IBM-322 Group Assignment Report

(Effectiveness of Sentiment Analysis for predicting election outcomes)

Submitted by – Group 27.

Suryansh Bhatnagar (21116092),

Chetan Sharma (21116031),

Rahul Negi (21116078), &

Yawalkar Ajinkya Ganpati (21116108)

Abstract: Twitter, a popular social media platform, has been a hotbed of political discussions during elections around the globe. The 2020 US Presidential election *(held on Tuesday 3 November, 2020)* was one of the most divisive and contentious in recent history. Twitter is a useful platform for understanding public opinion during the 2020 US Presidential election.

Sentiment analysis can be used to track changes in public opinion over time and to identify key events that influenced public opinion**.** This study analyzes the sentiment of tweets related to the 2020 US Presidential election to gain insights into public opinion. Using a VADER lexicon-based sentiment analysis approach, we analyzed over a million tweets related to the two main candidates, Joe Biden and Donald Trump.

Dataset Used for Analysis: *"hashtag_donaldtrump.csv"* and *"hashtag_joebiden.csv"* have been used for this project.

Original Source of Dataset:

https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets *Google Drive link:*

https://drive.google.com/drive/folders/1f6VjLuE1f_XIz2jK9LK_WYx1CZ-lnO9x?usp=sharing

About Sentiment Analysis and its use on tweets

Note: Here, we use VADER for Sentiment Analysis on data cleaned by various methods enlisted later in this report and illustrated in the Interactive Python Notebook

Sentiment analysis, also known as opinion mining, is a computational technique that involves determining and categorizing the sentiment expressed in a piece of text. With the increasing prevalence of social media platforms, sentiment analysis has become a crucial tool for understanding public opinion and consumer sentiment. One popular tool for sentiment analysis, particularly on short and informal texts like tweets, is the VADER (*Valence Aware Dictionary and sEntiment Reasoner*) sentiment analysis tool.

- ***Sentiment Analysis Fundamentals:*** Sentiment analysis is a natural language processing (NLP) task that involves classifying a piece of text as positive, negative, or neutral based on the emotions or opinions expressed. It has applications in various fields, including marketing, customer feedback analysis, and social media monitoring. Traditional approaches to sentiment analysis relied on machine learning algorithms trained on labeled datasets, but more recent methods, like rule-based approaches, have gained popularity due to their simplicity and efficiency.

- ***VADER Sentiment Analysis:*** VADER is a rule-based sentiment analysis tool designed for social media texts, making it well-suited for analyzing sentiments in tweets. Developed by researchers at the Georgia Institute of Technology, VADER employs a pre-built lexicon that assigns sentiment scores to individual words. It considers punctuation, capitalization, and conjunctions to understand the context and intensity of sentiments expressed in a given text.
  One key advantage of VADER is its ability to handle both polarities and intensity of sentiments. It not only classifies the sentiment as positive, negative, or neutral but also provides a compound score that represents the overall sentiment intensity. This feature is especially valuable in capturing the

nuances of sentiments expressed in short and informal texts like tweets.

- ***Application of VADER in Tweet Analysis:*** Twitter, being a microblogging platform, is a treasure trove of real-time public opinions and sentiments. Analyzing tweets using VADER can provide valuable insights for businesses, researchers, and policymakers. Here are some applications of VADER in tweet analysis:
    1. ***Political Opinion Tracking (Used here):*** During elections or major political events, VADER can be employed to analyze tweets related to candidates or political issues. This helps in understanding public opinion trends and predicting potential shifts in sentiment.
    2. ***Event Sentiment Analysis:*** VADER can be applied to analyze tweets related to events, such as conferences, sports games, or entertainment shows. Event organizers can gain insights into the audience's response and adjust strategies accordingly.
    3. ***Brand Sentiment Analysis:*** Companies can use VADER to monitor tweets related to their brand and gauge the overall sentiment of their customers. This information is crucial for managing brand reputation and addressing customer concerns promptly.
    4. ***Product Feedback Monitoring:*** Businesses can use VADER to analyze tweets containing feedback about their products. This enables them to identify areas for improvement and respond to customer concerns in a timely manner.

    and many more

VADER may struggle with sarcasm, irony, or context-dependent sentiments. Additionally, it requires periodic updates to account for emerging slang and language shifts on social media platforms.

Predicting election outcome using Sentiment Analysis

Dataset Pre-processing:

- This involved loading the dataset containing tweets which used hashtags of the two candidates and then processing our data to focus on relevant information for our analysis.
- Irrelevant columns are dropped from the original dataset, to remove information irrelevant for our analysis. The dataset is filtered to include only the columns required for our analysis.
- We utilized Pandas for data manipulation and filtering based on relevant criteria.

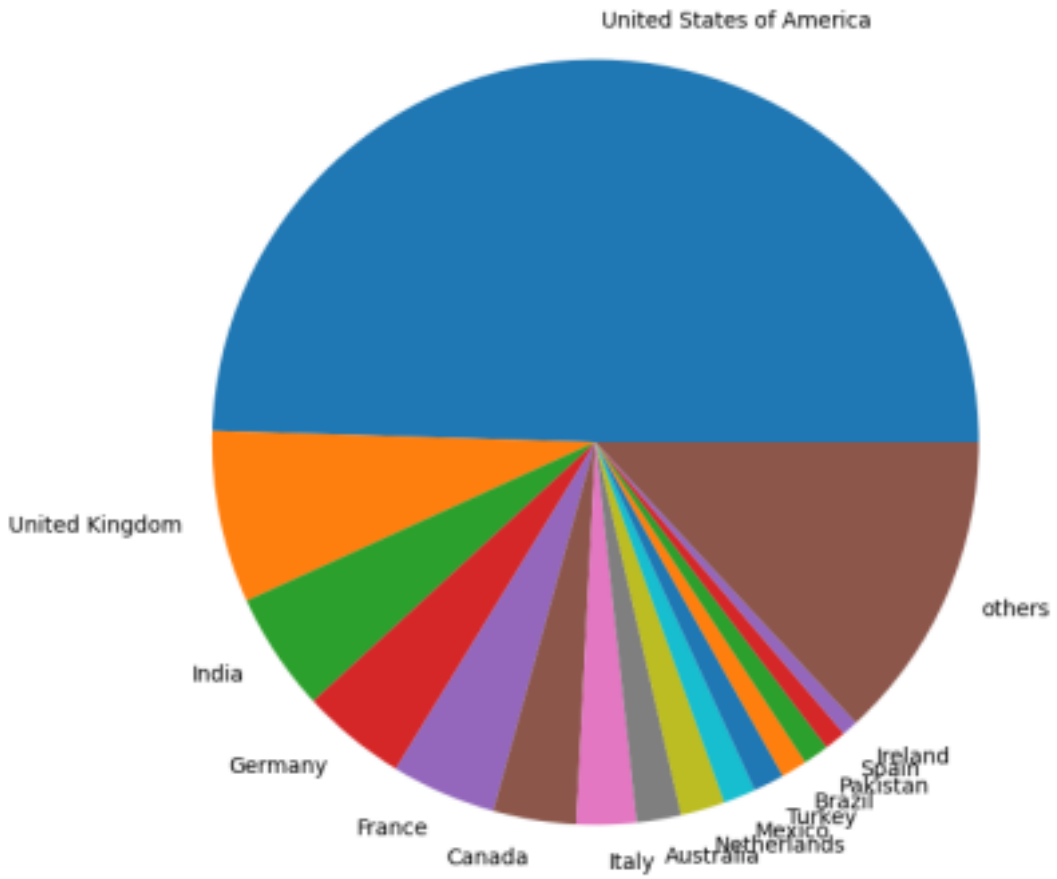
Provided Data Format:

Modified Data Format (after pre-processing):

	tweet	source	user_id	user_name	user_join_date	user_followers_count	country	state	state_code	Tweet_subject
0	#Elecciones2020	En #Florida: #JoeBiden dice ...	TweetDeck	3.606885e+08	El Sol Latino News	2011-06-23 15:33:45	1660.0	United States of America	Florida	FL	Trump
1	Usa 2020, Trump contro Facebook e Twitter: cop...	Social Mediaset	3.316176e+08	Tgcom24	2011-07-08 13:12:20	1067661.0	NaN	NaN	NaN	Trump
2	#Trump: As a student I used to hear for years,...	Twitter Web App	8.436472e+06	snarks	2007-06-26 05:56:11	1165.0	United States of America	Oregon	OR	Trump
3	2 hours since last tweet from #Trump! Maybe he...	Trumpytweeter	8.283559e+17	Trumpytweeter	2017-02-05 21:32:17	32.0	NaN	NaN	NaN	Trump
4	You get a tie! And you get a tie! #Trump 's na...	Twitter for iPhone	4.741380e+07	Rana Abtar - رنا أبتر	2009-06-15 19:05:35	5393.0	United States of America	District of Columbia	DC	Trump

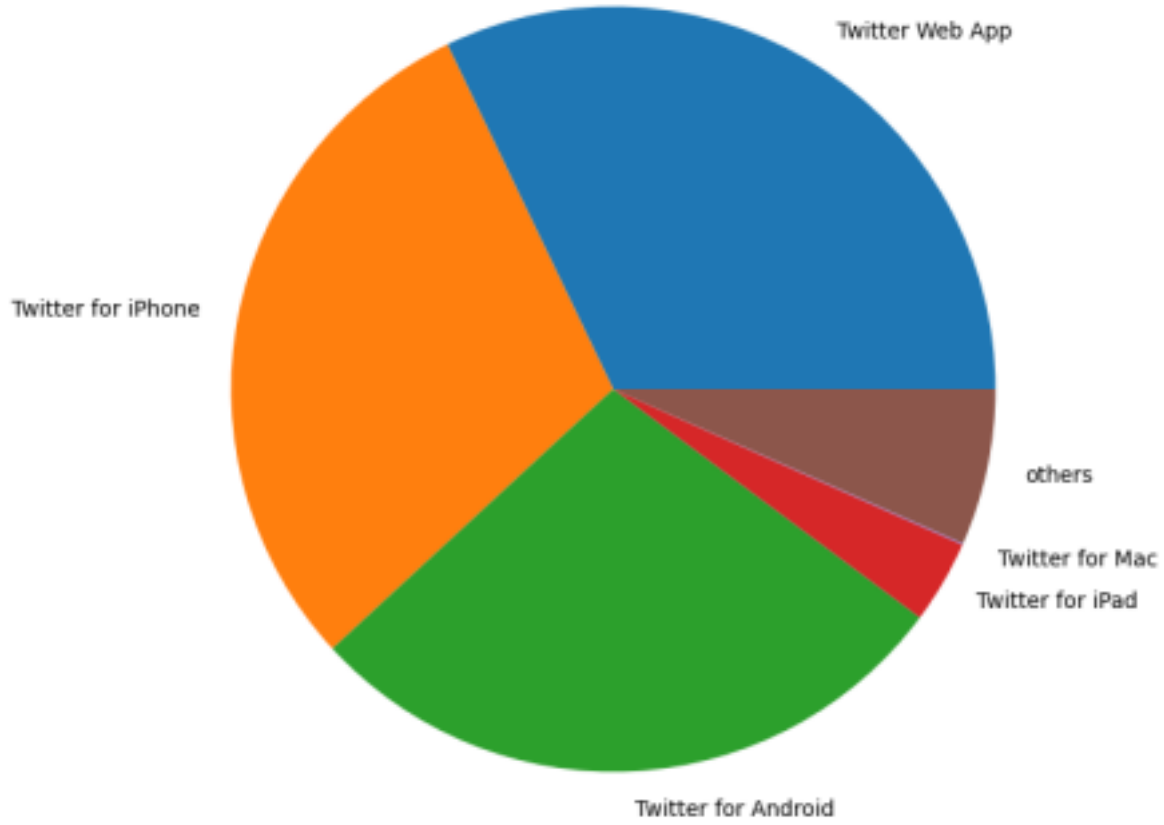
Country of origin of tweets being analysed:

• Visualizing the distribution of tweets by country of origin (We ignored tweets for which country of origin was not available)
• Created a pie chart showing the distribution. Later, we filtered out the tweets which originated from the United States of America to use for our analysis. • Utilized matplotlib library of python to generate the pie chart.



Tweet Source (Platform used for publishing tweet) Filtering

• Filtering tweets based on the source aimed to narrow down the analysis to specific platforms, providing insights into the sentiment from users on different devices.
• We successfully filtered tweets from the sources such as Twitter for iPad, iPhone, etc. removing the others (Sources other than Web App, iPhone, iPad, Android, Mac may represent tweets from bots)
• Created an array of target sources and filtered data using Pandas.

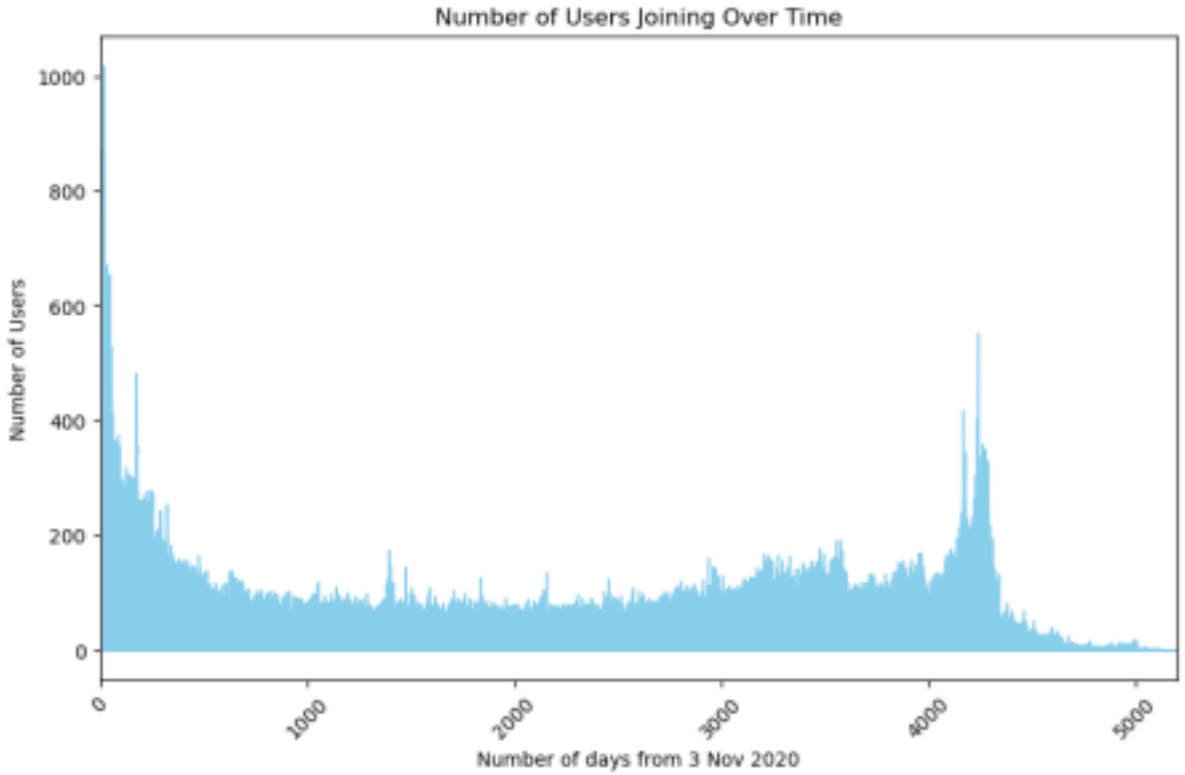


Data Filtering Based on Join Date:

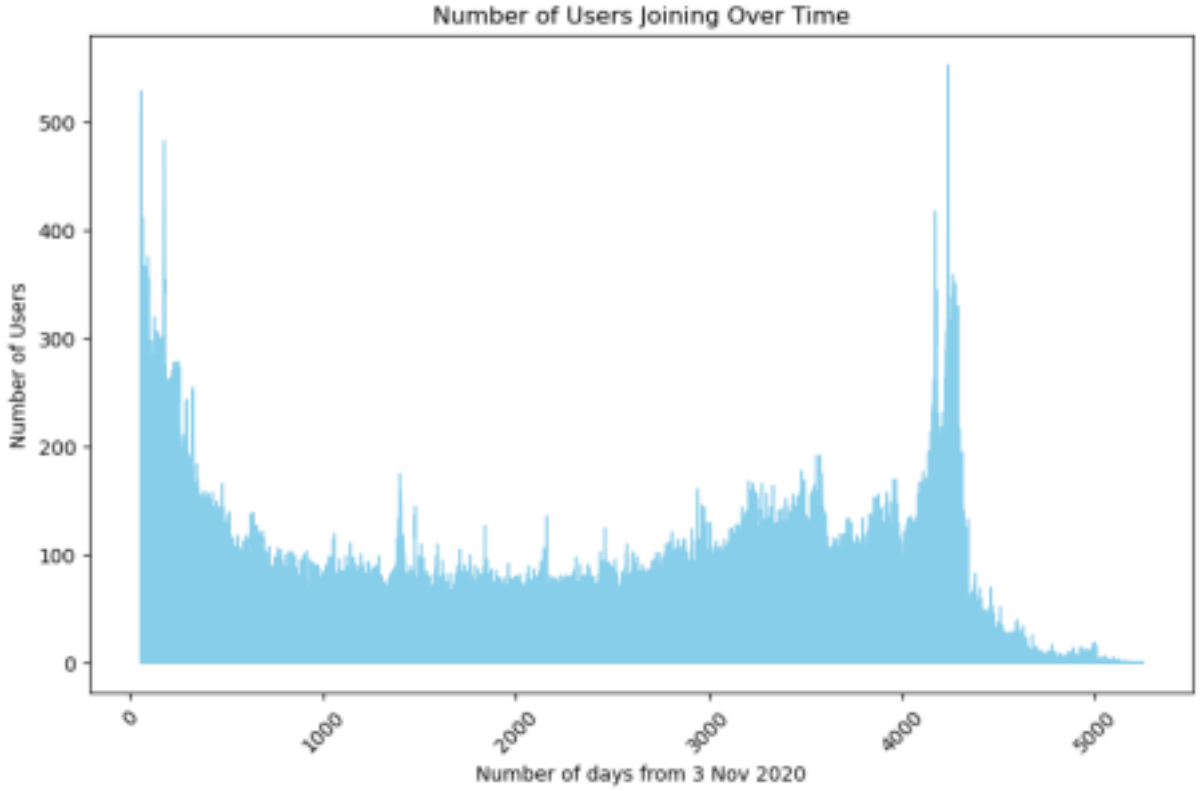
• We filtered data based on user join dates to identify potentially fake accounts. • Accounts which are created within a month of elections are considered to be accounts for spamming.

• Modified the DataFrame removing the users who joined after October 3, 2020.

Before removing the users who joined after October 3, 2020:



After removing the users who joined after October 3, 2020:

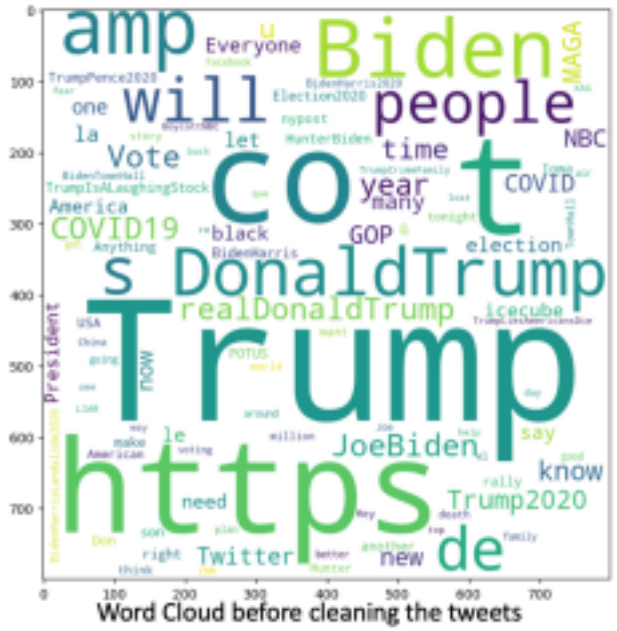


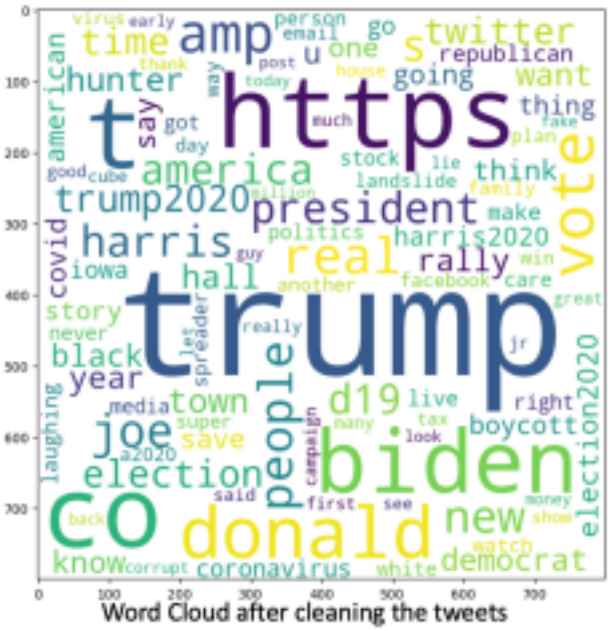
Use of WordCloud:

"WordCloud" simplifies the generation of visually appealing representations of textual data. The "WordCloud" library provides a straightforward way to generate a word cloud. The library allows users to customize various aspects of the word cloud, such as the color scheme, font, and the shape of the cloud itself. The frequency of each word in the text determines the size of the corresponding word in the cloud, with more frequently occurring words appearing larger.

In sentiment analysis, word clouds can reveal the most frequently used positive or negative words, providing insights into the overall sentiment of the text. Content summarization becomes more accessible as key terms are visually emphasized, making it easier to identify central ideas within a large body of text.

Thus we created word clouds for our data as well (the tweets)





Sentiment Prediction:

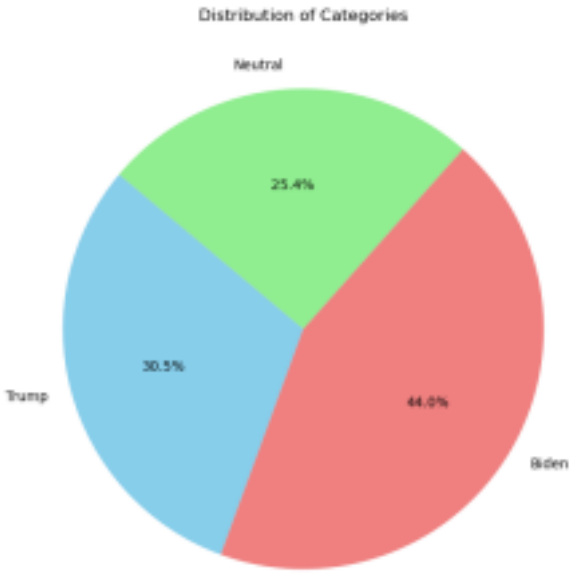
- Further cleaning tweet text allowed for a more refined sentiment analysis. Cleaned tweet text, removed stop words, and predicted sentiment labels. For cleaning the text, we:
    1. Removed signs like [@, #, _] and then converted *"snake_case"*, *"camelCase"*, and *"PascalCase"* to *"normal case"*

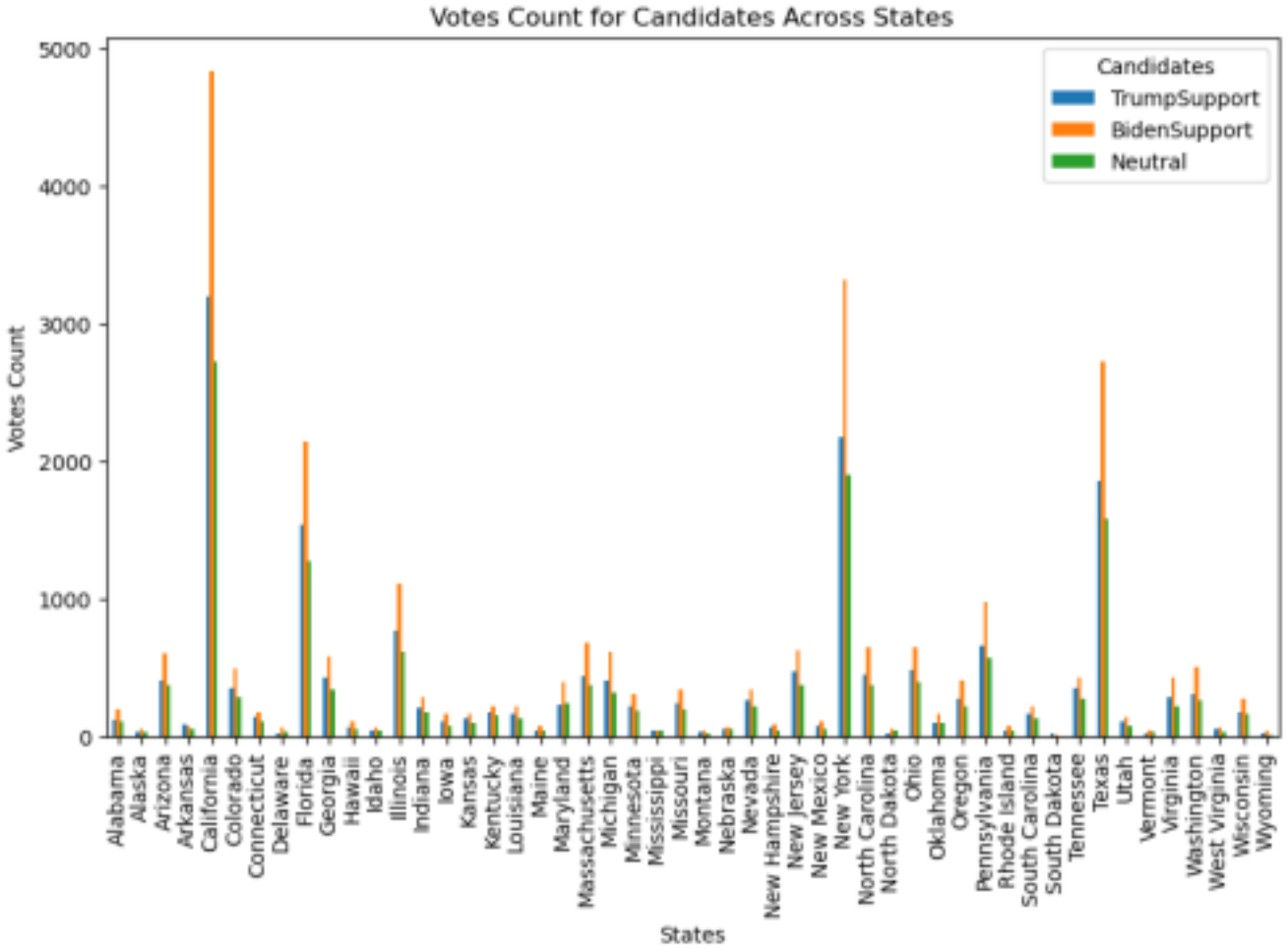
2. Expanded short forms such as *"can't"* , *"I'm"* to their corresponding full forms *"can not"* and *"I am"*
3. Tokenized the text and replaced the words preceded by *"not"* with their antonyms to make sure that *"not"* is not removed during text cleaning. 4. Removed the stop words utilizing the NLTK library.

- Utilized VADER for running sentiment analysis for each tweet and storing it in the corresponding column in the data frame.

Prediction of Election Results:

- First, we prepare our data frame to analyze probable vote share. This involves:
  1. Sorting the data frame by username of user (As username is unique for each user in Twitter)
  2. Creation of a new data frame to contain username (single occurrence, duplicates are merged), primary sentiment (taking average of sentiment for all the tweets done by each user), and a column indicating which candidate the user will vote for.
  3. The new data frame is filled by iterating through the sorted data frame, ensuring only unique entries, and filling in calculated average sentiment values (we add the sentiment values, for tweets with **"Trump"** as the subject, they are taken with positive sign, and for tweets with **"Biden"** as the subject, they are taken with negative sign. Finally, we divide this value by the number of tweets done by the user)
- Based on the sentiment value obtained for the user, we decide whether we can determine whom he will vote or not.
- For sentiment value in the range [-0.05, 0.05], we cannot decide whom the user will vote.
- For sentiment value greater than 0.05, we predict that the user will vote for **Trump.**
- For sentiment value less than 0.05, we predict that the user will vote for **Biden.**
- Thus, the column ***"Vote"*** is filled.
- The vote share for both the candidates is calculated. *The same is represented by a pie-chart. State wise estimation of Votes is also illustrated:*





Inference:

From the above model we can conclude that estimated share of voters are: 30.5% for Donald Trump and 44.0% for Joe Biden, while 25.4% of users are neutral. Based on the available data we can't predict whether neutral share of voters are inclined towards Donald Trump or Joe Biden.

Thus, we infer that Joe Biden is more likely to win the election.

Conclusion:

The model was able to provide us useful insights, which closely represent the actual scenario. The model predicted that ***"Joe Biden"*** is more likely to be the winner, with the vote share of **Biden** to be around 59% != !"#$%&'#$!#"&#)'*+# ,- .,'#/ -," 0%$#)

1223!"#$%&'#$!#"&#)'*+# ,- 4#5'"*6 .,'#/× 100'.

The actual winner was ***"Joe Biden"*** with vote share 51.3%.

Hence, the predictions made by the model were found to be sufficiently accurate.

Sources Of Error:

- Data obtained from twitter provides only an approximate representation of the actual public sentiment.
- The state-wise distribution of data was poor due to which, our model was unable to predict state-wise result accurately.
- Tweets are done very casually by users and hence it is very difficult to analyze them by the existing language models.
- VADER lacks the ability of contextual analysis, i.e., it does not consider the surrounding sentences or paragraph. As a result, it may misinterpret the sentiments of the sentence.

Extended Use:

We can also use Web Scraping to obtain data from social media sites. Access tokens are required for the same (which require a subscription).

You can also generate your own dataset using following code for future predictions

In [ ]:
```
import tweepy as tw
import pandas as pd
accesstoken = 'xyz'
accesstokensecret = 'abc'
apikey = '123'
apisecretkey = 'qwe'
auth = tw.OAuthHandler(apikey, apisecretkey)
auth.set_access_token(accesstoken, accesstokensecret)
api = tw.API(auth, wait_on_rate_limit=True)
#for Joe Biden
search_words = "Joe Biden"
#for Donald Trump
search_words = "Donald Trump"
date_since = #date from which you want to extract data
tweets = tw.Cursor(api.search_keys, q=search_words, lang="en", since=date_since).items(100)
tweets
tweet_details = [[tweet.geo, tweet.text, tweet.user.screen_name, tweet.user.location] for tweet in tweets]
tweet_df = pd.DataFrame(data=tweet_details, columns=["geo","text","user","location"])
tweet_df.head()
```

References:

- Vader Documentation
- Kaggle: Your Home for Data Science
- pandas documentation — pandas 2.1.3 documentation (pydata.org)
- Matplotlib documentation — Matplotlib 3.8.2 documentation
- NLTK :: Natural Language Toolkit
- NumPy Documentation
- WordCloud for Python documentation — wordcloud 1.8.1 documentation (amueller.github.io)
- 2020 United States presidential election - Wikipedia