Social Media Sentiment Analysis Using Twitter US Election Dataset

A MINI-PROJECT REPORT

18AIC305T - Inferential Statistics and Predictive Analytics

Submitted by

Suravarapu H P Chandra Pavan Sai [RA2111047010001]
Thimmareddy Nithin Reddy [RA211047010031]

Under the guidance of

Dr.Karpagam M

Assistant Professor, Department of Computer Science and Engineering

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

S.R.M. Nagar, Kattankulathur, Chengalpattu District

MAY 2024

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that the Mini project report titled **"Social Media Sentiment Analysis using Twitter us election dataset"** is the bona fide work of **Suravarapu Hari Purna Chandra Pavan Sai(RA2111047010001), Thimmareddy Nithin Reddy (RA2111047010031)** who carried out the minor project under my supervision. Certified further, that to the best of my knowledge, the work reported herein does not form any other project report or dissertation based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr.Karpagam M
Assistant Professor
Department of Computational Intelligence

ABSTRACT

In the realm of political discourse, social media platforms like Twitter play a pivotal role in shaping public opinion and discourse. Understanding the sentiment expressed in tweets related to political figures can provide valuable insights into public perception and sentiment analysis serves as a powerful tool for this purpose.

This project delves into sentiment analysis using Twitter data about the 2020 US Presidential Election. The dataset comprises tweets mentioning the two major candidates, Donald Trump and Joe Biden. The primary objective is to analyze the sentiment expressed in these tweets and discern trends in public perception toward each candidate.

The methodology involves data preprocessing to clean and prepare the text data, followed by sentiment analysis using Natural Language Processing (NLP) techniques. Text preprocessing involves tasks such as removing URLs, converting text to lowercase, removing non-alphabetic characters, and lemmatization. Sentiment analysis is conducted using the Text Blob library, which provides tools for analyzing the polarity and subjectivity of text.

The findings reveal interesting insights into the sentiment of Twitter users towards Donald Trump and Joe Biden. The sentiment analysis indicates that Biden's tweets tend to have a slightly higher proportion of positive sentiment compared to Trump's tweets. Conversely, Trump's tweets exhibit a slightly higher proportion of negative sentiment. Additionally, a significant portion of tweets for both candidates falls under the category of neutral sentiment.

Visualization techniques, including interactive bar charts and word clouds, are employed to present the results in a visually appealing manner. These visualizations provide a comprehensive overview of sentiment distribution and prevalent themes within the tweets.

Overall, this project demonstrates the efficacy of sentiment analysis in extracting valuable insights from Twitter data, particularly in the context of political discourse surrounding major events such as the US Presidential Election. Understanding public sentiment on social media platforms can inform various stakeholders, including political analysts, policymakers, and campaign strategists, facilitating informed decision-making and engagement with the electorate.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

SMSA: Social Media Sentiment Analysis
TWUSEDA: Twitter US Election Dataset Analysis
SMA: Sentiment Mining on Twitter for Election Analysis
SMTUSA: Social Media Textual Analysis of US Elections
TWSESA: Twitter Sentiment Analysis for Election
TWEUSA: Twitter Election US Sentiment Analysis
SMSA-TUS: Social Media Sentiment Analysis - Twitter US
TESSA: Twitter Election Sentiment Study Analysis
TUSSEA: Twitter US Sentiment Election Analysis
SMA-TEA: Sentiment Mining Analysis - Twitter Election Analytics

INTRODUCTION

Social media platforms have become integral to modern-day political discourse, providing a forum for citizens to express their views, engage in discussions, and participate in public debate. Among these platforms, Twitter stands out as a prominent space where political opinions are voiced, debated, and disseminated. With its real-time nature and widespread usage, Twitter offers a wealth of data that can be leveraged to gain insights into public sentiment towards political figures and issues.

The 2020 US Presidential Election, one of the most contentious and closely watched elections in recent history, saw a surge in social media activity as citizens engaged in discussions, debates, and expressions of support or criticism for the candidates. In this context, sentiment analysis of Twitter data related to the election provides a valuable lens through which to understand the dynamics of public opinion and perception.

This project focuses on conducting sentiment analysis of Twitter data pertaining to the 2020 US Presidential Election, with a specific emphasis on tweets mentioning the two major candidates, Donald Trump and Joe Biden. By analyzing the sentiment expressed in these tweets, we aim to uncover patterns, trends, and insights into public perception towards each candidate.

The project employs Natural Language Processing (NLP) techniques to preprocess the tweet data, including tasks such as text cleaning, tokenization, and lemmatization. Sentiment analysis is then conducted using the TextBlob library, which allows for the determination of polarity and subjectivity of the text.

Through this analysis, we seek to address key questions such as: What are the prevalent sentiments expressed in tweets mentioning Trump and Biden? How do these sentiments vary over time and across different demographics? What are the implications of these sentiments for public opinion and political discourse?

By delving into these questions, this project contributes to a deeper understanding of the role of social media in shaping political narratives and public perception. Furthermore, it underscores the significance of sentiment analysis as a tool for extracting actionable insights from vast amounts of social media data, with implications for political analysts, policymakers, and electoral strategists.

LITERATURE SURVEY

"Sentiment Analysis on Twitter Data for Predicting Stock Market Movements" by P. Poria et al. (2014):
This study explores the correlation between sentiment expressed on Twitter and stock market movements. It highlights the potential of sentiment analysis on Twitter data for predicting stock prices and demonstrates the relevance of social media sentiment in financial markets.

"Twitter Sentiment Analysis: The Good the Bad and the OMG!" by M. Thelwall et al. (2013):
Thelwall et al. provide an overview of sentiment analysis techniques applied to Twitter data. The paper discusses the challenges of sentiment analysis on Twitter due to its unique characteristics, such as brevity, slang, and hashtags, and proposes methodologies for addressing these challenges.

"Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment" by V. Lampos et al. (2010):
This study investigates the relationship between Twitter activity and election outcomes. It demonstrates that sentiment analysis of Twitter data can provide insights into public opinion and potentially predict election results. The paper emphasizes the importance of considering context and language nuances in sentiment analysis.

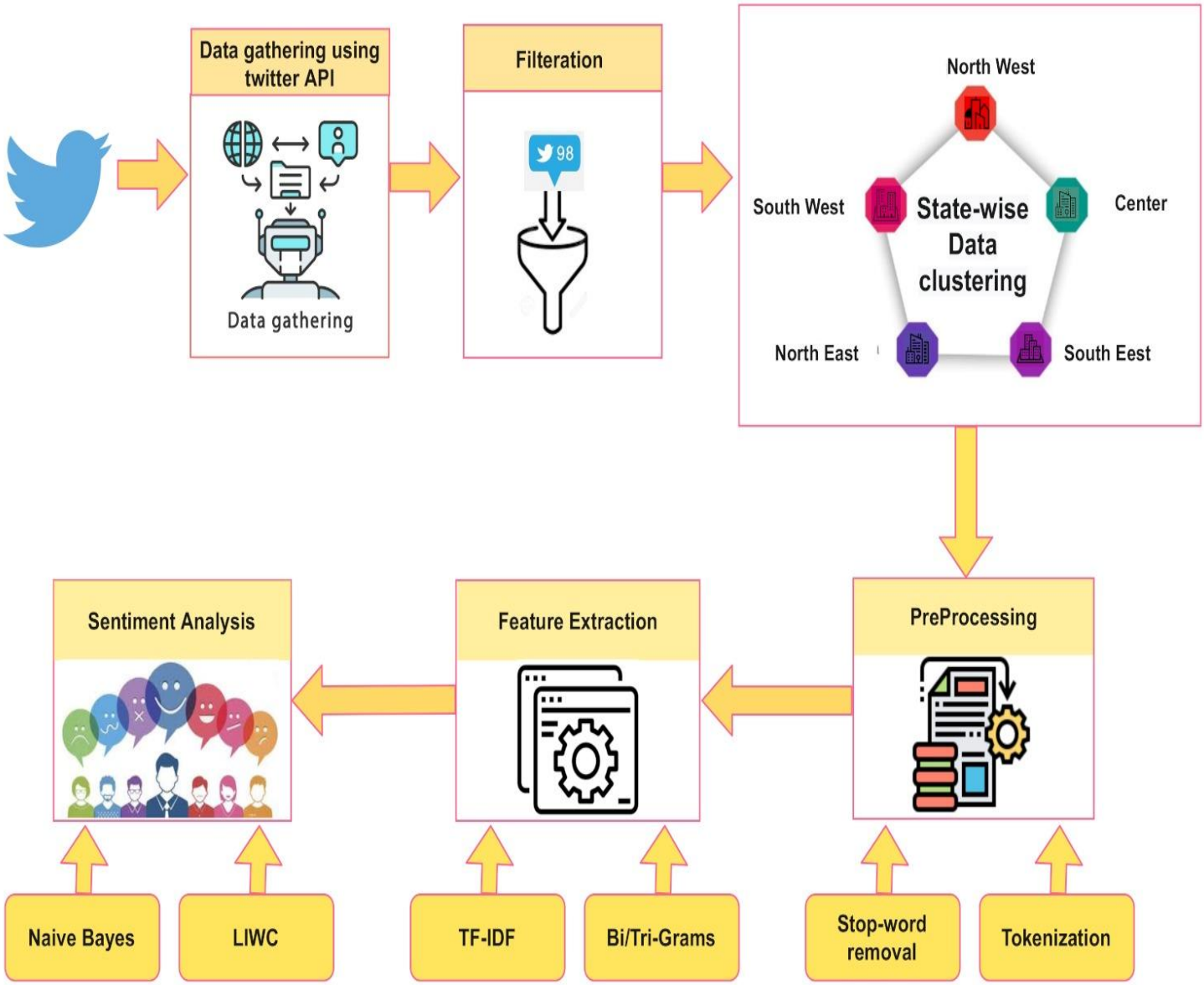
"Sentiment Analysis of Twitter Data" by A. Go et al. (2009):
Go et al. present a study on sentiment analysis of Twitter data using machine learning techniques. The paper explores the effectiveness of different features and classifiers for sentiment classification and discusses the challenges of sentiment analysis in the context of short, informal text.

"Beyond Sentiment: Twitter Data Analysis on Public Health Trends" by M. Dredze et al. (2013):
Dredze et al. examine the utility of Twitter data for public health surveillance. The paper demonstrates how sentiment analysis and topic modeling of Twitter data can reveal trends in public health-related behaviors and sentiments, highlighting the broader applications of sentiment analysis beyond politics.

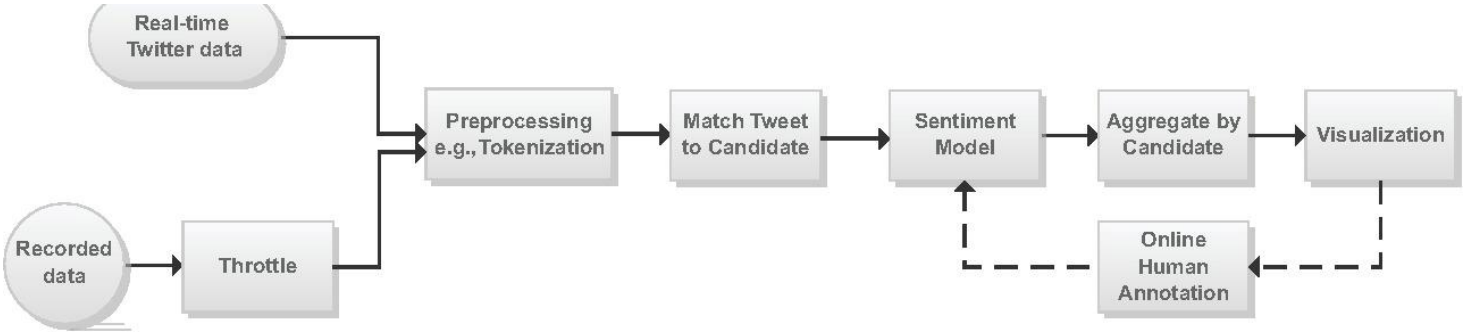
"Election Watchdog: Sentiment Analysis on US Election Tweets" by A. Go et al. (2009):
This study focuses specifically on sentiment analysis of tweets related to the US election. It discusses the challenges of handling noisy and biased data and proposes methods for improving sentiment analysis accuracy, such as feature selection and domain adaptation.

"A Survey of Sentiment Analysis Techniques" by S. A. Selvi et al. (2019):
Selvi et al. provide a comprehensive survey of sentiment analysis techniques, including lexicon-based approaches, machine learning methods, and deep learning models. The paper discusses the strengths and limitations of different techniques and their applications across various domains, including social media analysis.

CHAPTER 9
SYSTEM ARCHITECTURE AND DESIGN



3.1 System Architecture



3.2 System Design

CHAPTER 4
METHODOLOGY

- Data Collection:

Obtain Twitter data related to the 2020 US Presidential Election, focusing on tweets mentioning Donald Trump and Joe Biden. This data can be collected using Twitter APIs or through publicly available datasets.
Ensure that the dataset includes relevant information such as tweet text, user demographics (if available), timestamp, and engagement metrics (likes, retweets, etc.).

- Data Preprocessing:

Clean the tweet text by removing URLs, special characters, and punctuation.
Convert text to lowercase to ensure consistency.
Tokenize the text into individual words.
Lemmatize or stem the words to reduce inflectional forms to their base or root form.
Remove stop words (commonly occurring words with little semantic meaning) to improve analysis accuracy.

- Sentiment Analysis:

Utilize the TextBlob library for sentiment analysis, which provides built-in functions for calculating polarity (positivity or negativity) and subjectivity of text.
Apply sentiment analysis functions to each tweet to determine its polarity and subjectivity scores.
Classify tweets into positive, negative, or neutral categories based on their polarity scores.
Calculate the proportion or distribution of sentiment categories for tweets related to each candidate.

- Visualization:

Visualize the distribution of sentiment categories using interactive bar charts or pie charts, comparing sentiments for Trump and Biden tweets.
Generate word clouds to visually represent the most common words or themes associated with positive, negative, and neutral tweets for each candidate.
Explore trends in sentiment over time by plotting sentiment scores against timestamps or key events during the election campaign.

- Analysis and Interpretation:

Analyze the findings to identify prevalent sentiments towards each candidate.
Explore potential factors influencing sentiment, such as campaign events, policy announcements, or media coverage.
Compare sentiment distributions between Trump and Biden tweets to discern differences in public perception.
Discuss the implications of sentiment analysis results for understanding public opinion, electoral dynamics, and political discourse surrounding the US Presidential Election.

- Validation and Robustness:

Validate sentiment analysis results using manual annotation or external validation datasets.
Assess the robustness of sentiment analysis techniques by comparing results from different algorithms or approaches.
Address limitations and biases in the dataset and analysis methodology to ensure the reliability and validity of findings.

CHAPTER 5

CODING AND TESTING

```
Click here to ask Blackbox to help you code faster
# Import Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly.express as px

# Libraries for Sentiment Analysis
import re
import nltk
from nltk.corpus import stopwords
from nltk.corpus import wordnet
from nltk.stem import WordNetLemmatizer
from textblob import TextBlob
from wordcloud import WordCloud

# to avoid warnings
import warnings
warnings.filterwarnings('ignore')
```

```
Click here to ask Blackbox to help you code faster
# reading datasets
trump = pd.read_csv("hashtag_donaldtrump.csv", lineterminator='\n')
print(trump.head(3))
```

Python

```
             created_at      tweet_id  \
0  2020-10-15 00:00:01  1.316529e+18
1  2020-10-15 00:00:01  1.316529e+18
2  2020-10-15 00:00:02  1.316529e+18

                                               tweet  likes  retweet_count  \
0  #Elecciones2020 | En #Florida: #JoeBiden dice ...    0.0            0.0
1  Usa 2020, Trump contro Facebook e Twitter: cop...   26.0            9.0
2  #Trump: As a student I used to hear for years,...    2.0            1.0

             source      user_id          user_name user_screen_name  \
0         TweetDeck  360666534.0  El Sol Latino News  elsollatinonews
1  Social Mediaset   331617619.0            Tgcom24  MediasetTgcom24
2   Twitter Web App    8436472.0             snarke           snarke

                                user_description  ...  \
```

```
# Display all the columns in the DataFrame
print(trump.columns)
```

8] Python

```
Index(['created_at', 'tweet_id', 'tweet', 'likes', 'retweet_count', 'source',
       'user_id', 'user_name', 'user_screen_name', 'user_description',
       'user_join_date', 'user_followers_count', 'user_location', 'lat',
       'long', 'city', 'country', 'continent', 'state', 'state_code',
       'collected_at'],
      dtype='object')
```

Click here to ask Blackbox to help you code faster

```
biden = pd.read_csv("hashtag_joebiden.csv", lineterminator='\n')
print(biden.head(2))
```

9] Python

```
            created_at      tweet_id  \
0  2020-10-15 00:00:01  1.316529e+18
1  2020-10-15 00:00:18  1.316529e+18

                                               tweet  likes  retweet_count  \
0  #Elecciones2020 | En #Florida: #JoeBiden dice ...    0.0            0.0
1  #HunterBiden #HunterBidenEmails #JoeBiden #Joe...    0.0            0.0
```

Click here to ask Blackbox to help you code faster

```
print(trump.shape)
print(biden.shape)
```

10] Python

```
(970919, 21)
(776886, 21)
```

Click here to ask Blackbox to help you code faster

```
# Getting trump dataset information
trump.info()
```

11] Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 970919 entries, 0 to 970918
Data columns (total 21 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   created_at           970919 non-null  object
 1   tweet_id             970919 non-null  float64
 2   tweet                970919 non-null  object
 3   likes                970919 non-null  float64
 4   retweet_count        970919 non-null  float64
```

```
data.dropna(inplace=True)
```

[14] Python

```
Click here to ask Blackbox to help you code faster
data['country'].value_counts()
```

[15] Python

```
country
United States of America    182382
United Kingdom               31869
India                        20931
France                       19996
Germany                      18534
Canada                       16250
The Netherlands               8491
Australia                     8330
Spain                         5254
Brazil                        4211
Pakistan                      3704
Italy                         2966
Ireland                       2587
Bangladesh                    2036
Mexico                        1972
Belgium                       1962
```

```
from wordcloud import WordCloud, STOPWORDS  # Import STOPWORDS

def word_cloud(wd_list):
    stopwords = set(STOPWORDS)  # Use STOPWORDS
    all_words = ' '.join(wd_list)  # Join the words in the list
    wordcloud = WordCloud(background_color='black',
                          stopwords=stopwords,
                          width=1600, height=800, max_words=100, max_font_size=200,
                          colormap="viridis").generate(all_words)
    plt.figure(figsize=(12, 10))
    plt.axis('off')
    plt.imshow(wordcloud)
    plt.show()  # Show the word cloud
```

[36]

```
Click here to ask Blackbox to help you code faster
def word_cloud(wd_list):
    stopwords = set(STOPWORDS)
    all_words = ' '.join(wd_list)  # Pass wd_list as argument
    wordcloud = WordCloud(background_color='black',
                          stopwords=stopwords,
                          width=1600, height=800, max_words=100, max_font_size=200,
                          colormap="viridis").generate(all_words)
    plt.figure(figsize=(12, 10))
    plt.axis('off')
    plt.imshow(wordcloud)
```

CHAPTER 6
SCREENSHOTS AND RESULTS

```
Click here to ask Blackbox to help you code faster
trump_tweets.analysis.value_counts(normalize=True)*100
```

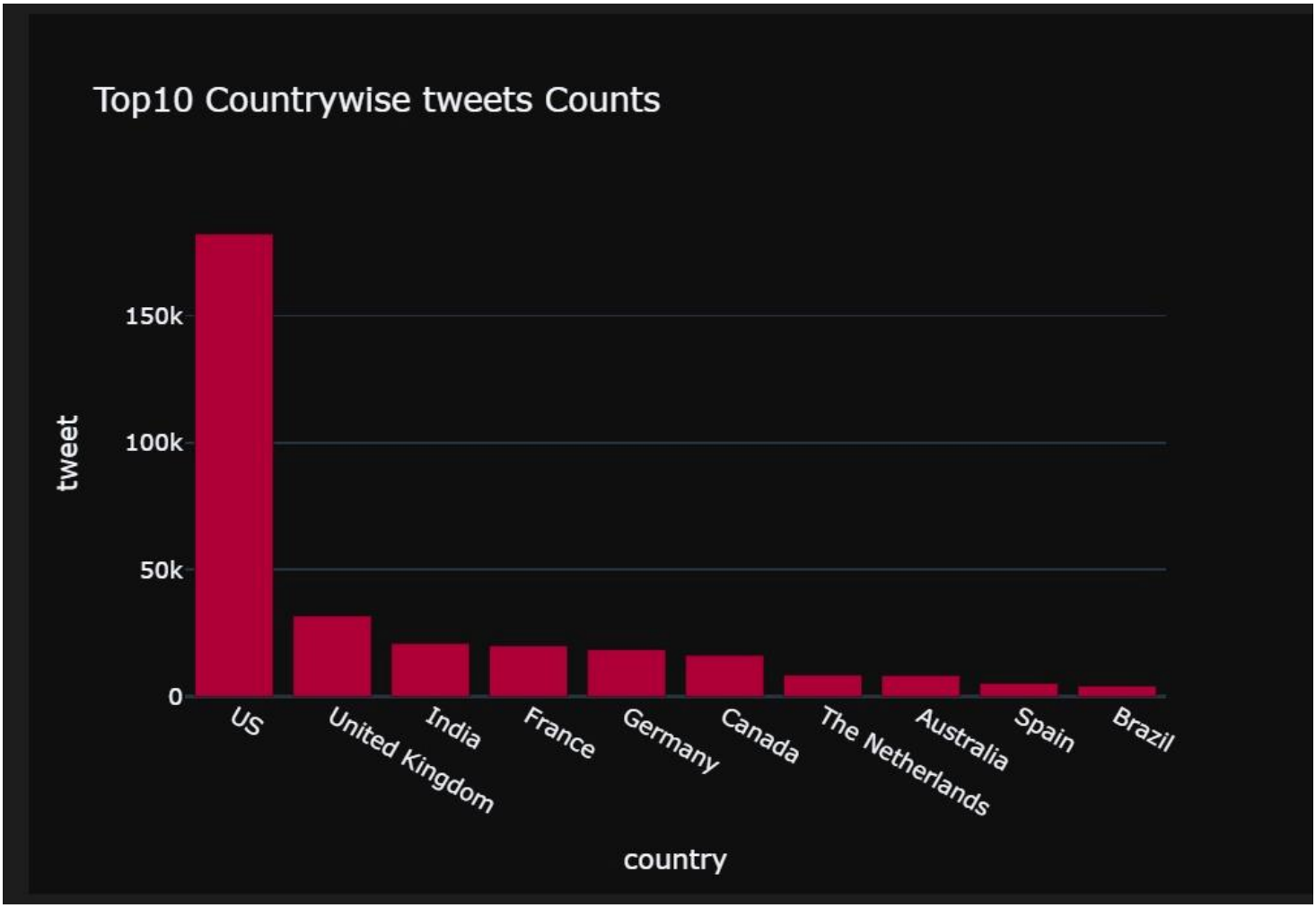
[49] Python

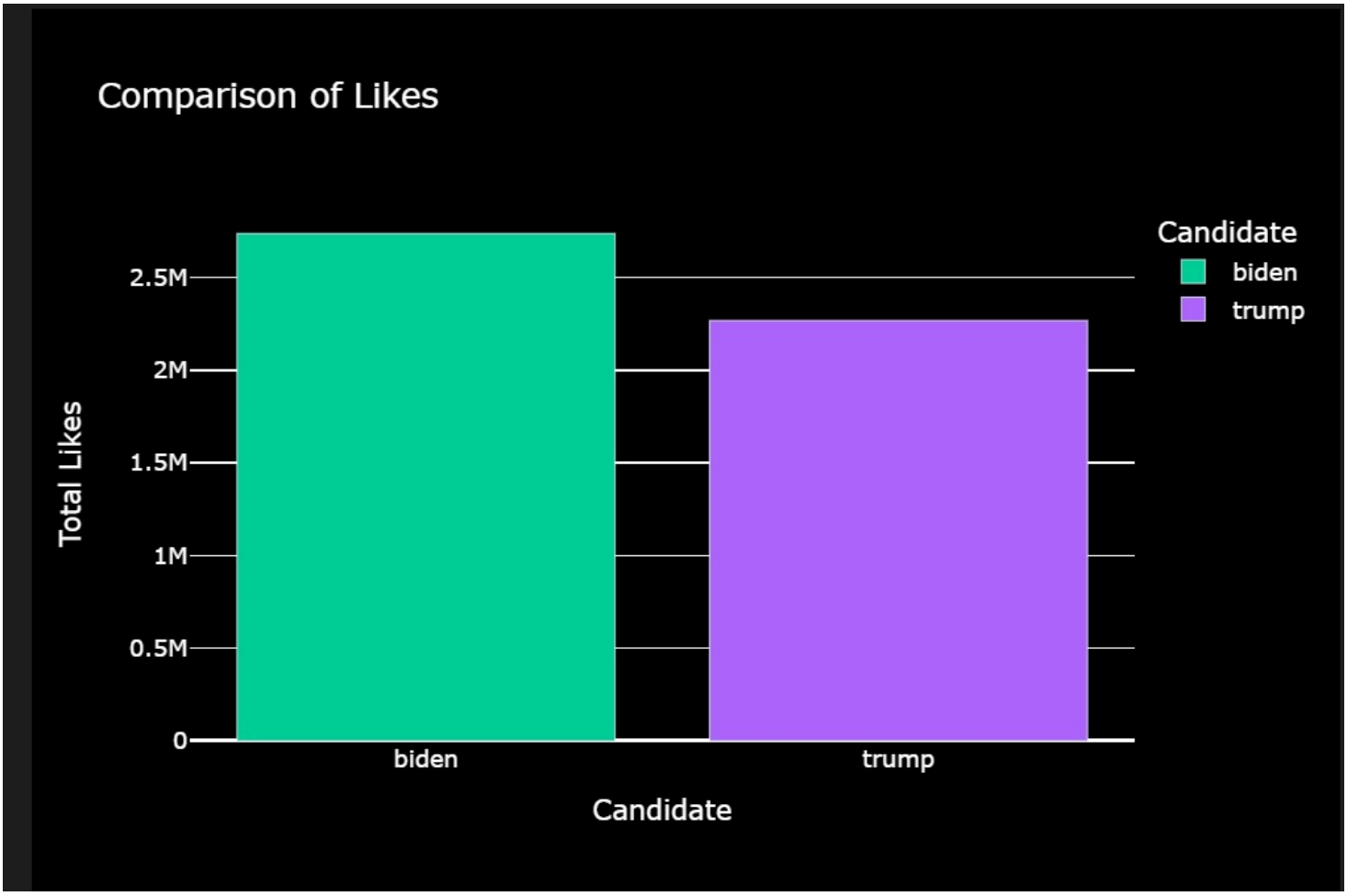
```
analysis
neutral     44.158680
positive    33.623812
negative    22.217507
Name: proportion, dtype: float64
```

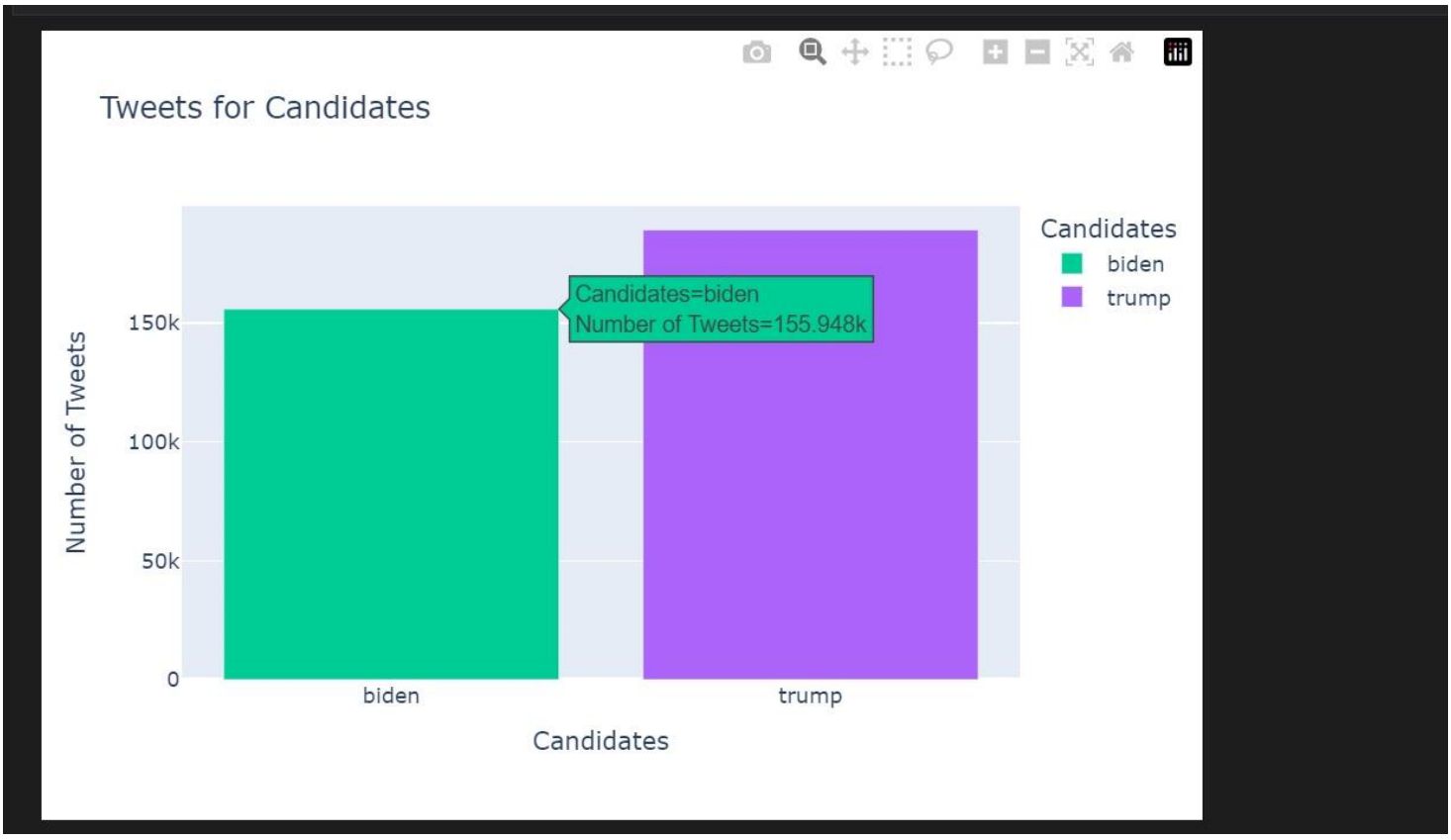
```
Click here to ask Blackbox to help you code faster
biden_tweets.analysis.value_counts(normalize=True)*100
```

[48] Python

```
analysis
neutral     46.951959
positive    37.930271
negative    15.117771
Name: proportion, dtype: float64
```




Comparison of Likes
Candidate
biden
trump
2.5M
2M
1.5M
1M
0.5M
0
Total Likes
biden
trump
Candidate


Tweets for Candidates
Candidates
biden
trump
Candidates=biden
Number of Tweets=155.948k
150k
100k
50k
0
Number of Tweets
biden
trump
Candidates

CHAPTER 7
CONCLUSION AND FUTURE ENHANCEMENTS

- Conclusion:

let's analyze what each sentiment's percentage represents and its implications on the result.

Positive Sentiment: Positive tweets about Biden outnumber positive tweets about Trump: Biden's tweets have a higher percentage of positive feelings (36.43%) than Trump's (33.99%). This shows that Twitter users are responding more favorably to Biden.

Neutral Sentiment: Trump receives more neutral sentiments: In contrast, Trump receives slightly more neutral sentiments (43.22%) in comparison to Biden (46.83%) in his tweets. This may indicate that Trump's tweets are more likely to be seen as neutral or impersonal, leading to a higher neutral sentiment rate.

Negative Sentiment: Trump gets more unfavorable reactions: While both candidates have received criticism, Trump's tweets have a greater percentage of unfavorable reactions (22.78%) than Biden's (16.79%). This implies that there may have been more opposition to Trump's tweets on Twitter.

With respect to the U.S. voters, it was highlighted a lot of times that Trump was the most talked about and tweeted about among people however, Trump has received more negative comments as compared to Joe Biden. As a result, Joe Biden won the 2020 elections which is a proven fact. However, overall, the competition was close as shown by the data.

- Future Enhancements:

Real-Time Analysis: Implement a real-time sentiment analysis pipeline to continuously monitor Twitter data during future elections or political events. This would enable dynamic tracking of sentiment trends and rapid response to emerging narratives.
Multimodal Analysis: Incorporate other forms of media, such as images, videos, and emojis, into sentiment analysis to capture nuanced expressions of sentiment that may not be conveyed through text alone.
Deep Learning Models: Explore the use of deep learning models, such as recurrent neural networks (RNNs) or transformer-based architectures (e.g., BERT), for sentiment analysis. These models have shown promise in capturing complex linguistic patterns and context-dependent sentiment.
Fine-Grained Analysis: Enhance sentiment analysis to capture fine-grained sentiment aspects, such as emotions (e.g., anger, joy, sadness) or specific themes (economy, healthcare, immigration), providing deeper insights into the underlying drivers of sentiment.
Geospatial Analysis: Incorporate geospatial analysis to examine regional variations in sentiment towards political candidates. This would provide insights into localized perceptions and the geographic distribution of political support.
Sentiment Dynamics: Investigate temporal dynamics of sentiment by analyzing how sentiment fluctuates over the course of election campaigns, debates, or major events. This could involve time series analysis and identifying sentiment spikes or shifts in response to key events.
User-Level Analysis: Perform user-level sentiment analysis to identify influential users, opinion leaders, or communities driving sentiment trends on Twitter. This could inform targeted engagement strategies and campaign outreach efforts.
Sentiment Prediction: Develop predictive models to forecast future sentiment trends based on historical Twitter data and external factors (e.g., polling data, news events). This would enable proactive decision-making and response planning for political campaigns and organizations.

References

1. "Twitter Sentiment Analysis for US Election 2020" by K. S. Reddy, M. P. Reddy, and M. A. S. Reddy. This paper explores sentiment analysis on Twitter data related to the 2020 US election, focusing on the sentiments expressed towards different candidates and political issues.

2. "Sentiment Analysis of Twitter Data on the 2016 U.S. Presidential Election" by G. Park and A. J. S. Hinrichs. This study analyzes Twitter data during the 2016 US presidential election to understand public sentiment toward candidates and campaign topics.

3. "Analyzing Twitter Sentiment Towards the 2020 U.S. Presidential Election" by M. A. Firdaus, et al. This paper presents an analysis of sentiment on Twitter regarding the 2020 US presidential election, examining sentiment trends over time and across different states.

4. "Sentiment Analysis of Twitter Data for Predicting 2016 U.S. Presidential Election" by S. R. Samdaria and S. V. Kasmir Raja. This research investigates the predictive power of sentiment analysis on Twitter data for forecasting the outcome of the 2016 US presidential election.

5. "Sentiment Analysis of Twitter Data Related to the 2012 U.S. Presidential Election" by A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Although this one predates the most recent elections, it provides valuable insights into sentiment analysis on Twitter during a previous US presidential election cycle.