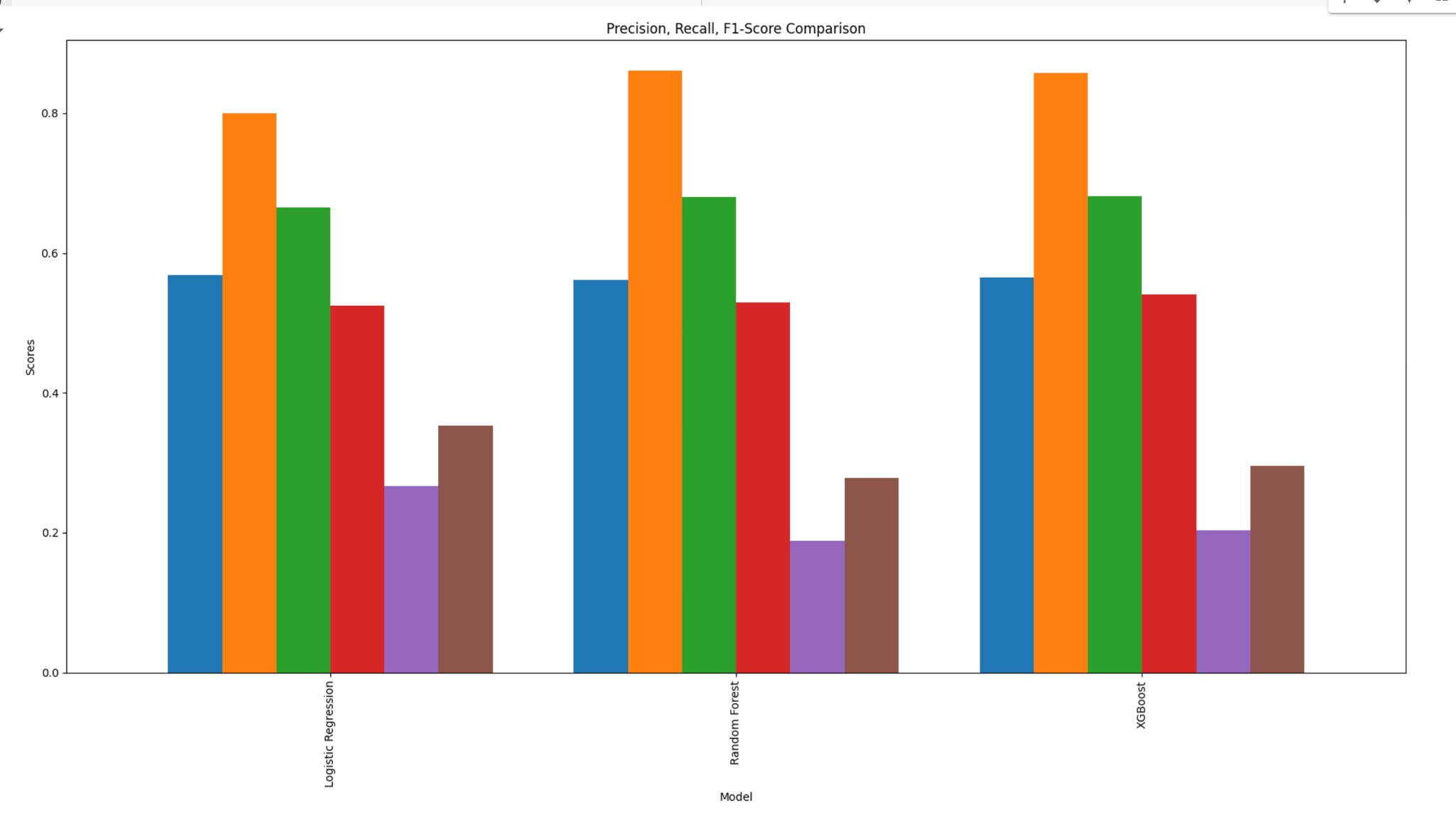
Our anticipated results for this are that it will capture shifts in public opinion and common emotions (e.g., support, criticism, frustration) associated with candidates. Outcome build a model capable of forecasting election results based on tweet sentiment. Identify impactful hashtags, topics, and tweets that shaped voter perspectives. Apply the developed model to 2024 election data to evaluate changes in social media’s influence over time. Mitigate biases caused by echo chambers, bots, and skewed data representation for more accurate predictions.

**Our Results**









### 

### 

### 

### 

### 

### 

### Conclusion

This study applied machine learning models and NLP techniques to analyze tweets about Donald Trump and Joe Biden, revealing several key insights. Overall, tweets about Joe Biden had slightly more positive sentiment compared to those about Donald Trump. Sentiment trends fluctuate daily, often in response to political events or media coverage, with periods of extreme sentiment reflecting significant public engagement.

In terms of model performance, XGBoost outperformed both Logistic Regression and Random Forest, showing the highest accuracy and recall for sentiment classification. Logistic Regression provided a reliable baseline, while Random Forest offered robustness against overfitting.

The analysis demonstrated that sentiment analysis can effectively track shifts in public opinion over time, providing valuable insights for political campaigns and public engagement strategies. XGBoost, in particular, proved to be a powerful tool for identifying patterns and trends in social media sentiment.

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### References and Citation

1. [US Election 2020 Tweets Dataset](https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets/data?select=hashtag_joebiden.csv)
2. Ramteke, Jyoti, et al. "Election result prediction using Twitter sentiment analysis." *2016 International Conference on Inventive Computation Technologies (ICICT)*, IEEE, 2016.
3. Batra, Payal Khurana, et al. "Election result prediction using twitter sentiments analysis." *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, IEEE, 2020.