Political Analytics on Election Candidates and their Parties in Context of the US Presidential Elections 2020

by

Rakshit Sorathiya

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Science (M.Sc.) in Computational Sciences

The Faculty of Graduate Studies

Laurentian University

Sudbury, Ontario, Canada

THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE
Laurentian Université/Université Laurentienne
Faculty of Graduate Studies/Faculté des études supérieures

Title of Thesis
Titre de la thèse — Political Analytics on Election Candidates and their Parties in Context of the US Presidential Elections 2020

Name of Candidate
Nom du candidat — Sorathiya, Rakshit

Degree
Diplôme — Master of Science

Department/Program
Département/Programme — Computational Sciences

Date of Defence
Date de la soutenance May 27, 2021

APPROVED/APPROUVÉ

Thesis Examiners/Examinateurs de thèse:

Dr. Kalpdrum Passi
(Supervisor/Directeur(trice) de thèse)

Dr. Ratvinder Grewal
(Committee member/Membre du comité)

Dr. Ramesh Subramanian
(Committee member/Membre du comité)

Dr. Sanjay Madria
(External Examiner/Examinateur externe)

Approved for the Faculty of Graduate Studies
Approuvé pour la Faculté des études supérieures
Tammy Eger, PhD
Vice-President Research (Office of Graduate Studies)
Vice-rectrice à la recherche (Bureau des études suypérieures)
Laurentian University / Université Laurentienne

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

I, **Rakshit Sorathiya**, hereby grant to Laurentian University and/or its agents the non-exclusive license to archive and make accessible my thesis, dissertation, or project report in whole or in part in all forms of media, now or for the duration of my copyright ownership. I retain all other ownership rights to the copyright of the thesis, dissertation or project report. I also reserve the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that this copy is being made available in this form by the authority of the copyright owner solely for the purpose of private study and research and may not be copied or reproduced except as permitted by the copyright laws without written authority from the copyright owner.

ABSTRACT

The availability of internet services in the United States and rest of the world in general in the modern past has contributed to more traction in the social network platforms like Facebook, Twitter, Instagram, YouTube, and much more. This has made it possible for individuals to freely speak and express their sentiments and emotions towards the society. Social media has also made it possible for bringing people closer by making the world a global village. There are influencers who promote products on social media platforms and politicians run their campaigns online for broader reach. Social media has become the fuel for globalization. In 2020, the United State Presidential Elections saw around 1.5 million tweets on Twitter specifically for the Democratic and Republican party, Joe Biden, and Donald Trump, respectively. The tweets involve people's sentiments and opinions towards the two political leaders (Joe Biden and Donald Trump) and their parties. The computational study of beliefs, sentiments, evaluations, perceptions, views, and feelings conveyed in text is known as sentiment analysis. The political parties have used this technique to run their campaigns and understand the opinions of the public. It has also enabled the modification of their campaigns accordingly. In this thesis, during the voting time for the United States Elections in 2020, we conducted text mining on approximately 1.5 million tweets received between 15th October and 8th November that address the two mainstream political parties in the United States. We aimed at how Twitter users perceived for both political parties and their candidates in the United States (Democratic Party and Republican Party) using VADER (Valence Aware Dictionary and sEntiment Reasoner) a sentiment analysis tool that is tailored to discover the social media emotions, with a lexicon and rule-based sentiment analysis. The results of the research were the Democratic Party's Joe Biden regardless of the sentiments and opinions in the in Twitter showing Donald Trump could win.

KEYWORDS:

Sentiment Analysis, Twitter, U.S. Elections 2020, Text Mining, Data Mining, Lexical Analysis, Positive Polarity, Negative Polarity, Tweets, Word Tokenization, Word Cloud, Donald Trump, Joe Biden, Electoral College, Voting/Elections.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my thesis supervisor Dr. Kalpdrum Passi for the constant support of my master's study and research, for his patience, motivation, and extensive knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a more desirable advisor and mentor for my master's study.

I would also like to thank my committee members for reviewing this thesis and serving on the defense committee.

Last but not the least, I would like to thank my loving and caring parents, brother, my dear partner and all my friends for presenting me with constant support and continuous encouragement throughout my years of study and researching and writing this thesis. This accomplishment would not have been possible without them.

Thank you.

DEDICATION

I would like to dedicate my thesis to my beloved grandparents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

NLP	Natural Language Processing
NLTK	Natural Language Toolkit
IDE	Integrated Development Environment
CPU	Central Processing Unit
API	Application Programming Interface

CHAPTER 1

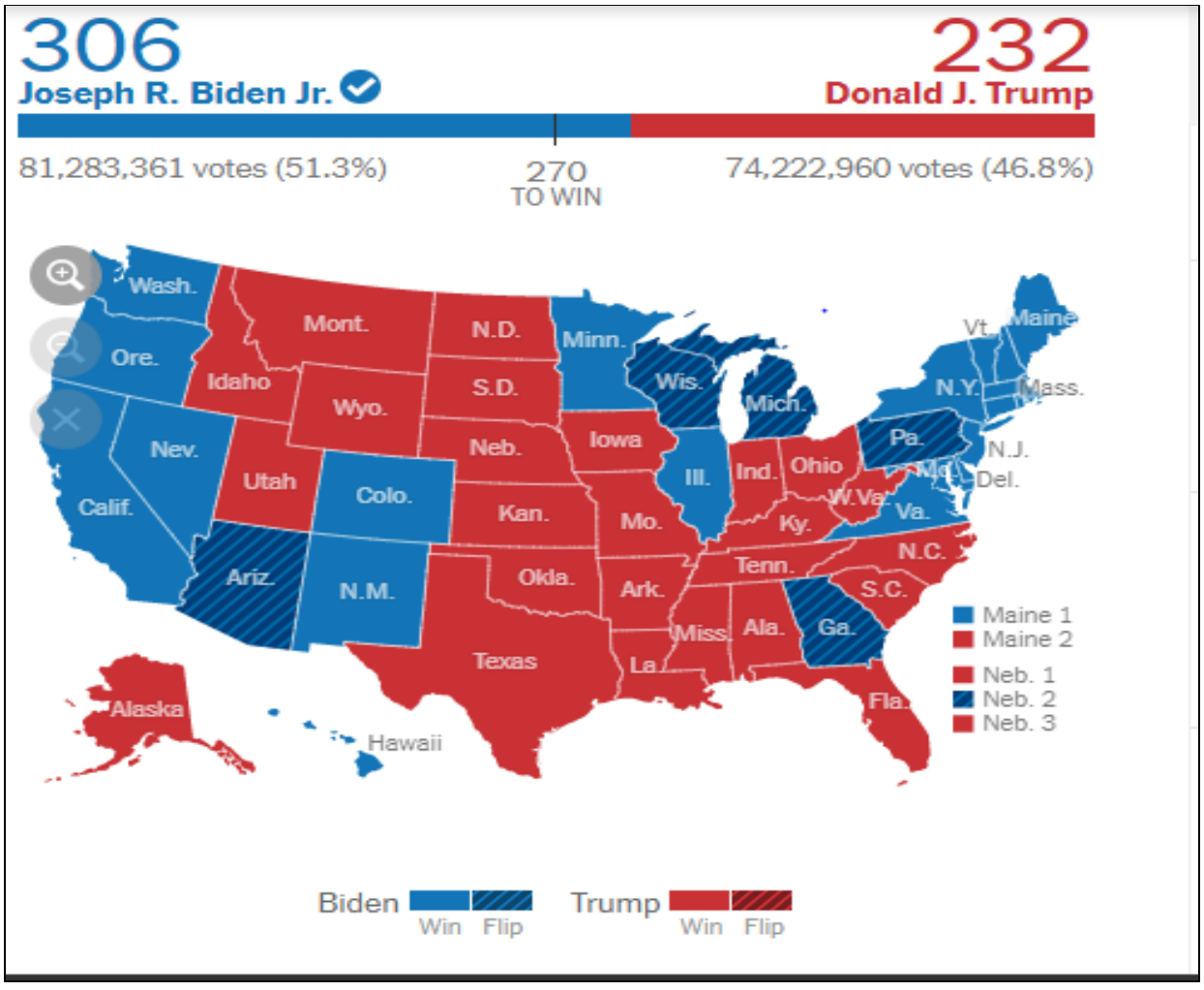
INTRODUCTION

The United States of America held their 59th quadrennial elections on 3 November 2020, with Joe Biden flagging the Democratic while Donald Trump took up the Republican flag. Although other parties like the Libertarian Party and The Green Party, the race was between the two political giant parties, namely, Republicans and the Democrats. The Democrat's move of choosing Kamala Harris as Biden's running mate was also influential as she was the first African American, Asian-American and first female to be selected as the vice-presidential candidate in US history. The US presidential elections are held after every four years, and these years must be divisible by four. The United States Electoral College is the body responsible for the polls, and the formula for choosing this body is determined by each state in the way they deem fit.

The number of qualified voters in the 2020 presidential elections was approximately 239.2 million voters [28]. Those who voted totaled at least 159.8 million Americans, an increase of 23.2 million voters from the previous elections in 2016, which had 136.6 million votes. Since 1900, the voting rate for registered voters in the 2020 elections was the highest in the US Presidential elections history, with the highest number of ballots counted in the election. The number 159.8 million voters constituted 66.8% voter turnout, which accounts for more than half of America's eligible citizens. The pandemic resulted in a doubling of the ballots cast in person or by mail before 2016.

The eligible citizens for voting opt for electors who would ultimately vote for the president during the election process. The number of electors is proportional to California's population

with the most, 55 electors and Washington DC, Alaska and Dakota having only three electors. The US has 538 electors, and for one to become president, he/she must accumulate half of this, which is 270 votes. When an elector wins the ordinary votes, he/she is then awarded all the state's electoral votes. All the electors finally make up the Electoral college responsible for electing the president [7]. The victorious party, led by Joe Biden, won 306 electoral votes and 81.2 million total votes, accounting for 51.3 percent of all votes cast in the United States.



Source:
https://www.nytimes.com/interactive/2020/12/14/us/elections/electoral-college-results.html

Figure 1.1 The Electoral Votes per state [31]

1.1. Influence of Social Media on Elections

Since emergence of Social media in 1996, it has been able to reach over half of the world's 7.7 billion population. The last decade saw the number of social network platforms triple from 970 million to 3.81 billion users between 2010 and 2020 [11][12]. According to Kepios's analysis, the United States had 6.9 million new users joining social media between 2019 and 2020, an equivalent of 6th place worldwide [15]. The United States has 70% of its population using social media, with another 83% of this population being 13 years and above. The number accounts for 231.47 million active social media users as of 2020, with the most popular platforms being 73% and 69% for YouTube and Facebook, respectively. In terms of gender, men are more inclined to YouTube and Twitter, while women are particularly interested in Snapchat and Pinterest. Generally, women are more prevalent in social media use than men; 76% of females and 72% of males use social media platforms altogether. In terms of race, the whites are more prevalent users with 73%, followed by Hispanic at 70% and 69% for the Blacks. Millennials and Gen Z are the most users, 82% and 90%, respectively [15].

The figures state that the effect of social media on elections cannot be understated. For instance, Donald Trump's win in 2016 was majorly attributed to his use of Twitter. Although there was a steep margin between Trump and Biden in terms of followership on Twitter, significantly, the two candidates had increased engagements all through the election period. Even though Donald Trump has 87 million followers, while Joe Biden has 11 million followers, reports showed that the tweets with the highest performance from Biden had doubled in terms of interactions compared to Trump's equivalent tweets. This is regardless of Trump's higher follower account, which shows the follower account's irrelevance [38].

After the debate held between Trump and Biden, social media users discussed the two candidates and reports show that Biden and Trump had 6.6 million mentions with Biden taking up 72% of those mentions [38]. Biden had a total of 511,000 mentions on the actual day of debate while Trump had a total of 244,000 mentions. The sentimental breakdown had Biden taking up 14% positive sentiments and 48% negative ones while Trump had 10% and 49% positive and negative, respectively [38].

More first-time young voters have been using social media for a significant part of their lives. The students may have first known about elections from their professors, and had their opinions influenced by social media, which accounts for 20% of young adults. Instagram and TikTok went ahead on influencing their opinions of the candidates they wanted. This is because most young adults who were eligible to vote are loyal influencers. The influencers affect not only their views concerning a brand or a product but also their political views. With the number of first-time voters having doubled in 2020 compared to the same in 2016, their vote is most likely to be influenced by social media content.

The United States campaigns had to rely so much on social media, primarily due to the Covid-19 pandemic, which limited their ability to hold physical campaigns. That is to add, social media was essential in reaching out to the crucial undecided voters. According to the polls after the debate, the percentage of undecided voters was 38% and 41% for Biden and Trump respectively [38].

Generally, according to experts social networking platforms had a major effect on new voter registration, community awareness, and early voting in the United States. As of October 29, 2020, several states announced massive new voting registrations, with a total of 99.7 million early votes cast. This showed a two-thirds improvement on the overall number of ballots cast in

2016. Campaigns such as "Get Out the Vote" (GOTV) on social platform are some of the actions that campaigners used to get the voters to turn up in large numbers to vote [29].

Although there was a positive turn up for the citizens to vote due to social media, there was a downside. There was a rise in misinformation and disinformation on social media platforms. On the days nearing the elections, there was adverse manipulation of video clips and content fabricated to make candidates seem less competent posting these clips in super active accounts to appear to be popular to affect the polls [29]. According to several investigations, Russian operatives used social media to intimidate votes in favor of Trump ahead of the 2016 presidential election. [2].

1.2. The Motivations

Trump runs his own Twitter account and has been an active user since the presidential elections of 2016 as well as 2020. He used Twitter to share his sentiments and give details about his campaigns to his supporters [20]. Biden also made extensive use of Twitter to broadcast his campaign news and opinion on policies' changes, although someone else managed his account. He also used social media to announce Kamala Harris as his running mate and talk to her in public. A survey showed that the most active users constitute 10%, generating 92% of popular tweets. The democrats are the highly engaged users comprising 69%, while the Republicans are 26%. The democrats post twice as many posts as the republicans per month [34].

Facebook was used to get donations from the public and people to sign up for candidate email lists. Facebook was considered the best platform for ads because of its ability to give direct feedback. Compared to its competitors, the demographic range for Facebook is broader, encompassing more significant proportions of the elderly [34].

Social media trends such as the Black Lives Matter Movement, the Trump administration and handling the Covid19 pandemic brought forth an urgency for the eligible voters to cast their votes for the presidential candidate they perceived to be favorable. As it had been in earlier elections where in-person voting was more popular, with the surge in the pandemic and people staying at home, mail-in voting took the lead [30]. Staying at home also contributed to people relying on social media as a source to keep up with elections proceedings.

1.3. The Objective

The aim of this research is to demonstrate the relationship between public opinions and viewpoints expressed on social media sites and their effect on the 2020 US presidential elections. The objective is mainly on how Twitter was relevant in influencing the users' political orientation towards certain political parties, with 1.5 million tweets collected from 15 October 2020 to 8 November 2020. The thesis mainly focuses on user tweets targeted to Joe Biden of the Democratic Party and Donald Trump of the Republican Party.

The sentiment analysis according to the Oxford dictionary, "The method of computationally characterizing and categorizing sentiments conveyed during a piece of text to see if the writer's attitude against a given subject, product, and much more is positive, negative, or neutral." The thesis will apply python programming and lexicon rule-based approaches to predict the tweets' sentiments using SentiWordNet and WordNet. The thesis will use Valence Aware Dictionary and sEntiment Reasoner (VADER) to analyze the social media texts.

In order to better understand the popularity of Joe Biden and Donald Trump, the thesis will conduct exploratory data analysis and use data visualization techniques to analyze tweet volumes for the candidates during the period stipulated.

1.4. Methodology

Figure 1.2 shows the flow for Twitter sentiment analysis, starting from the first extraction of tweets from the source to filtering out the data, preprocessing, sentiment analysis, and the results. The VADER will help analyze the sentiments and identify the tweets as either positive, negative, or neutral.

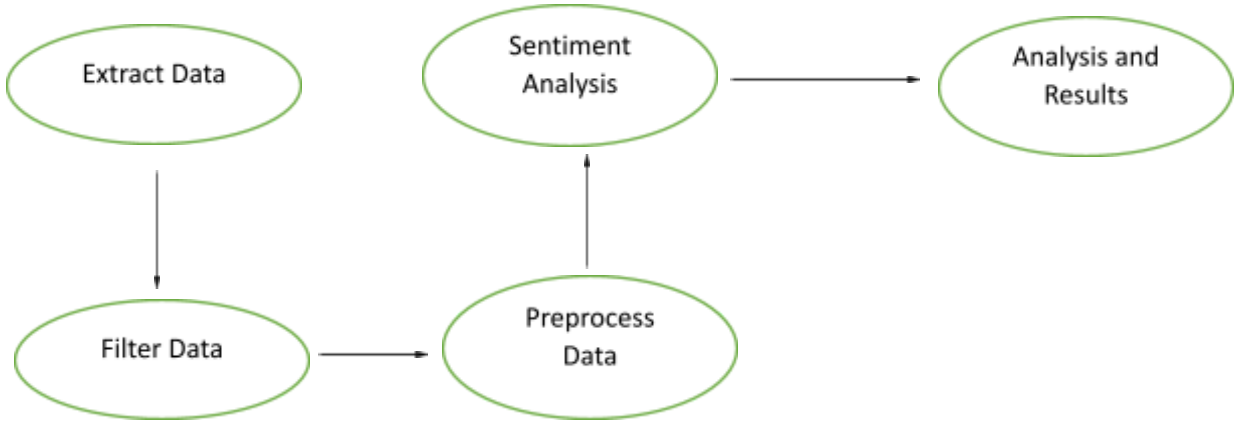


Figure 1.2. Flow diagram for Twitter sentiment analysis

CHAPTER 2

LITERATURE REVIEW

Sentiment analysis of tweet data is presumed to be a much more difficult problem than regular text analysis, i.e., review documents. The widespread use of odd and casual expressions contributes to this issue, the short length of tweets, and twitter's rapid linguistic evolution. In Twitter sentiment analysis, a significant amount of work is carried out, followed by the feature-based approaches.

2.1. Related Work

In this section, we are persisting in discussing related works about predicting and analyzing the result of an election using Twitter. We marked that researchers use a different way concerning this problem. A few researchers try to find the political or ideas preference of a user, then compare it to the election, and others use judged tweets related to the future election and view out vote preference of the user using that data.

Previous research has demonstrated that studying these sentiments and trends can provide useful findings that can be used to predict public opinion on elections and government policies [32] [36] [39]. In [32], the author examined sentiments (positive, negative, or neutral) as well as emotions (anger, sorrow, and a variety of other emotions) about the many leading party candidates, and endorsed that, and calculated a distance scale. The distance calculation indicates the proximity of political parties; the greater the distance, the more likely they are to have similar political relations. However, [36] and [39] show that Twitter knowledge will facilitate the prediction of election polls and derive valuable data regarding public opinions.

Four procedures were discussed by Das and Bandopadhyay [13] to conclude the sentiment of a word. The first method of assessing the sentiment was to suggest an immersive game that demonstrates the words and their polarity. The polarity was generated by a bilingual dictionary of English and Indian languages in the second approach. The polarities were implied using WordNet in the third approach. They use pre-annotated corpora to assess polarity in the fourth approach. Das and Bandopadhyay [13] recognized concerned style in the Bengali corpus. To execute sentence-level annotation, they divided the words into six feeling classes including three intensities classes.

Analysis has acquired a rule-based model for sentiment analysis called VADER. They found that with a sentiment lexicon and several syntax rules, their model could exceed both individual human raters and machine raters and machine learning methods [21]. The VADER model resulted in an open-source application [22]. A transposed version of the VADER application and lexicon will be used in this report and will be further introduced Chapter 5.

The ideas that people express in social media are also related to having an impact on people’s choices of political parties [14]. Some investigations have proved to predict political election results with sentiment analysis of tweets: using the lexicon method for the Swedish elections [39]. In the study concerning the Brazilian elections, Oliveira et al. [14] analyzed whether sentiment analysis of Twitter data could be used to assess voters' political choices in the same way as public opinion surveys do. Their results were positive.

The dictionary-based approach's main technique is to start with a small selection of manually labelled opinion terms and then expand it by browsing through a wider range of texts like WordNet [10] [11]. The new words are then combined with the first collection of opinion words, and the process is repeated until no further words can be identified. The most important

drawback of this approach is that it is primarily based on corpora, and we do not necessarily have access to many opinions in terms for domain. It is vital to know that not every word in a lexicon represents a positive or negative opinion about an entity. [26]. The Corpus-based technique is mostly used in two scenarios: discovering new sentiment words from a domain corpus using a given list of existing opinion words and constructing a sentiment lexicon from another corpus [26]. This strategy is less feasible than the dictionary-based method since it would necessitate a corpus of all English words [10]. Depending on the methodology used, the corpus-based method is distributed along the statistical and semantic paths.

To discover new potential tweets, we first used a lexicon-based approach, followed by a chi-square analysis. A lexicon-based approach is mixed with a rule-based approach in our research. There are several modulation schemes, but they all rely on manually defined collections, or labelled data, for supervised learning. The method described by Barbosa et al. [12] after eliminating the original tweets and classifying the first category as positive or negative, the algorithm classifies in subjective and objective terms.

Jose and Chooralil [35] a new sentiment analysis method in 2015. They analyzed tweets using a new approach. They looked at how to retrieve information and facts from tweets using lexical tools like WordNet, SentiWordNet, and word sense disambiguation. They also suggested to use a negation handling approach in the pre-processing data stage to achieve maximum performance. They explored a range of tools because of the author's creative vision. Twitter's Streaming API was used to implement the process of data collection.

Tumasjan et al. [39] examined that in the 2009 German national parliament election, Twitter and tweet opinion had a significant impact. In the weeks prior to the election, activists sent 104,003 tweets. Linguistic Inquiry and Word Count, a program developed in 2007, was used to derive

sentiment from tweets. Political discourse is normally dominated by a narrow pool of heavy users (80+ tweets), according to the writers. This research has discovered that the number of tweets represents the election results.

However, we discovered in our dataset that the tweets we concern about each candidate and their political parties at given time was challenging for analyzing their political sentiment. Because a full-text sentiment analysis solely takes into consideration what was the user's sentiment.

In this thesis, we have used sentiment analysis techniques to uncover the opinions, sentiments, and emotions of user tweets for respective parties and candidates. We propose a rule-based lexicon approach that uses a dictionary of WordNet to calculate the polarity of the tweets (positive, negative, and neutral). However, just using simple WordNet and Corpus-based approach may not give accurate results when working with social media text. We have used VADER [22] (Valence Aware Dictionary and sEntiment Reasoner), which is particularly sensitive to feelings shared on social media platforms. VADER [22] not only determines the Positivity and Negativity ratings, but also the degree to which a sentiment is positive or negative.

Taboada et al. [26] introduced a lexicon-based method for sentimental analysis. To complete the research job, the author used dictionaries of positive and negative polarized terms. By integrating intensifiers and negation terms, a semantic orientation calculator was created based on these dictionaries. The authors [21] compared the techniques for analyzing a piece of natural language text based on the sentiment conveyed in it, i.e., whether the overall feeling is negative or positive. They utilized machine learning (Support-Vector Machine model) together with domain-specific lexicons to put into effect a set of methods for categorizing features and determining polarity in product reviews.

The use of Twitter by politicians and their campaigns is a significant topic to study. When Barrack Obama was campaigning in 2008, he used Twitter, which finally defined the role of Twitter in the political battles [39] [40]. A similar investigation was carried out into the use of Twitter by members of the US Congress during their election campaigns. According to findings, members of Congress regularly posted information on Twitter about their political views on a variety of topics and problems concerning their constituents. [41] [42].

In the early 2000s, a sentiment analysis study was initiated. Several tools for analyzing views and feelings from public opinion streams (blogs, forums, and industry websites) have been developed since then. On social platform like Twitter, where people express their opinions on a variety of topics, a growing issue has emerged. [43] [44]. Most of these efforts are focused on one of two approaches: Lexical-based or Machine Learning approach.

CHAPTER 3

DATA EXTRACTION

This chapter discusses the different ways of extracting raw information about people's sentiments in the many different social media channels such as Facebook, Pinterest, and Twitter. The emphasis will majorly be on Twitter, the pros and cons of its features, the use of Twitter API. This chapter gives a detailed description of how Twitter has used a social media platform, its characteristics, and the tweets collected. The chapter will also describe the features of the extracted tweets and filtration of extracted tweets.

3.1. Other Social Media Platforms

Facebook is a website created to allow users to connect with other users, commonly referred to as friends online. The website enables people to share their posts in the form of videos, music and written articles. The platform has allowed people to share their thoughts more intensely and freely with anyone they may like [41].

Facebook was created in 2004 by Mark Zuckerberg, who was at Harvard University and had from then grown into a worldwide brand. With over 2.3 billion users, the website has increased by ensuring that it offers privacy to its users [41]. For instance, a user is required to send another user a "friend request" to see that user's posts. Once a user accepts the other's friend request, they are branded as friends, and therefore they can view each other's timelines and the timelines of their friends. The privileges of being friends are that one can view the posts of the other freely.

Friends on Facebook can communicate freely in public or send private messages. However, the reactions to one's posts can be known since there are options of "like" with a thumb up emoji,

"unlike" with a thumb down emoji, "love" with a heart emoji and the like, which "friends" click to air out their sentiments. There is also a comment section in every post that a user posts, allowing users to air out their views in words. Besides, users can "like" pages of people, businesses, football clubs, which allows the user to post comments and receive updates regarding that individual liked page with ease. Facebook gained popularity because of giving the users the facility to be in contact and be social. It is easier to reach someone from miles away just by one post [41].

Facebook's popularity influenced because of its vast base which is both a good and a bad thing. A good thing is the candidates did have a pool of people in one place and thus made communication easier. However, this very thing can be a challenge because one post can mislead the people regardless of any news or events.

Between March 01, 2020 and the election day, Facebook could register 4.5 million people to vote and sign up 100,000 people as poll workers. Facebook ads also helped sensitize people on their civic duties regarding elections, leading to 140 million people visiting the Voting Information center with 33 million people visiting on the election day only [16].

During the period in question alone, Trump's Facebook page had skyrocketed to 130 million comments, reactions, and shares, while those of Biden's official page was only 18 million. The previous month, Trump was still leading with 86 million engagements, while Biden had 10 million [36].

Instagram is a social networking service created in 2010 by Mike Krieger, Kevin Systrom and Facebook in 2012. The application provides the functionality of sharing/posting photos and

videos publicly or to the approved followers only. With over a billion active users and 140 million users from the US alone, Instagram played a huge role in the US election campaigns [9].

Influencers and celebrities are the ones who have dominated in Instagram and are usually paid to promote products of brands. Politicians turned to Instagram to capture the millennials' attention, who are the platform's majority users. Donald Trump had 60 million comments and likes on his posts while Biden had 34 million engagements only. In the previous month, Trump had 39 million engagements, while Biden had 13 million [36].

Trump had the most outstanding engagements on Instagram and Facebook in early October when he was hospitalized for having COVID outbreak. Well-wishers were so supportive on his pages sending likes, comments and wishes to him for quick recovery.

YouTube is a service that allows video sharing/uploading and allows users to comment, like, watch, share and download offline any uploaded videos. Users can get notifications when one uploads a video by subscribing to that user’s channel.

According to SocialBlade, Trump also took lead on YouTube with his video for the last thirty days getting 207 million views while Biden had only 29 million. Trump used ads about accusations against Hunter, Joe Biden's son to increase engagements on his YouTube channel. Those ads got approximately 22 million views on a single day [36].

Facebook and Instagram could be categorized as platforms that people use to connect to their friends and families. In Facebook, people share videos, and generally their lives updates with their friends and families. In Instagram, people post their highlight reel in the form of photos and videos in addition to following influencers [17].

However, Twitter is a platform people use to share their views and opinions on trending news. In as much as Twitter is also used to connect with friends, it is majorly custom made to connect to a larger audience, the world, where people are given 280 characters or less to share with the world what is happening and how they feel about it [17].

One peculiar feature that makes Twitter stand out among other social media platforms is that Twitter offers an API (Application Programming Interfaces). The API enables the developers to integrate Twitter with other applications. The developers use the APIs to extract public tweets and replies by searching specific keywords or requesting a sample from the specific Twitter accounts. Crawler alongside other archiving websites that would be used to extract tweets for all languages. However, the other social media platforms do not have APIs that could be used for extraction, making Twitter the only platform used to extract the required data.

3.2. About Twitter

Twitter was created in March 2006 and later launched in July the same year by Jack Dorsey, Evan Williams, and Noah Glass. It is a social platform that helps people to connect with posts known as "tweets". Unlike the unregistered ones, registered Twitter users can interact with the tweets by retweeting, posting, and liking. Initially, the tweets were restricted to 140 characters but were later doubled in 2017 for non-CJK languages (a collective term used for Korean, Japanese, and Chinese). However, for the videos and audio tweets, the time limit is 140 seconds. Twitter can be accessed from its website, or one can download the application using their mobile device.

Twitter has its headquarters in San Francisco, California, with over 25 outlets worldwide with approximately 5,000 employees. Twitter has over 300 million users, with approximately 275

million active monthly users. The United States is ranked first with active users worldwide, with 67.8 million active users as of 2020 [11].

Twitter users must follow other Twitter followers for them to access their content in their timelines. The users either follow individuals or organizations that have similar personal interests. The Former President, Barrack Obama, is among the people with the highest number of followers on Twitter.

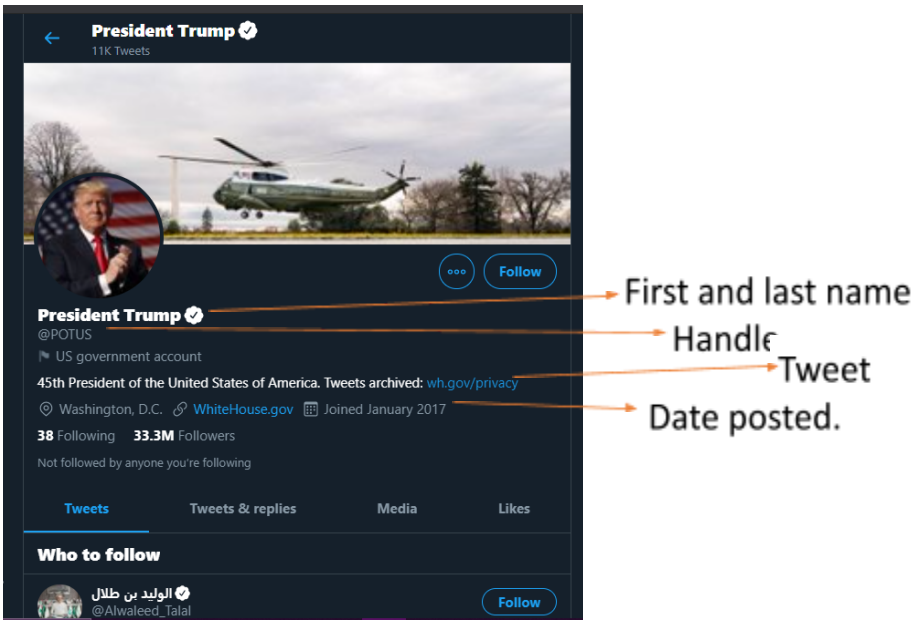
The social media ads, that both candidates used were targeted to all ages. Biden's campaigns were targeted mostly on women than men; for every $2 spent on men, they spent $3 on women since most women were likely to identify themselves as Democrats than Republican. Trump spent primarily on Facebook ads for 55 years and older while Biden spent mostly on people aged 35-44. The campaigns spent less on people between 18-24 simply because this category has a high likelihood of showing up to vote regardless [37].

3.3. Characteristics of Twitter Data

For people who love social networking, Twitter comes in handy because it is free, fun, and fast. Its popularity comes from the fact that it is immediate; in case of a question or a concern, all one needs to do is attach a hashtag (#) to a tweet and join other people conversing about the same topic. For instance, in the case of presidential campaigns, all that one would require is to write #elections 2020, #trump or #Biden. You can now join the conversation and see the tweets retweets, posts, videos, and news concerning that hashtag.

Twitter is not limited by geography, conversation, or time. The only restriction is the 280 characters rule that restricts the users on the character count. However, this is not a bad thing because it keeps the user focused on summarizing what they want to mean in a nutshell.

The symbol "@" is also a significant symbol when it comes to Twitter. A user will use the symbol when they want to mention another user or any specific account on Twitter. For instance, when one posts "@Biden", they mention him or call him out, and thus the account, or user being mentioned will get a notification that someone mentioned them in a comment.



Source : https://twitter.com/realDonaldTrump

Figure 3.1 Public View of Candidate's Profile

Besides, a tweet contains metrics that have been summarized to show how a particular tweet faired; the likes, dislikes, retweets, date, and time the comments were posted and the user who tweeted. The thesis has focused on extracting the text part, which contains the users' opinions and sentiments.

3.4. Extracting Twitter Data

We used Tweepy, a library from Python, to extract the data from Twitter. The thesis applied Tweepy instead of its closest rival Twint because Tweepy is the official Python library authorized

to access Twitter API. In contrast, Twint has not been officially authorized to access the API. Thus, it uses a scrapping tool customized to grab tweets and eliminate the API's restrictions [22].

Tweepy's documentation is excellent and has tutorials for authentication and streaming, and API references for every method used. However, Twint documentation has no proper detailed specifications, and thus the package seems to have outrun the documentation. The limitation encountered by using Tweepy - is that brought about by the API. Twitter API has limitations for extraction of only the latest 3,200 tweets in a timeline [22].

The data downloaded is 1.5 million tweets that were collected for three weeks. This research emphasized on tweets that mentioned the two running candidates from the Democratic and Republican parties. The data was extracted from a JSON object (data interchange file format), a semi-structured data file format.

The information obtained from the JSON object includes hashtags, URL profile image, counts of followers, friends and statuses, tweet text and locale among other useful variables with regards to a tweet and the user profile. The table 3.1 gives a detail of the keywords extracted from each JSON object.

Table 3.1 Keywords extracted from JSON object.

Columns	**Data Types**
Favorite Count	Integer
Retweet Count	Integer

Quote Count	Integer
User Location	String
Hashtags	String
User Screen Name	String
Created At	Timestamp
Tweet Locale	String
Tweet Text	String
Tweet ID	Integer
User Mention Screen Name	String

3.5. Dataset and variables

In the thesis, we used Python Twitter API Tweepy to extract the data for the period between 15 October and 8 November; the data collected was 1.5 Million tweets. We used specific keywords to filter in user tweets relevant to the US presidential elections in 2020. A comprehensive list of keywords was set in place to extract people's sentiments for each leader and party. The keywords listed helped in extracting both positive, negative, and neutral sentiment for the candidates in question. The structure for the presidential elections was as follows mentioned in Table 3.2.

Table 3.2 Structure of Presidential Elections

Nominee	Donald Trump	Joe Biden
Party	Republican	Democratic
Home State	Florida	Delaware
Running Mate	Mike Pence	Kamala Harris

The information below shows a list of keywords used for extracting user tweets related to each party and leader. We used the #(explore) to extract the unique hashtags for the candidates.

For Donald Trump and Mike Pence (Republican Party)

(#Republican, #DonaldTrump, #voting, #Trump, #HarrisCounty, #MAGA2020, #TrumpIsANationalDisgrace, #TrumpVirusDeathToll, #TrumpCovid, #EndTrumpChaos, #TrumpTaxReturns, #DumpTrump2020, #TrumpLies)

For Joe Biden and Kamala Harris (Democratic Party)

(#BidenLies, #BidenLies, #VoteBlueToSaveAmerica, #VoteBlue, #VoteBidenHarris2020, #BidenHarrisToSaveAmerica)

For anything else related to elections, we used the following keywords:

(#ExGOP, #GOPSuperSpreaders, #AmericasGreatestMistake, #TrumpVsBiden, #PresidentialDebate, #PresidentialDebate2020, #Election2020, #TrumpBidenDebate, #Propaganda, #USPresidentialDebate2020, #USElection2020, #BountyGate, #BLM,

#BlackLivesMatter, #Elections, #VoteLibertarian, #Opinion, #CountryOverParty, #nhPolitics, #VoteForAmerica, #PlatinumPlan, #PresidentialDebates2020, #Debate, #SuperSpreader)

CHAPTER 4

DATA PRE-PROCESSING

4.1. Introduction

Data pre-processing is a significant aspect in any text mining analysis—the action precedes before carrying out Natural Language Processing (NLP). The absence of data pre-processing will have difficulties calculating any text data's polarity, resulting in incorrect results. In Chapter 3 (Data Extraction), we collected linguistic data from Twitter, meaning the data can have multiple dialects in a single tweet. It is also possible for a single tweet to contain large amounts of noise such as slang, repeated words, URLs, images, and unnecessary whitespaces. These kinds of noises make the performance of sentiment analysis seem almost impossible. Therefore, it is mandatory to remove the noise and irregular data, so that the actual meaning of a tweet is correctly decoded and understood.

Data pre-processing helps reduce the dataset's overall size since the noisy data is removed, resulting in a reduced overall processing and execution time. It also improvises the accuracy of calculating the sentiment of the tweets. The algorithm implemented in the python programming language is discussed in the following section of this chapter.

4.2. About Python

Python is a programming language that Van Rossum created in the year 1991. It is an interpreted, high-level, general-purpose programming language that focuses more on code readability [40]. It supports several programming paradigms, including functional programming, object-oriented programming (OOP), and procedural programming. Python is an open-source language that has

an extensive and thorough library that users can explore. One of the significant features is the dynamic late resolution (late binding) that allows variables and method names to bind during a program's execution [40].

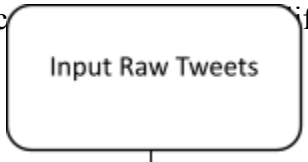
Python has the open-source libraries and has a strong community. The language has the most advanced libraries, such as Pandas. Pandas is an open-source data manipulation and analysis library that is fast, efficient, scalable, and simple to use. Pandas help to summarize data using aggregations and groups easily. It comes with inbuilt plotting functions, which allow the user to visualize the data while analyzing. Besides Pandas, Python also has other libraries used in exploratory data analysis and modellings, such as NumPy, open NLP, SkLearn, SciPy, and VADER.

This thesis used Python 3.7.6 version for data extraction, data pre-processing, sentiment analysis, and data visualization. The thesis used NumPy, nltk, Plotly, word cloud, Vader, and pandas as the open-source libraries. The thesis also used 'Pycharm' IDE 2020 as an interface for the implementation, code compilation, debugging, and error handling. Python has many other features that make writing code easy.

4.3. Data preprocessing methodology

The data set available on the 2020 Presidential Elections for the United States contain a text field that maps the user tweet. The tweets are noisy and have many irregular texts, memorable characters, and could also contain whitespaces and unnecessary punctuations.

We calculate the POSITIVE, NEGATIVE, and NEUTRAL polarity of the user tweets collected for the two candidates. The pre-requisite step is to clean each user tweet. Pre-processing data normalises linguistic data, reduc ifies the use of vocabularies when



interpreting emotions from tweets [46]. Figure 4.1 shows the high-level view for data pre-processing.

Figure 4.1 Overview of Data Preprocessing

Table 4.1 describes the algorithm implemented in Python for cleaning and removing noise from each user tweet.

Table 4.1 Data Preprocessing algorithm

Input- User Tweets

Output- Processed and cleaned user tweets.

For each user tweet in the dataset:

1. Remove all URLs or https:// links using regular expression methods.
2. Substitute the word 'username' for all '@username.'
3. Exclude all #Hashtags and Retweets from the posts
4. Look for two or more characters that are repeated and replace them with the character itself.
5. Exclude all additional special characters from the tweets (: | []; : - + ( ) >?! @ # percent *,). Also, remove all white spaces.
6. Replace wrong contractions with custom python dictionary containing mapping of correct English contraction.
7. Replace keywords related to political parties.

Return processed tweet

The algorithm describes what happens to each user tweet. It first removes any URL present, then removes all symbols and RT plus any special characters. Besides, the algorithm checks for the repetition of characters. If any, it removes the repeated characters. Lastly, it removes the stop words, extra white spaces, punctuations and replaces false contractions with the correct ones. The next section discusses pre-processing in detail and with examples.

4.4. Pre-processing Steps

In this section, pre-processing steps are broken down into sub-steps and explained in detail using a user tweet example.

4.4.1. Removing Links/URL

Removing any link or URL is the very first step of cleaning a tweet. The user tweets can include a URL that does not add up to the tweet's overall context. In our approach, web content is not considered for the calculation of the polarity of a tweet. Therefore, removing any links or URLs will not affect the overall meaning of a tweet. However, it will reduce the overall size of the dataset. Python "re" package is loaded, used to clean the textual data using a regular expression. The example shown in Table 4.2 contains a link starting with "https://." After this step, the processed tweet will not include the URL.

Table 4.2 Example of removing URL from user Tweet.

User Tweet	RT @DonaldTrump: "Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…https://trumptweetwWi
Step 1: Processed Tweet	RT @DonaldTrump: "Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…

4.4.2. Removing Hashtags and Username Symbols

The second step of cleaning the tweet deals with removing @ and # symbols from the user tweets. @ in a user tweet is used in mentioning a user account, and # is used by users posting the tweet for supporting the trend in Twitter. Twitter internally uses these two symbols, @ for identifying which user account is mentioned in the tweet and # for finding what is trending in Twitter.

The symbols @ and # have not added up to a user tweet's overall context when calculating its sentiment. Thus removing @ and # can help reduce the complexity in finding the overall polarity of a tweet. The regular expression of Python is used in eliminating such characters and symbols. Table 4.3 shows an example of how the tweet transforms after the second step.

Table 4.3: Example of removing @ and # from user Tweet.

User Tweet	RT @DonaldTrump: "Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…https://trumptweetwWi
Step 1: Processed Tweet	RT @DonaldTrump: "Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…
Step 2: Processed Tweet	RT DonaldTrump: "Tonight, FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…

4.4.3. Removing Retweet (RT) Character

The third step of data pre-processing entails removing the character "RT," which means a user tweet is retweeted in the Twitter dictionary. Twitter users retweet an original tweet to increase their reach across the community or add their opinion on the original tweet. To identify whether a tweet is retweeted, Twitter user the unique character "RT" followed by the original tweet. What matters in our analysis is the text of the user tweet, not the special characters. Thus, it is in our best interest to remove the special characters. The removal further reduces the complexity in calculating the polarity/sentiment of the user tweet. Table 4.4 shows the processed tweet after step 3.

Table 4.4 Example of removing RT character from user tweet.

User Tweet	RT @DonaldTrump: "Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…https://trumptweetwWi
Step 1: Processed Tweet	RT @DonaldTrump: "Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…
Step 2: Processed Tweet	RT DonaldTrump: "Tonight, FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…

Step 3: Processed Tweet	DonaldTrump: "Tonight, FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…

4.4.4. Removing extra white spaces and additional special characters

The fourth stage in pre-processing is to remove all unnecessary white spaces and special characters. Most user tweets contain different white spaces and other special characters like brackets, +, - signs, percentage signs, asterisks, and more. These characters neither have specific meaning nor add positivity or negativity to the user tweet's overall context. Most special characters are added as a suffix to a particular emotion. For instance, "happy!" the user adds the exclamation mark to emphasize the word "happy." Still, in a lexicon-based approach wherein each word in the user tweet is compared with the corpus, there is a fair likelihood that the word "happy!" (with (!)) would diminish the polarity of the optimistic answer, converting it into a neutral expression, resulting in an incorrect outcome. Table 4.5 shows the processed tweet after step 4.

Table 4.5 Example of removing extra white spaces and special characters.

User Tweet	RT @DonaldTrump: "Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…https://trumptweetwWi

Step 1: Processed Tweet	RT @DonaldTrump: "Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…
Step 2: Processed Tweet	RT DonaldTrump: "Tonight, FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…
Step 3: Processed Tweet	DonaldTrump: "Tonight, FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"…
Step 4: Processed Tweet	DonaldTrump Tonight FLOTUS and I tested positive for covid19 We will begin our quarantine and recovery process immediately We will get through this TOGETHER

4.4.5. Removing repeated Characters in a word

The fifth step is about removing repeated characters in a word. Users on Twitter are not always formal in their tweets. They repeat characters while expressing their emotions. For instance, one may tweet, "Thankssssss" the extra "S" is unnecessary as they do not belong to any lexical corpuses and therefore could mislead the user tweet's overall context. Thus, it is essential to remove the repeated characters from a word wisely. It is also vital to make sure that the characters removed do not alter a word's meaning. For instance, a tweet such as "Goooooood" would be incorrect if it were changed to "God" instead of "Good.'

Therefore, before removing repeated characters from the word, one must implement a recursive WordNet lookup. WordNet is a lexical database or corpus that provides similar words of a particular word called Synsets. Synsets are groups of words and their synonyms that express similar polarity. Therefore, a WordNet lookup would ensure the removal of the correct number of repeated characters.

4.4.6. Replacing words with Contraction

This is the sixth step in pre-processing. Contractions such as "I'll," "won't," "shouldn't" are common in tweets because they must adhere to the 280-word-count limit. Usually, the lexical corpus does not include the contractions. Instead, it contains several bigrams like "I will," "do not," "should not," which often determine the polarity of the tweets.

It is, therefore, essential to replace a contraction with its equivalent bigrams. This is achieved by identifying the pattern to be replaced, and then each instance of the specified pattern is replaced by a corresponding substitute string. Python uses regular expressions to identify the contractions and later develops a dictionary where the key is the identified contraction in the user tweets, and the value is the replacement. Table 4.6 shows the example of a user tweet processed by replacing the contraction "shouldn't" with "should not."

Table 4.6: Example of replacing contractions from a user tweet.

User Tweet	@JoeBiden we **shouldn't** repeat the same mistake of the previous elections
Processed Tweet	JoeBiden, we **should not** repeat the same mistake of the previous elections

4.4.7. Replacing keywords of Political parties and their leaders

In the United States, there are indeed two parties in conflict: The Republican Party and the Democratic Party. For the candidates running for the nomination of President, the two parties are seen interchangeably. Donald Trump can be interchange with The Republican Party, whereas Joe Biden can be interchanged with The Democratic Party.

This chapter exclusively focused on cleaning the raw tweets extracted. The algorithms implemented in each step in earlier sections helped filter the tweets and remove unwanted noise from the data. The approach also significantly reduced the dataset's size, which will help calculate the user tweet's polarity efficiently, and execution time will be on the lower side.

CHAPTER 5

SENTIMENT ANALYSIS

5.1. Introduction

Human beings are social beings, no matter how far they are from each other with different cultures and upbringings. The rise of the internet and globalization has turned the world into a small global village. The introduction of social platforms such as Facebook and Twitter have made interaction among people from different places possible and thus resulted in the base of these platforms' immense growth. Twitter has a total of 300 million active monthly users. The media have become popular among people because they have facilitated accessible communication and free expression of opinions and emotions. Twitter also allows the use of common languages such as English, Hindi, French, Spanish, Japanese, etc. Communication employing language is referred to as Linguistic Communication.

Communication is unique as each form has its distinct structure, grammar, part of speech etc. Natural language processing (NLP) is a technique for analyzing and manipulating linguistic content. By analyzing sentence structure and measuring sentiment polarity using lexical tools or corpuses like WordNet, SentiWordNet, and more, NLP can process and interpret massive quantities of unstructured data.

The technique involves word tokenization, stemming, lemmatization, and stop word analysis to extract the sentence's core structure. When breaking down the sentence, different parts of expression such as nouns, adverbs, prepositions, adjectives, conjunctions, and interjections are regarded, and negations are tested since they can reverse the polarity of the sentences. The

results are then compared with lexicon resources or corpuses to assign a weight to each word in the sentence and calculate the sentiment intensity and polarity. Figure 5.1 shows the proposed architecture of the sentiment analysis using the Natural Processing Toolkit available in the python programming language.

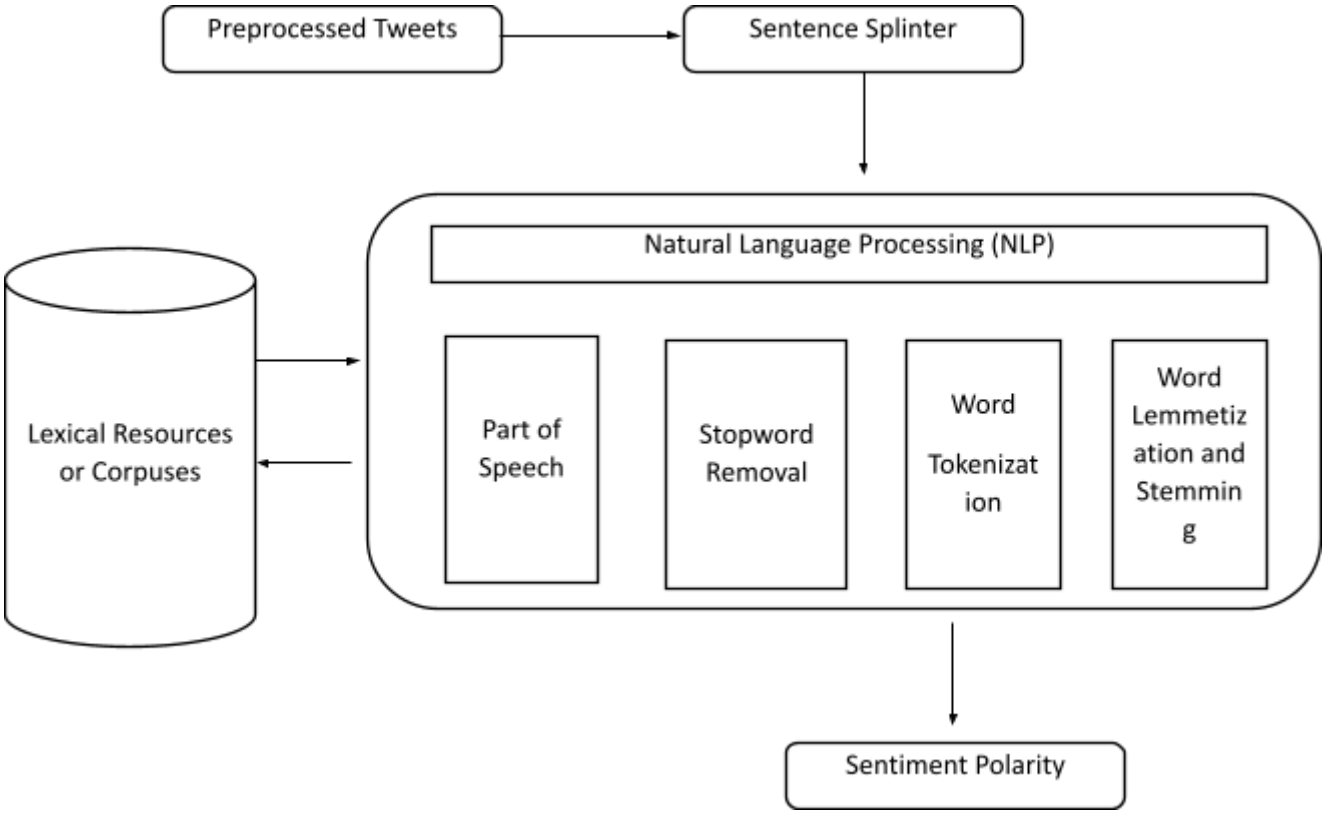


Figure 5.1: NLP-based sentiment analysis using a processed architecture

5.2. Natural Language Processing Toolkit (NLTK)

In Python, Natural Language Processing Toolkit (NLTK) is a platform for linguistic data processing that developers, scholars, and analysts can use. NLTK is an open-source series of modules and corpora that allows people to study and conduct detailed study in NLP (Natural Language Processing). [42]. NLTK is accessible easily for Windows, Mac OS X, and Linus. NLTK is a community-driven, free, open-source project. [42].

NLTK provides an interface to various lexical resources such as WordNet, SentiWordNet and even advanced corpuses like VADER, which is used for processing and analyzing social media texts. Combined lexical resources provides functionality like word tokenization, stemming, tagging, parsing, stop words filtering, etc.

The following subsections will discuss more NLTK libraries and functionality in detail.

5.2.1. Tokenization

Preprocessing of the extracted dataset from Twitter leaves the raw dataset free of noise and unwanted text. The noise and the unwanted text do not contribute to the sentence's overall context. A sentence refers to a combination of multiple words, sentence, emojis and much more. Each individual sentence has its own outcome expression/meaning. In social media platforms, each word in a user tweet can have an emotion or a sentiment attached to it. Each user expresses this sentiment in terms of different textual format and languages.

Tokenization is the decomposition a big amount of text into smaller chunks or individual words termed as tokens. These tokens are immensely beneficial for pattern recognition and are used as a basis for stemming and lemmatization. Tokenization could also be used to override sensitive data elements with non-sensitive elements. To achieve this objective, it is vital to comprehend each text pattern. Several modern-day applications like Text Classification, Intelligent Chatbots, Sentiment Analysis, Language Translation, Speech Recognition and much more uses tokenization in Natural Language Processing.

The “NLTK tokenize” sentences module, which is comprised of mainly two sub-modules, is a crucial part of the Natural Language Toolkit. The two sub modules are “word tokenize” and “sentence tokenize” and are explained next.

1. Word Tokenize

Natural Language Toolkit (NLTK) is a package in Python is utilized for word tokenization. It has an abstract method called "word_tokenize()" that splits the sentences into constituent words. For better text comprehension in machine learning applications, the performance of word tokenization can be translated to a Data Frame. It can also be used as a preliminary step for text cleaning operations including punctuation elimination, numeric character removal, lemmatization and stemming. To train the model and predict, machine learning techniques need numeric data. Word tokenization becomes a significant portion of the content (string) to numeric data transformation. Table 5.1 demonstrates the processing of tweet by an example for Word Tokenization in NLTK.

Table 5.1 Example of tokenization of a Word

Processed Tweet	from nltk.tokenize import word_tokenize text = " The democratic party leading polls Washington" print(word_tokenize(text))
Tokenized Tweet	['Democratic', 'party', 'leading', 'polls', 'Washington’]

2. Sentence Tokenize

The abstract method “sent_tokenize()”, the sub-module for Word Tokenization is available for Sentence Tokenization. A self-evident question might arise that why sentence tokenization is required when we have the alternative of word tokenization. For some instances during tokenization, we are required to count average number of words per phrase/sentence. For accomplishment of such processes there is a requirement of both NLTK Sentence tokenizer as

well as NLTK Word tokenizer to calculate the ratio. Since the response will be numerical, such performance is useful for processing the data. Table 5.2 shows the steps used in tokenizing a sentence and Table 5.3 describes the example of tokenization of sentence.

Table 5.2 Algorithm for tokenizing sentences

Input- Cleaned Tweets **Output-** Tokenize sentences For each user tweet in dataset: Tokenize the sentence into words. **Return** Tokenize list of words

Table 5.3 Example of tokenization of a sentence

Processed Tweet	from nltk.tokenize import sent_tokenize text = "HURRAY! The democratic party leading polls in Washington" print(sent_tokenize(text))
Tokenized Tweet	['HURRAY!', 'The democratic party', 'leading polls in Washington’]

5.2.2. Word Stemming and Lemmatization

The techniques for normalizing the terms in the sentence are word stemming and lemmatization. They apply to the ability to find a text's source or basis.

Stemming

The method of developing morphological variants of a root/base word is known as stemming. Stemming algorithms or stemmers are words used to describe stemming programs. A stemming

algorithm basically reduces the words to its basic meaning by eliminating the suffixes. In natural language processing (NLP), stemming is an important aspect of the pipelining process. Tokenized words are loaded into the stemmer.

The two possible errors to be encountered in stemming process are over-stemming and under-stemming:

1. **Over-Stemming**

Over-stemming occurs when two separate stemmed words are stemmed inside the same root. False positives can also be related to over-stemming.

2. **Under-Stemming**

When two words have the same root but have different stems, this is defined as under-stemming. False negatives may be represented as under-stemming.

There are multiple stemming techniques available. Below mentioned are some of the widely popular techniques:

1. **Porter's Stemmer:** Porter's stemmer uses a dictionary of English suffixes that are made up of a mixture of smaller and simpler suffixes. This stemmer is renowned for its quickness and ease for perceived usefulness. It is mainly used in data mining and information retrieval problems. Its implementations, however, are restricted to English words.
2. **Lovins Stemmer:** Lovins stemmer has considerably larger dictionary of suffixes than Porter stemmer. It has efficiently substituted space for time and because of its large suffix collection, it only requires two major steps to delete a suffix compared to eight steps required by the Porter algorithm.

3. **Dawson Stemmer:** The Dawson Stemmer is similar to Lovins stemmer in that it expands the method by increasing the number of suffixes. These suffixes are preserved and arranged compared to their last letters and length, making retrieval easier and thus more efficient. They are grouped as a series of distinct character trees for easy access. The Dawson Stemmer has over 1000 suffixes.
4. **Snowball Stemmer:** The Snowball Stemmer, unlike the Porter Stemmer, can also map non-English words. Snowball Stemmers is a multi-lingual stemmer since it supports other languages. The Snowball stemmer come from the NLTK kit as well. This stemmer, which processes short strings and is built on a programming language called "Snowball," is the most used stemmer. Snowball Stemmer, also known as Porter2 Stemmer, is a much more powerful stemmer than Porter Stemmer. The Snowball Stemmer has a faster computing pace than the Porter Stemmer because of the improvements proposed.

The Snowball Stemmer algorithm is used to remove the suffix from a term to isolate the root. It uses the stemming method to remove various suffixes such as –ED, -ING, -ION, -IONS, and gives the term a more abstract meaning by eliminating the same suffixes from the language. It is the advanced version of Porter stemmer. It is almost universally accepted as better than Porter stemmer. Table 5.4 demonstrates the stemming of words by Snowball stemmer.

Table 5.4 Example of Snowball Stemmer

Word	Stem
Cared	Care
Fairly	Fair
Easily	Easy

Singing	Sing
Sportingly	Sport

Lemmatization

In stemming a portion of the word is chopped at the end to extract the stem of the word. Several algorithms are used to determine how many letters should be deleted, but these algorithms do not have precise understanding of the context of the word of the language in which it is used. The algorithms in lemmatization, on the contrary, look up the meaning to work out what a word means before reducing it to its root word, or lemma. As a result, a lemmatization algorithm would recognize that the word "better" is derived from the term "good", hence the lemme is "good". A stemming algorithm, on the other side would be incapable of doing the same. Over-stemming or under-stemming may arise, and the word 'best' may be simplified to 'bet' or 'bett', or merely kept as 'better'. However, there is no way to minimize it to its root word 'good' by stemming. This is the key contrast between stemming and lemmatization. Table 5.5 demonstrates with example the implementation of Lemmatization over stemming.

Table 5.5 Example of using Lemmatization over Stemming.

'Caring' → Lemmatization → 'Care'
'Caring' → Stemming → 'Car'

Lemmatization is the process of converting a word's structural form to its root; it is the dictionary form of a word. NLP does complete morphological analysis to identify each word's lemma correctly. Word Lemmatization in NLTK uses a WordNet database, a corpus of synonyms or a thesaurus. The significant difference between lemmatization and stemming is that lemmatization

also considers the part of speech. Both the techniques are used to process the sentences in the Twitter dataset before prediction. Table 5.6 describes the algorithm used in performing stemming and lemmatization using NLTK in Python and Table 5.7 describes an example of Word Stemming and Lemmatizing.

Table 5.6 Algorithm of Word Stemming and Lemmatizing

Input- Tokenized Sentences **Output-** Stemmed and Lemmatized lists of words For each word in **tokenized_words:** **IF** word_length> 2: Call snowballStemmer for stemming words Call WordNetLemmatizer for lemmatizing words **Return** Stemmed and Lemmatized list or words

Table 5.7 Example of Word Stemming and Lemmatizing

Processed Tweet	Tonight, FLOTUS and I tested positive for COVID-19 we will begin our quarantine recovery process immediately we will get through this TOGETHER
Tokenized Tweet	[tonight, FLOTUS, and I, **tested**, positive, for, COVID-19, we, will, begin, our, quarantine, **recovery**, process, immediately, we, will, get, though, this, TOGETHER]
Stemmed & Lemmatized Tweet	[tonight, FLOTUS, and I, **test**, positive, for, COVID-19, we, will, begin, our, quarantine, **recover**, process, immediately, we, will, get, though, this, TOGETHER]

5.2.3. Removing Stop Words

Words like "he/she", "it", "I", "is", "the", "we", "be", "what", "where", etc. are considered as stop words in the NLTK. It is generally believed that excluding stop words has little detrimental effect on assessing the text's polarity and reduces the total scale of data. Table 5.8 shows a processed tweet's example after removing the stop words from the original tweet.

Table 5.8 Example of removing stop words from a user tweet.

User Tweet	Tonight FLOTUS and I tested positive for COVID-19 we will begin our quarantine recovery process immediately we will get through this TOGETHER
Processed Tweet	Tonight FLOTUS tested positive COVID-19 will begin quarantine recovery process immediately will get through this TOGETHER

The previously discussed processes make the dataset ready for performing sentiment analysis and calculating polarity for each user tweet. The steps ensure that the dataset is ready and allows the sentiment analyzers to determine each user tweet’s polarity effectively and efficiently. The following sections in this thesis will discuss the VADER sentiment analyzer's use.

5.3. Sentiment Lexicons

A sentiment lexicon is a set of lexical features (e.g., words) that are classified as positive or negative depending on their semantic orientation [5]. One can also manually create and validate the list of opinion-bearing features, which most times prove to be more reliable for the development of sentiment lexicons, which is very time-consuming too. These reasons contribute to researchers having to use predefined and modelled sentiment lexicons.

5.3.1. Semantic Orientation (Polarity-based) Lexicons

Linguistic Inquiry and Word Count (LIWC) is a computer software that analyses text samples for relational, cognitive, procedural, and process elements. It employs the almost 4500 words vocabulary, which is divided into 76 divisions. LIWC is one of the most reliable and validated tools. The researchers use this lexicon to calculate social media text's sentiment polarity. The lexicon developed by LIWC was used to derive political sentiment indicators from tweets [39]. They are also used to estimate when people will develop depression based on text from social media [43]. They use Twitter posts to assess the emotional variability of pregnant mothers. They use Facebook status notifications to gauge community satisfaction and use text messages to separate positive and sad dating couples based on their engagement and communication. [44].

Despite its extensive usage in analyzing sentiment in social media texts, LIWC does not consider sentiment-bearing lexical elements such as acronyms, initialisms, emotions, or slang, which are critical in social media text research [21][45]. Furthermore, LIWC is incapable of detecting variations in emotion strength in the document. For example, "The football match was great!" conveys a more positive sentiment than "The football match was okay." LIWC would score both sentences equally, while such differences matter most in social media sentiment analysis as it demands a more detailed look at each word in the text.

5.3.2. Sentiment Intensity (Valence-based) Lexicons

The ability to calculate the sentiment intensity would help researchers and analysts do a more rigorous and in-depth study instead of just settling with the binary polarity of a particular text's positive or negative sentiment. Valence-based lexicons allow us to find the sentiment intensity in text. It is crucial and helpful for analysts and researchers to understand how a product's sentiment

intensity has changed over time. Due to these reasons, having a lexicon with strong valence would be helpful as it allows analysts and researchers rhetorical analysis of the change in the user interest over time.

The extension of the Wordnet database is SentiWordNet which provides the text's Valence strength. Each word in the sentence is attached with a numerical score that ranges between 0-1. These scores are calculated using a highly complex semi-supervised algorithm. Although it is not the standard gold resource like WordNet, LIWC, etc., it helps correctly determine the strength of sentiment for most domains. NLTK in Python provides the interface for SentiWordNet to use analysts and researchers. But often, SentiWordNet fails to resolve the sentiment intensity for social media text correctly.

5.3.3. VADER

The major limitation of the two approaches discussed was their inability to handle social media texts. Vader promises to leverage the benefits of rule-based modelling to calculate text sentiments by adding social media style texts yet can efficiently handle textual data from other domains. Such is achieved by adding additional lexical features commonly used lexical resources such as SentiWordNet, WordNet, LIWC etc., to express sentiment in social media text (emotions, slang, acronyms). As pointed outby Hutto, in the social media realm, the VADER lexicon works extraordinarily well [21]. The correlation coefficient indicates that VADER (r=0.881) matches ground reality as well as individual human raters (r=0.888) (aggregated group mean from 20 human raters for sentiment intensity of each tweet). Although the corpus used in VADER has more than 9000 lexical features, which is higher than other gold standard linguistic resources, it is also fast enough to use for streaming data online and has no performance lag.

The VADER sentiment lexicon measures the polarity and strength of emotions conveyed in social media posts. It gives a 'normalized, weighted composite score' also known as a compound score: the sum of each word's valence scores and their normalized set of (-1,1), where -1 is the most negative and +1 is the most positive.

In this thesis, after the comparative study of all the types of lexical resources, VADER has been used to calculate every user tweet's sentiment polarity and intensity. The user tweets are then grouped into positive, negative, and neutral categories depending on the threshold value given by VADER. In this research, the value of 0.1 is used. The below illustration shows the rules for labelling each user tweet into a sentiment category based on the threshold.

1. Positive Sentiment: compound score>=0.1
2. Negative Sentiment: compound score <= -0.1
3. Neutral Sentiment: -0.1 < compound score < 0.1

The succeeding chapter of this thesis will focus on data analysis and the data's results. A comparative study is conducted for the two candidates and their political parties based on the volume of tweets, Change in sentiments, etc., from 15th October to 8th November 2020.

CHAPTER 6

DATA ANALYSIS AND RESULTS

In this chapter, the objective is to examine the collective Twitter data gathered over a month of campaigning and show the findings of sentiment analysis for each political leader and his or her political party.

6.1. Data Distribution

To understand the transition and analyze campaign growth, the data was extracted of user tweets from 15th October 2020 to 8th November 2020. The dataset contains approximately 1.5 million tweets, which is after filtering the tweets based on the keywords listed in the data collection section, after conducting data preprocessing by cleaning the raw tweets and mapping the tweets to the party and its candidate. Table 6.1 show the distribution of tweets by party and candidate.

Table 6.1 Number of Tweets by Candidate

Candidate/Party	Number of Tweets
Joe Biden (Democratic Party)	1,83,734
Donald Trump (Republican Party)	3,12,746

The distribution of user tweets shows that during the election campaigns, the Republican Party had most public opinions, which shows people expressed more interest towards the Republican Party.

Figure 6.1 show that most of the tweets have been originated from USA followed by UK and India but in each of these countries there is more popularity of Trump over Biden. In this research, the data cleaning was performed for the filtering the tweets by countries having tweeted more than 10000 tweets in dataset to perform sentiment analysis. After filtration of tweets by countries, sorting with highest number of tweets in ascending order we got tweets from the eight essential countries: United States, United Kingdom, India, Germany, France, Canada, Italy, and Australia.

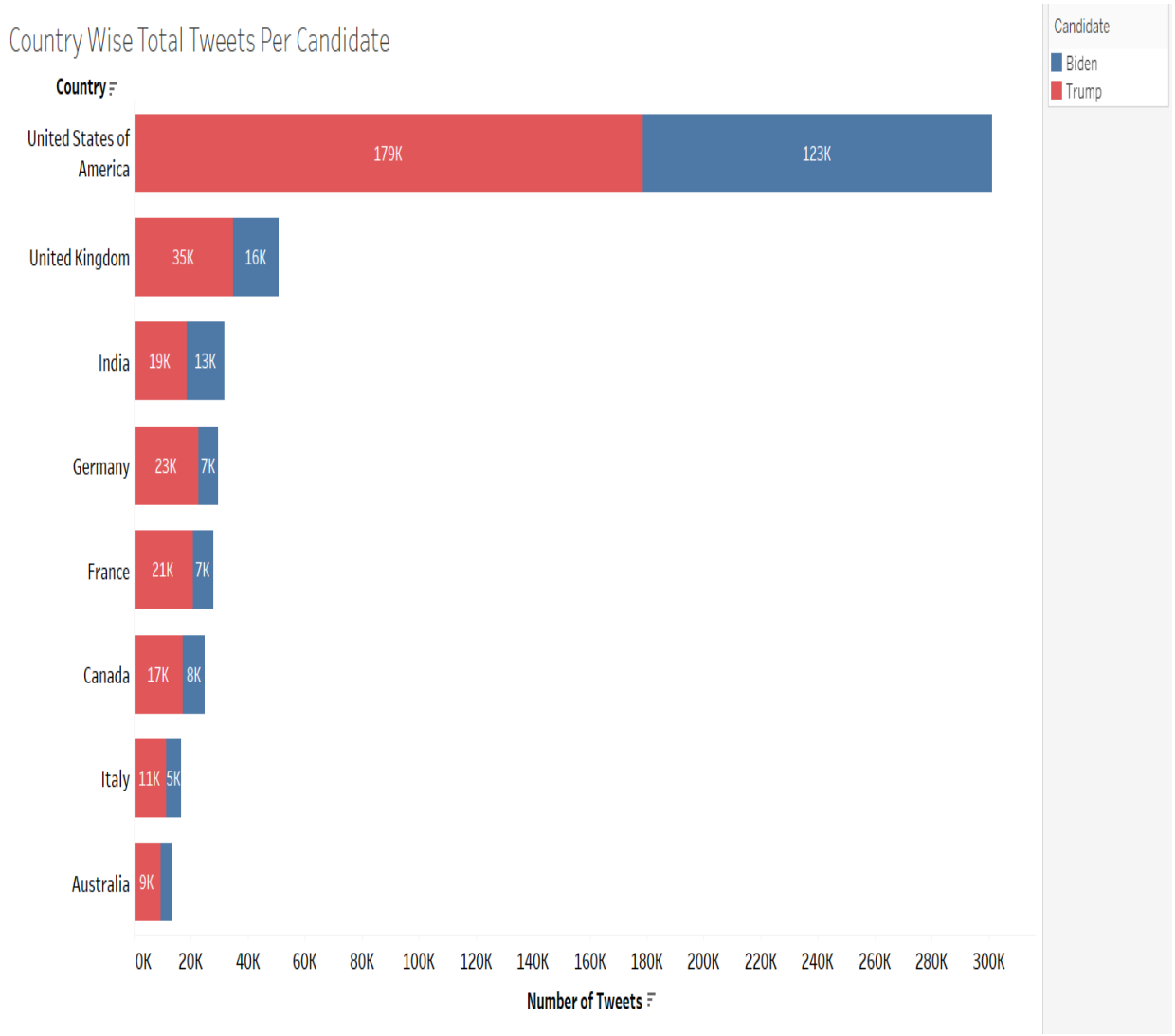


Figure 6.1: Distribution of Tweets by country for each candidate

Table 6.2 Distribution of Tweets by country for each candidate

Countries	Donald Trump	Joe Biden	Total Tweets
United States	179K	123K	302K
United Kingdom	35K	16K	51K
India	19K	13K	32K
Germany	23K	7K	30K
France	21K	7K	28K
Canada	17K	8K	25K
Italy	11K	5K	16K
Australia	9K	4K	13K

6.2. Analyzing Popularity of Party and its Candidate

Analyzing the popularity of parties and candidates, the Twitter sentiment analysis is used to study the user sentiment changing over time. Parties and their leaders were required to understand this shift in sentiment to adjust and change their campaigning strategies over the period during the elections. Moreover, it is also important to relate this to the change in the volume of user tweets during the elections for each party and its leader/candidate.

6.2.1. Volume Analysis

Figure 6.2 presents the frequency of user tweets changing over the last month during election time for the two parties considered for comparison.

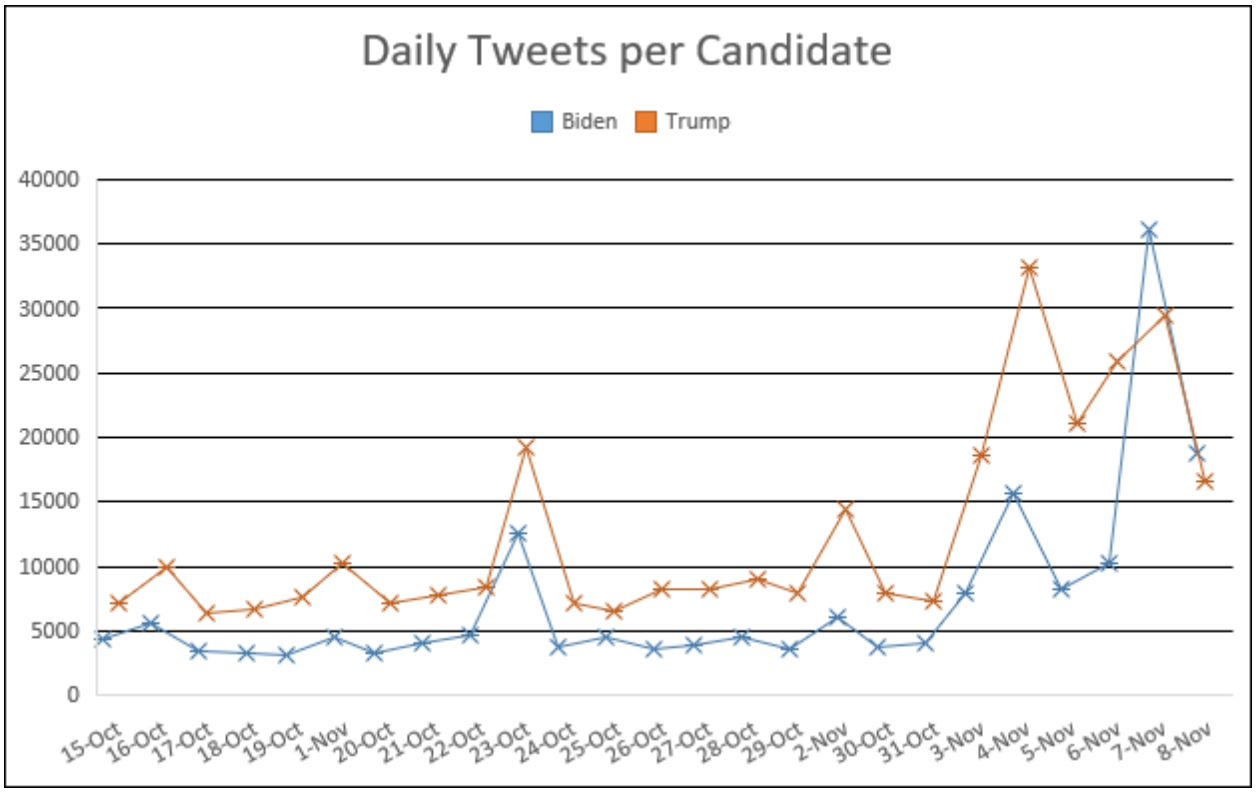


Figure 6.2: Change in Frequency of tweets for each Candidate

It can be clearly observed from Figure 6.2 that the frequency of user tweets is more for the Republican Party than it is for the Democratic party. Moreover, the frequency of the user tweets decreased for the Republican Party over the month as the Democratic Party gained popularity during election on 3rd November. The trend is similar in the frequency of the user tweets for the candidates, Joe Biden, and Donald Trump. This can be easily observed from the Figure 6.3 and Figure 6.4.

6.2.2. Analyzing Change in Sentiment

This section focuses on analyzing the change in sentiment of parties and their candidates during election campaigns. As mentioned earlier, the election campaigns kicked off in March but were more heated between the month of October and November. Figure 6.3 and Figure 6.4 displays the daily change in sentiment for each candidate during the election period from 15th October until 8th November.

Since the Republican Party was already elected and in office since 2016, it was expected to have higher overall sentiment between October and November. The plot in Figure 6.3 and Figure 6.4 show that the Republican Party has a higher sentiment score than the Democratic Party. However, the Democratic Party had more positive sentiments during the election campaign period than the Republican Party.

A similar trend is observed for change in sentiment for each candidate of the parties over the campaigning period. Figure 6.3 show that the overall sentiments of user tweets for the candidate Joe Biden increased as the elections approached. The candidate appointed by the Republican Party, Donald Trump, had more negative sentiments on the user tweets.

From the plots in Figure 6.3, it can be clearly stated that the Democratic Party and its leader Joe Biden have a stable overall sentiment and popularity among people throughout the election period. However, Figure 6.4 show that there were overall high fluctuations observed in sentiments for Republican party and its candidate Donald Trump especially when the election was about to happen on 3rd November.

Figure 6.5 show the polynomial trendline indicating the number of tweets with sentiments over the time frame 15th October to 8th November along with sentiments. It can be observed that even though Democratic party and its candidate Joe Biden received few sentiments (tweets) from

public compared to Republican party and its candidate Donald Trump, but they were leading in terms of positive sentiments. There was also a drastic change observed for Democratic party and its candidate Joe Biden in terms of positive sentiments (tweets) just few days before the election day.

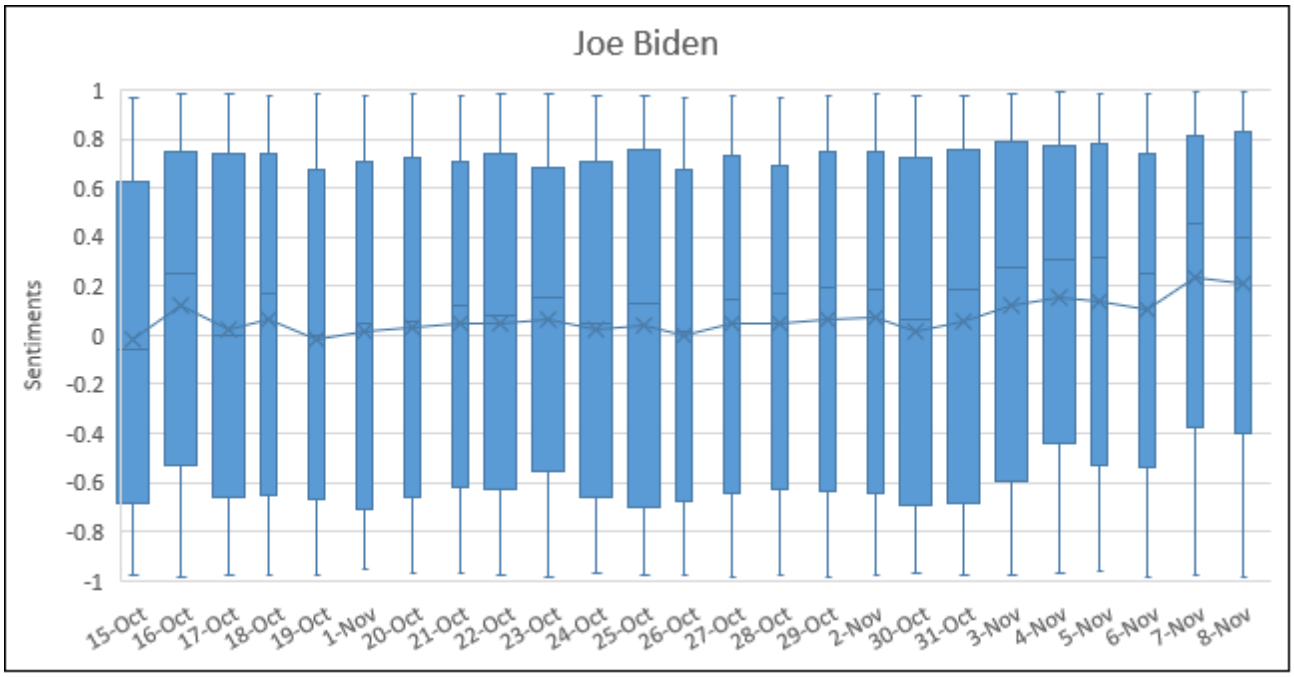


Figure 6.3: ***Daily Change in Sentiment for Joe Biden***

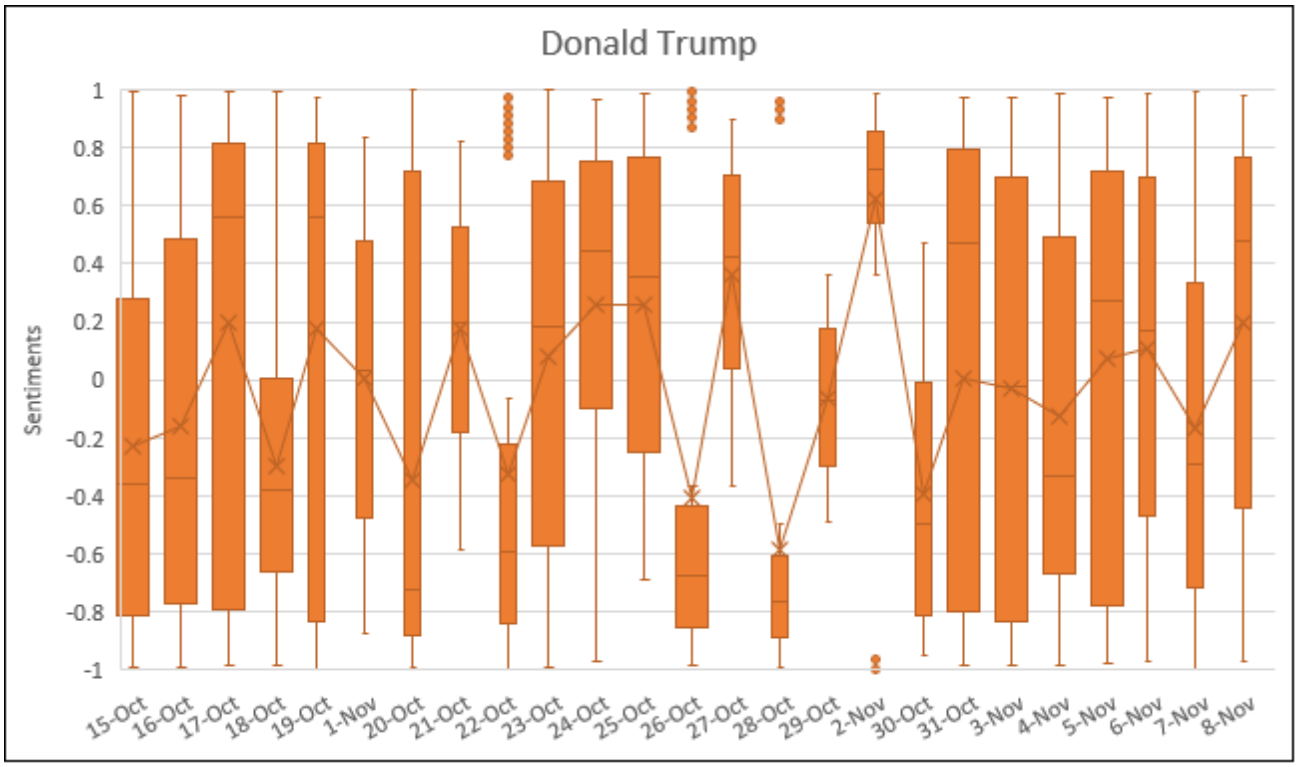


Figure 6.4: Daily Change in Sentiment for Donald Trump

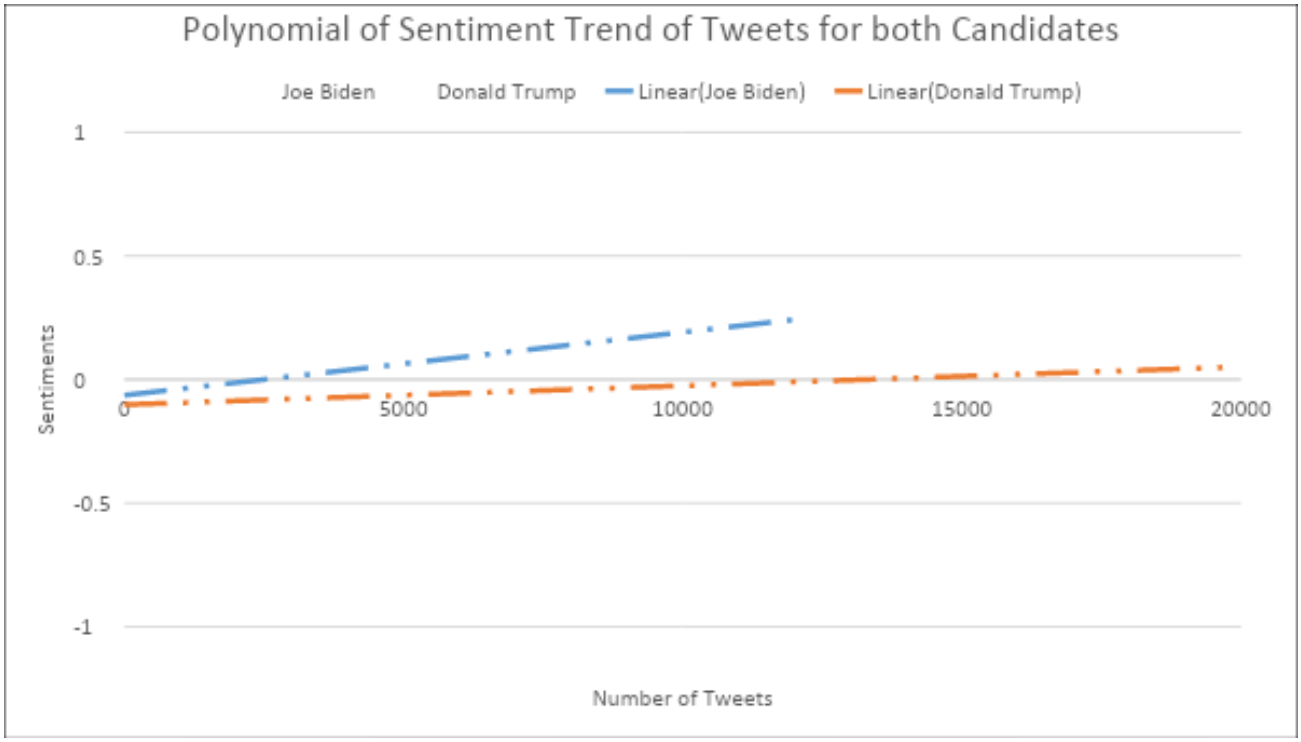


Figure 6.5 Polynomial Sentiment Trend of the tweets for Donald Trump and Joe Biden

6.2.3. Analyzing popularity in terms of Positive Sentiments of Joe Biden and Donald Trump

Figure 6.6 show the distribution of the sentiment for the two candidates. The distinctness of their popularity is interesting. The map (Figure 6.6) shows that Biden was more popular than Trump in southern states of America while Trump was popular in the northern states of America. This trend could be attributed to the fact that Republicans have more supporters in the northern states while the southern states are populated by the minority groups of people supporting the Democrats.

In this research, it is observed that The Republican party who was leading in terms of sentiments in key states namely Alaska, South Dakota and Mississippi won the electoral votes as well by 52.8%, 61.8% and 57.5% respectively. On the contrary, it is noticed that the three key states Florida, Texas, and Ohio where people expressed more sentiments towards Democratic party, but Republican party was leading in electoral votes with 51.2%, 52.1% and 53.3% respectively.

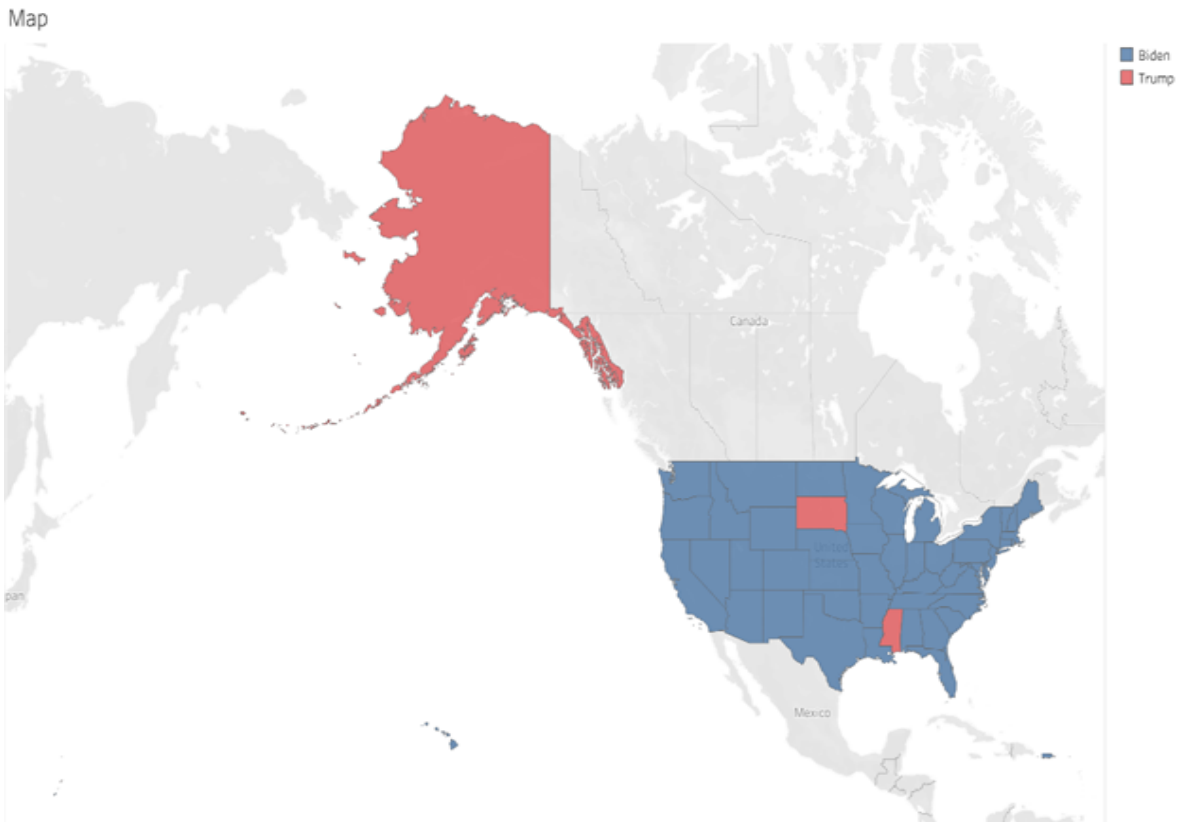


Figure 6.6 Candidates leading in terms of Positive Sentiments in different Provinces

6.2.4. Analysis of Sentiments in Key States with Vote Share

The distribution in Figure 6.7 show that the sentiments for Biden in key states were relatively average. There was a huge range between the negative and the positive sentiments. The people in these key states had more positive things to say about Biden. On the other hand, the overall sentiment for Trump in these states was high. The range between the positive and negative sentiment was very low. Florida took the lead in terms of positive sentiment for Trump. This could be attributed to the fact that this is his home state. The state of Pennsylvania had the highest number of negative sentiments for Trump.

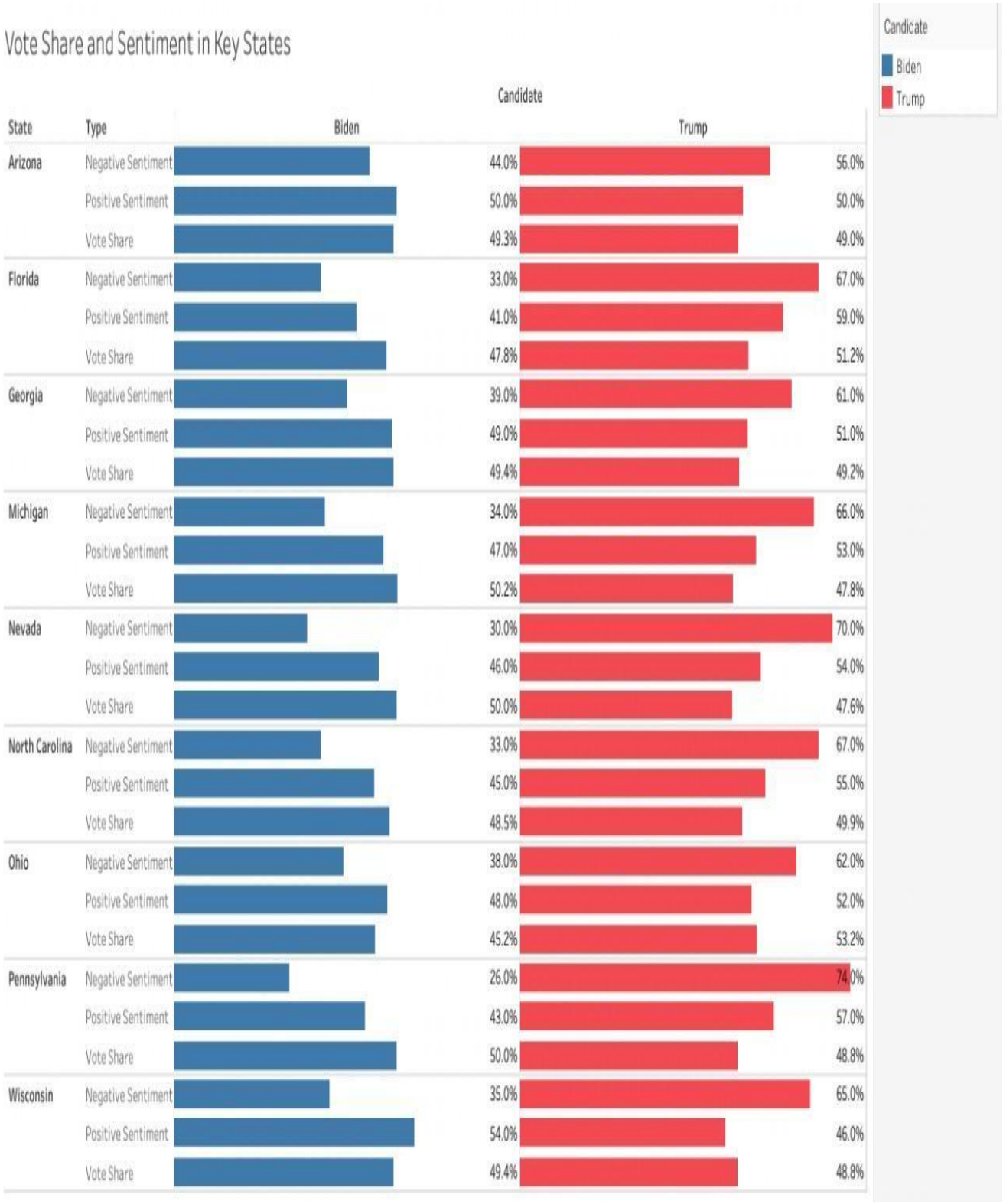

Vote Share and Sentiment in Key States
Candidate
Biden
Trump
State
Type
Biden
Trump
Arizona
Negative Sentiment
44.0%
56.0%
Positive Sentiment
50.0%
50.0%
Vote Share
49.3%
49.0%
Florida
Negative Sentiment
33.0%
67.0%
Positive Sentiment
41.0%
59.0%
Vote Share
47.8%
51.2%
Georgia
Negative Sentiment
39.0%
61.0%
Positive Sentiment
49.0%
51.0%
Vote Share
49.4%
49.2%
Michigan
Negative Sentiment
34.0%
66.0%
Positive Sentiment
47.0%
53.0%
Vote Share
50.2%
47.8%
Nevada
Negative Sentiment
30.0%
70.0%
Positive Sentiment
46.0%
54.0%
Vote Share
50.0%
47.6%
North Carolina
Negative Sentiment
33.0%
67.0%
Positive Sentiment
45.0%
55.0%
Vote Share
48.5%
49.9%
Ohio
Negative Sentiment
38.0%
62.0%
Positive Sentiment
48.0%
52.0%
Vote Share
45.2%
53.2%
Pennsylvania
Negative Sentiment
26.0%
74.0%
Positive Sentiment
43.0%
57.0%
Vote Share
50.0%
48.8%
Wisconsin
Negative Sentiment
35.0%
65.0%
Positive Sentiment
54.0%
46.0%
Vote Share
49.4%
48.8%

Figure 6.7 Sentiment in key states and Vote Share Distribution

Table 6.3 show the Positive and Negative sentiment comparison with votes share on key states.

Table 6.3 Vote Share Comparison with Positive and Negative Sentiment

	Joe Biden			Donald Trump		
States	**Negative Sentiment**	**Positive Sentiment**	**Vote Share**	**Negative Sentiment**	**Positive Sentiment**	**Vote Share**
Arizona	44.0%	50.0%	49.3%	56.0%	50.0%	49.0%
Florida	33.0%	41.0%	47.8%	67.0%	59.0%	51.2%
Georgia	39.0%	49.0%	49.4%	61.0%	51.0%	49.2%
Michigan	34.0%	47.0%	50.2%	66.0%	53.0%	47.8%
Nevada	30.0%	46.0%	50.0%	70.0%	54.0%	47.6%
North Carolina	33.0%	45.0%	48.5%	67.0%	55.0%	49.9%
Ohio	38.0%	48.0%	45.2%	62.0%	52.0%	53.2%
Pennsylvaniya	26.0%	43.0%	50.0%	74.0%	57.0%	48.8%
Wisconsin	35.0%	54.0%	49.4%	65.0%	46.0%	48.8%

6.2.5. Tweet Distribution for each Candidate in Key States

Figure 6.8 displays the user tweets distribution in key states, Florida, had the highest number of sentiments which is approximately 10,234 tweets for Trump while Pennsylvania had 5,086 tweets. The Republican Party received additional tweets than the Democratic Party. The state of North Dakota had the least number of engagements during the campaign. People generally had little to say about other parties in all the states. The states of Arizona, Georgia, and Wisconsin tied in the number of engagements for the two parties.

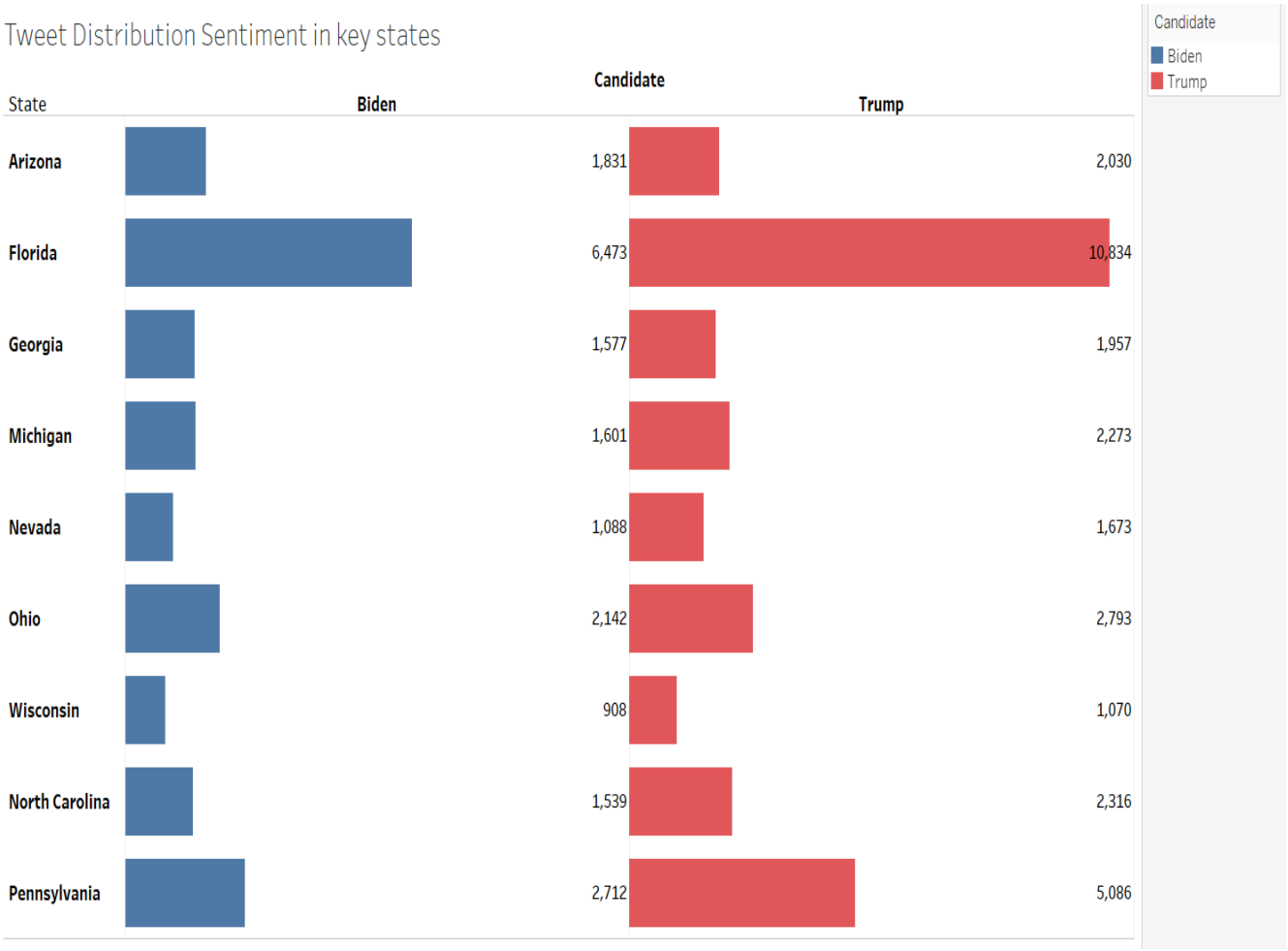


Figure 6.8 Tweets Distribution in key States

Generally, Trump had more tweets compared to Biden although, Figure 6.7 show that Trump had both his sentiment almost tying. The contrary stats here observed is that even though Donald

Trump was leading in the number of tweets from Joe Biden in Florida and Pennsylvania, Table 6.2 show that the vote share results were bit contrary for both provinces. Donald Trump was leading in Vote share in Florida with 51.2% to Joe Biden's 47.8%, while Joe Biden was leading in Pennsylvania with 50.0% with Donald Trump's 48.8%. This clearly shows that Donald Trump also received more negative sentiments along with overall sentiments compared to Joe Biden.

6.2.6. User Tweets detected by filtering 'Fake' keyword.

Figure 6.9 show the distribution of fake tweets for both Biden and Trump. Trump had a total of 4244 fake tweets while Biden had 1517 fake tweets. These fake tweets acted as channels of misleading the people that the individual candidates had traction while in real sense they did not. These tweets were identified based on the keywords/hashtags in the tweets posted and shared by users in twitter.

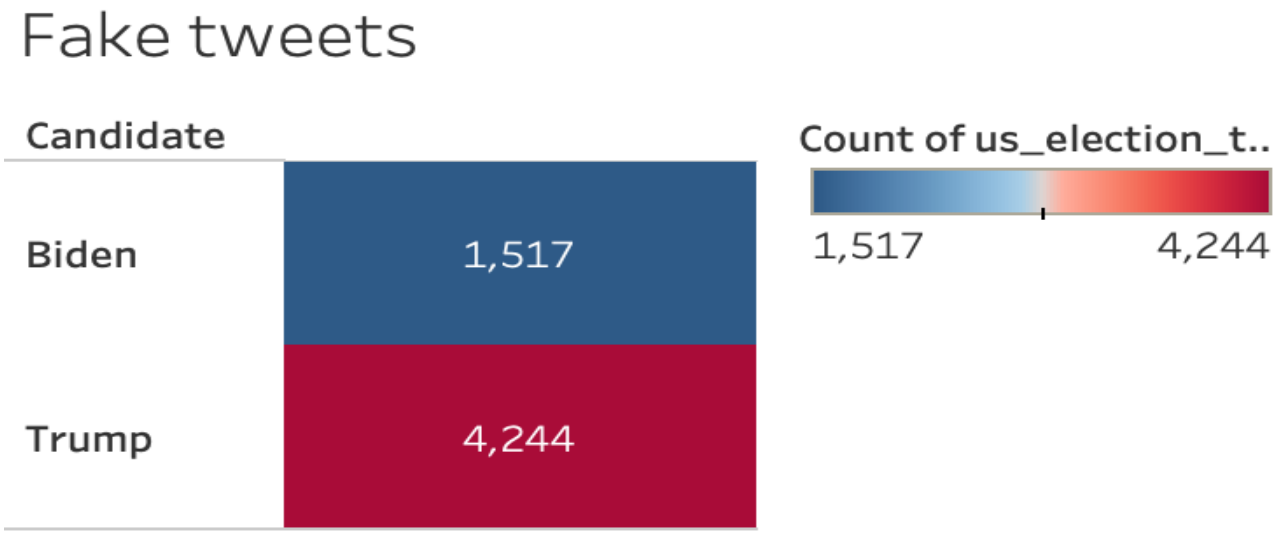


Figure 6.9 Number of tweets detected by 'Fake' keyword for each candidate

6.2.7. Votes Share Distribution in key States

There was an interesting trend in the percentage of votes in North Dakota. Although, the state was not engaging so much in tweets, the voters turned up in large numbers to vote. Donald Trump was found to be the leading candidate in North Dakota with 65.1% of vote share (235,595 votes) compared to Joe Biden with only 31.8% of vote share (114,902 votes). The state also had a great turn up for those people that voted for other parties. There was a landslide win in North Dakota for the Republican Party. However, in the other states, the parties did not have such a huge range or difference. Figure 6.10 show the distribution of votes share in key states for Democratic party, Republican party and other parties.

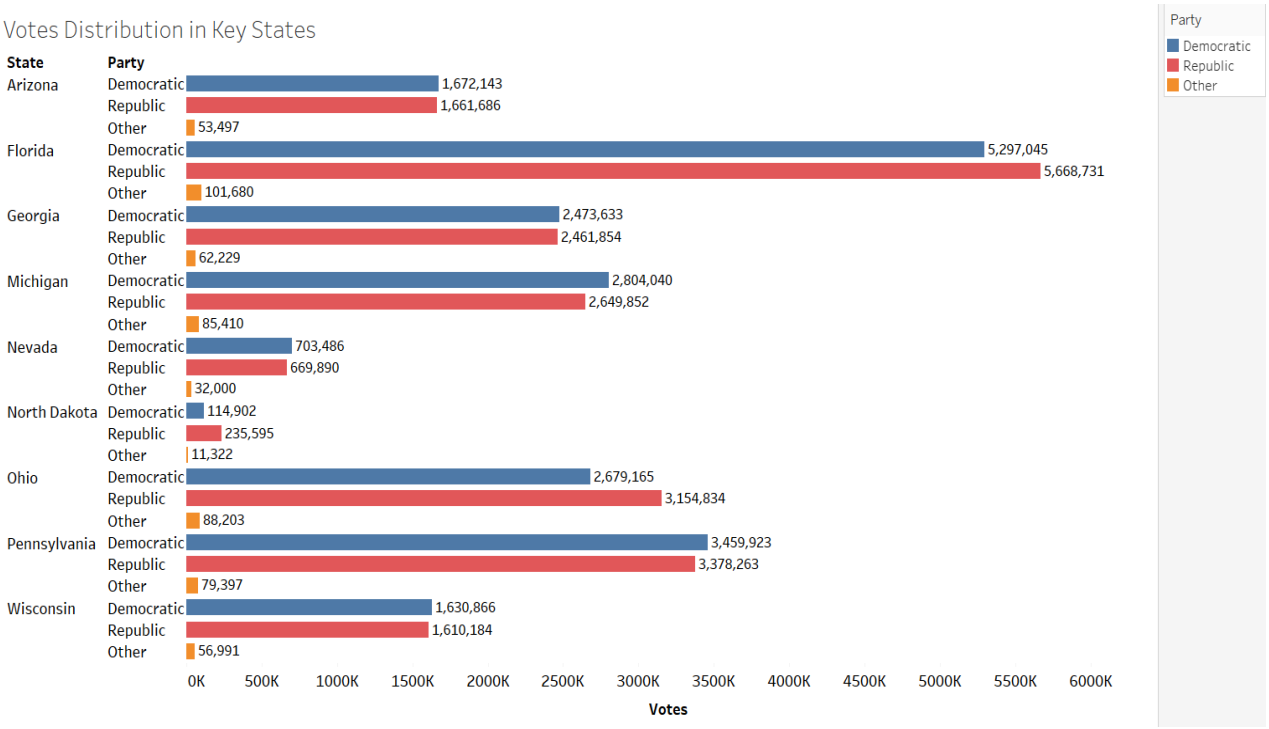


Figure 6.10 Votes Share Distribution in Key States

The Democratic Party won in Arizona by a very close margin of 49.4% compared to Republican party with 49.06%. Democratic party recorded 1,672,143 votes and the Republican party recorded 1,661,666 votes. The other two marginal wins for Democratic party against Republican

party was Georgia with 49.5% of votes (2,473,633 votes) against 49.26% of votes (2,461,854 votes) and Wisconsin with 49.45% of votes (1,630,866 votes) against 48.82% of votes (1,610,184 votes) respectively. While, the Republican party managed to take lead on other key provinces like Michigan with 50.62% of votes (2,804,040 votes), Nevada with 50.06% of votes (703,486 votes) and Pennsylvania with 50.02% of votes (3,459,923 votes) respectively. Figure 6.11 show the distribution of votes share in percentage for Democratic party, Republican party and other parties.

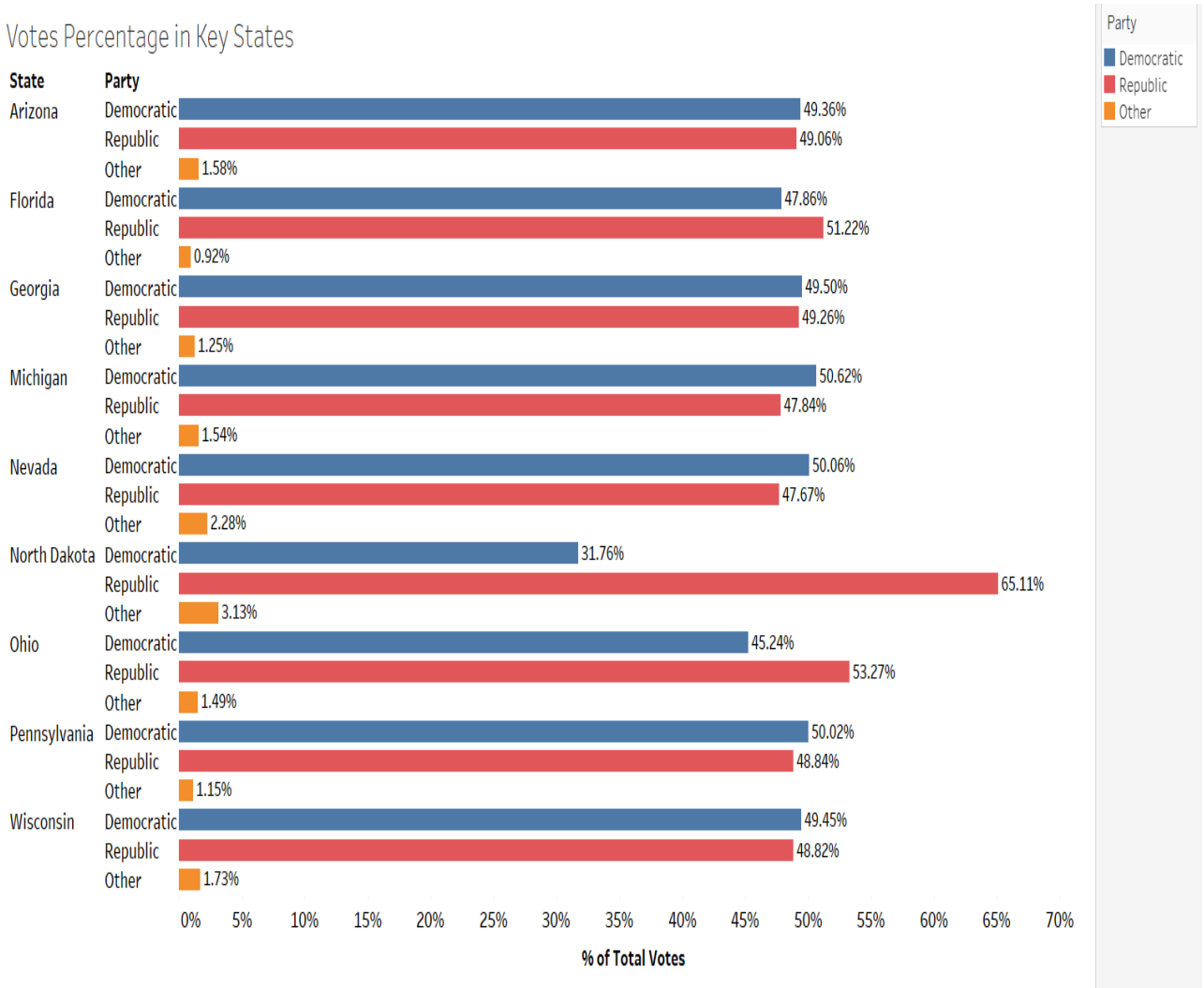


Figure 6.11 Votes Share Distribution in Percentage in Key States

6.2.8. Daily Sentiment for Individual Candidate

Figure 6.12 show the number of sentiments expressed for Donald Trump from October 15 to November 8. The trend for both the positive and negative sentiments for Trump seemed to tie in the entire campaign period. The only time that there was an increase in positive sentiments for Trump was in early November when he shared about his being tested positive for Covid-19, there were also few negative sentiments during this time.

In Figure 6.12, it is observed that Donald Trump received average negative sentiment of 0.19, positive sentiment of 0.18, and neutral sentiment of 0.61. This shows that Donald Trump was almost near about equal in terms of negative and positive sentiments while he received more neutral sentiments comparatively. It is also observed that just before the election day (3rd November) there was a slight fall observed in terms of neutral sentiments for the Trump and rise in positive and negative sentiments.

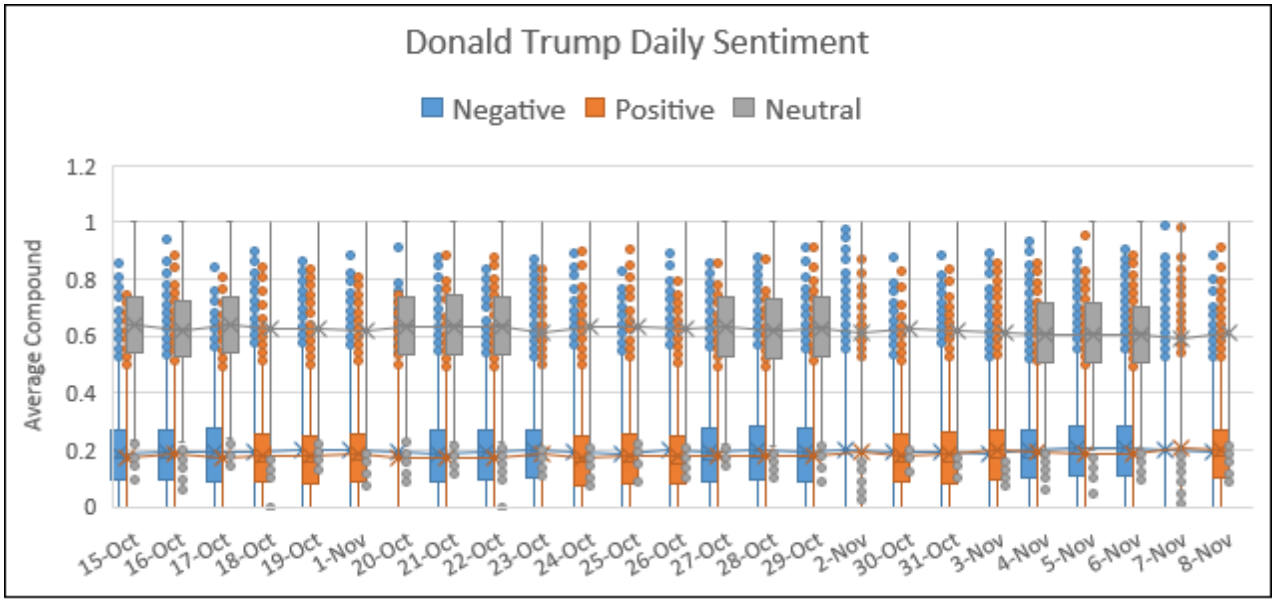


Figure 6.12 Donald Trump Daily Sentiments

Figure 6.13 show the number of sentiments expressed towards Donald Trump from October 15 to November 8. There was a general low engagement in tweets for Biden during the beginning of the campaigns. However, as the day of the elections approached, his popularity increased positively. He had less negative sentiments and more positive sentiments.

In Figure 6.13, it is observed that Joe Biden received average negative sentiment of 0.17, positive sentiment of 0.20, and neutral sentiment of 0.63. This demonstrates unlike Donald Trump; Joe Biden was low in negative sentiments and strong in positive sentiments. In terms of neutral sentiment, both candidates Joe Biden and Donald Trump were almost similar for whole time (15th October to 8th November) even though Joe Biden received overall low tweets compared to Donald Trump during this time frame. It is also observed that just before election day (3rd November) there was a slight fall observed in terms of neutral sentiments for the Trump and rise in Positive and Negative Sentiment.

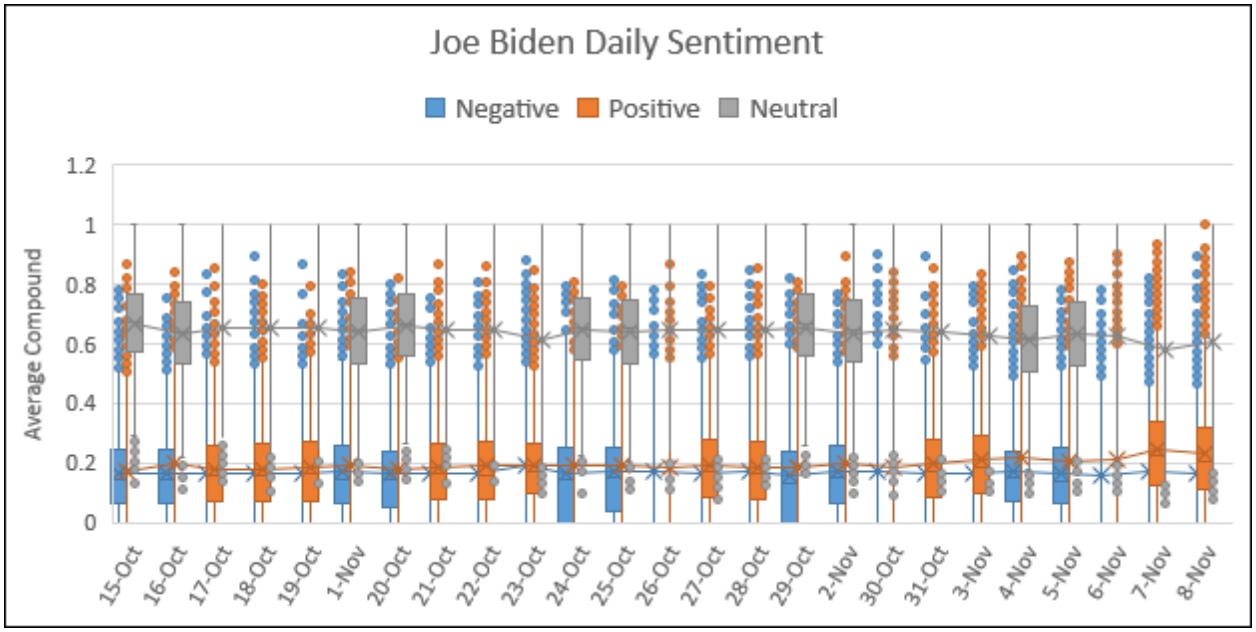


Figure 6.13 Joe Biden Daily Sentiments

6.2.9. Word Cloud

Word clouds are the easiest and most valuable way to grasp the text's general context. Moreover, it is used to depict the frequency of various words or phrases in the user tweet. Word Cloud can be helpful in easily identifying the words or phrases that are commonly used by the public to annotate political parties and their leaders. It can also help unravel the political agenda of each of the parties and how the public is responding to them with its opinions on Twitter. In this section, the research seeks to perform a comparison of word clouds for the two political parties: The Democratic Party and The Republican Party. Figure 6.14 and 6.15 show the word cloud for the Democratic Party and Republican Party, and their leaders Joe Biden and Donald Trump, respectively.

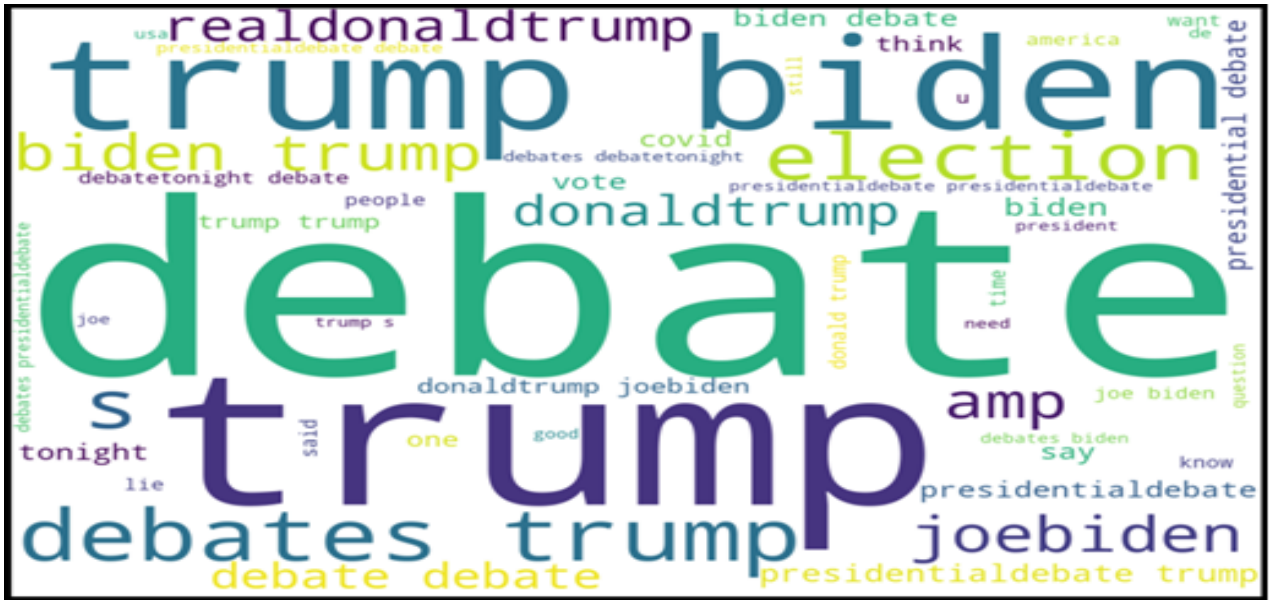


Figure 6.14 Word Cloud Change for Donald Trump

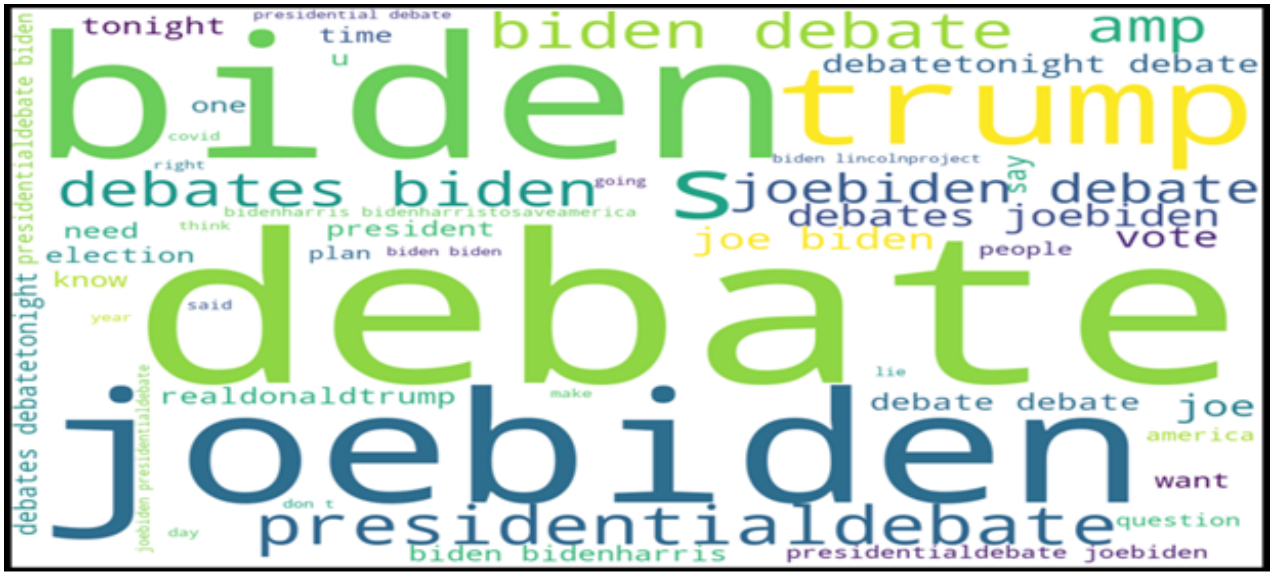


Figure 6.15 Word Cloud Change for Joe Biden

6.3. Results

The sentiment analysis was conducted using VADER for calculating the popularity and sentiment intensity of text. Table 6.4 and Table 6.5 show the distribution of sentiment score for the party and its candidate in percentage, respectively.

Table 6.4 and Table 6.5 show that the Republican party has more positive sentiments (56%) than its candidate Donald Trump (49%). On the contrary, the Republican party has fewer negative sentiments of 69% than its candidate Donald Trump with 73% negative sentiments. The citizens are more neutral about the Republican Party with 55% compared to Democratic Party with 45%. In the overall distribution of the positive, negative, and neutral sentiments, the Republican party had more sentiments than the Democratic Party. This is because the Republican Party stood as the acting party during the election. Table 6.4 and Table 6.5 also show that the distribution of the tweet sentiments for the individual candidates was generally lower than that of the party. These distributions can be interpreted to mean that the parties are more popular in the United Stated than the candidates that flag them. There is a wide range between the overall tweet sentiments for the Republican Party and its running candidate Donald Trump. This trend is unlike the range between the Democratic Party and the running candidate Joe Biden.

Table 6.4 Distribution of Tweet Sentiment by Party

Party	**Positive**	**Negative**	**Neutral**
Republican Party	56%	69%	55%
Democratic Party	44%	31%	45%

Table 6.5 Distribution of Tweet Sentiment by Candidate

Candidate	**Positive**	**Negative**	**Neutral**
Donald Trump	49%	73%	48%

Joe Biden	51%	27%	52%

Table 6.6 show the distribution of votes by age group. It can be observed that younger people (age 18-44) voted more for democratic party whereas elder people (age 45 and above) voted more for the Republican party [47]. The US citizens of age group 45-64 have the highest ratio of participation for voting with 38% voters, followed by the second age group of 30-44 with 23% of voters. This demonstrates that individuals in the 30-64 age range were most likely to use social media sites to voice their opinions on the parties and politicians running for President of the United States in 2020.

Table 6.6 Distribution of Votes by Age group

Age Group	**Republican Party**	**Democratic Party**
18-29 17% of voters	36%	60%
30-44 23% of voters	46%	52%
45-64 38% of voters	50%	49%
65-Over 22% of voters	52%	47%

Figure 6.16 show the distribution of sentiments (Positive, Negative and Neutral) for Donald Trump and Joe Biden in form of a Pie chart. Since the number of tweets extracted for each party and candidate differs, comparing the actual count of tweets for positive, negative, and neutral support will be erroneous. Figure 6.16 show that Joe Biden had 40% positive sentiment, 36%

neutral sentiment and 24% negative sentiment, whereas Donald Trump had 35% positive sentiment, 30% neutral sentiment and 35% negative sentiment.

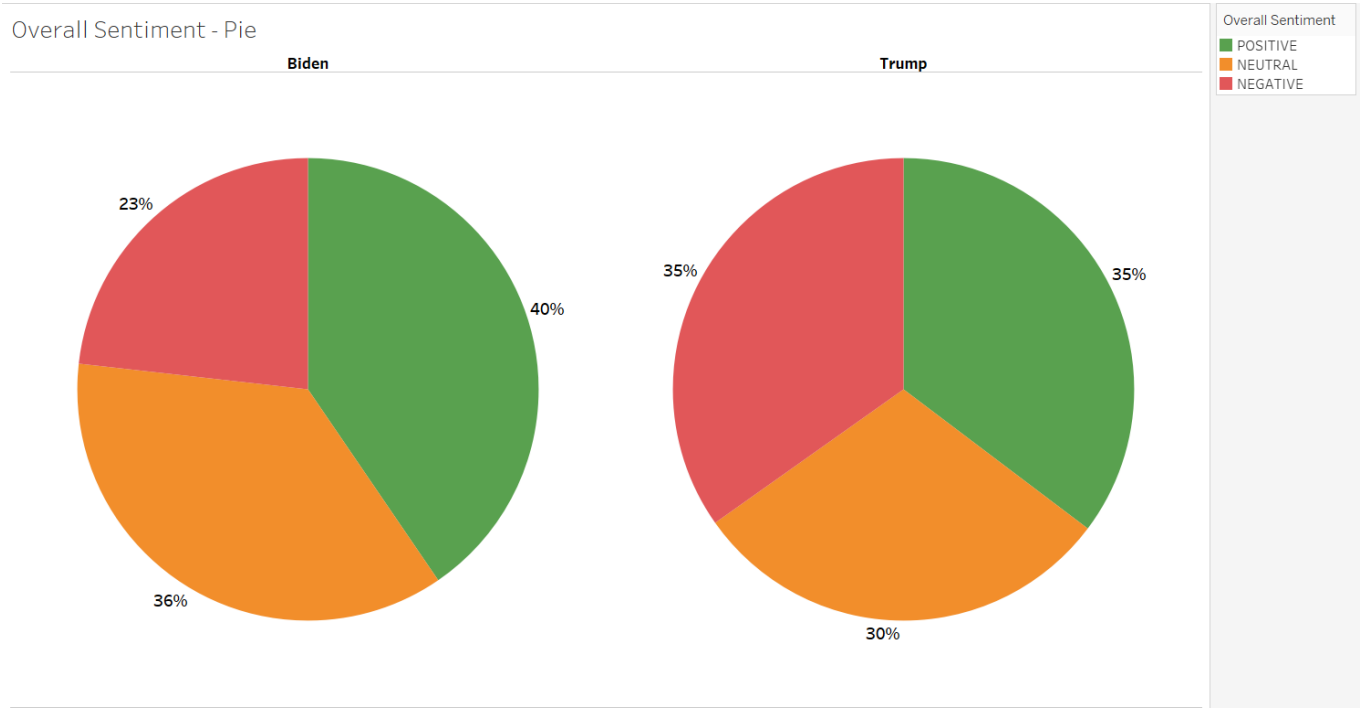


Figure 6.16 Sentiment Distribution for Each Candidate

Considering the sentiments in Table 6.4, the Republican party was leading in terms of positive, negative, and neutral sentiments with the fair bit of margin than Democratic party. However, Republican party was also leading in terms of negative sentiments compared to positive sentiments. From Table 6.7 it can be concluded, although people expressed their sentiments more towards Republican party and its candidate Donald Trump, but more than half of the votes was bagged by Democratic party and its candidate Joe Biden.

Table 6.7 show that out of 538 members of electoral college, Democratic party managed to win 306 electoral votes (270 electoral votes needed to win) and Republican party won 232 electoral votes. The Democratic party and Republican party tied in terms of Senate votes with each party

bagging 50 votes. The House votes also favored the Democratic party to take a lead by 222 votes compared to Republican party who won 213 votes.

Table 6.7 Distribution of seats won in the United States Presidential elections 2020.

Candidate	**Party**	**Electoral Votes**	**Senate Votes**	**House Votes**	**Popular vote**	**Percentage equivalent**
Joe Biden	Democratic	306	50	222	81,268,924	51.3%
Donald Trump	Republican	232	50	213	74,216,154	46.9%

CHAPTER 7

CONCLUSIONS & FUTURE WORK

7.1. Conclusions

In this research, we conducted a Twitter sentiment analysis for the 2020 United States Presidential Elections for two parties, namely, The Republican Party and the Democratic Party, together with their candidates Donald Trump and Joe Biden. We extracted user tweets from Twitter archiver from 15th October to 8th November 2020 based on keywords and hashtags related to the trend observed during the election period. It is important to perform a series of pre-processing steps to remove noise from textual Twitter info. The steps included cleaning up raw tweets and mapping the tweets to the referenced party and the candidate. These overall helped to generate more accurate tweets, which helped in precise calculations of the polarity of tweets.

In this research, the model was proposed to conduct sentiment analysis on a linguistic dataset, emphasizing the use of the VADER sentiment analyzer, which is specially calibrated to measure polarity of text articulated in social media, rather than a conventional dictionary-based solution that uses WordNet for positive, negative, and neutral polarity keywords. It was seen that people had higher positive sentiments and emotions for the Republican party but not for his candidate Donald Trump.

However, the results of the sentiment analysis do not follow the actual election results. The Democratic won the election regardless of what the tweet sentiments had shown. With this, we

can conclude that sentiment analysis and opinion mining at times does not help in understanding how people are responding to the campaign. It can also mislead when it comes to predicting which candidate would lead in the election race. In this new age of the internet, its accessibility and ease of use have given politicians and their parties an avenue whereby they can take full advantage of it by professionally targeting user groups for running their campaigns and leading the election race.

7.2. Future Work

The future scope of this research can include changes in the data extraction techniques by using advance big data frameworks like Apache Spark, as such complex event processing engines are capable of processing huge amounts of data. Due to advances in technology and growing digitalization data, the size of social media texts is growing manifolds and using Apache Spark or Spark Streaming will help us perform the extraction, transformation, and load in near real-time. This can be achieved by running spark jobs in a cluster environment which can be created on AWS and Azure cloud services.

The focus of this research was on accurately classifying data as positive, negative, or neutral. Nonetheless, it might be prudent to try identifying between subjective and objective tweets, since this would be an exemplary initial filter. Furthermore, the system addresses the issue of tweet grouping based on the inclusion or absence of a keyword that we termed "the main attribute," but a more sophisticated method may be built that considers meaning or even synonyms. It should increase not only the number of manually labelled datasets, but also the scope of inputs. The emphasis of this research was on Twitter, although other social platforms may play a vital role in sentiment analysis. Additionally, datasets with a more balanced distribution of tweets from diverse factions should be established.

REFERENCES

1. A. Das and S. Bandyopadhyay, “SentiWordNet for Bangla,” Knowledge Sharing Event-4: Task, Volume 2, 2010
2. ABC News. (, 2020). What social media giants are doing to counter misinformation this election. Retrieved 13 January 2021, from https://abcnews.go.com/Technology/social-media-giants-counter-misinformation-election/story?id=73563997
3. Abroms, L. C., & Craig Lefebvre, R. (2009). Obama's wired campaign: Lessons for public health communication. Journal of Health Communication, 14(5), 415–423.
4. Adam Bermingham and Alan F Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011), pages 2–10.
5. B. Liu, Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012).
6. Baumgartner, J. C., Mackay, J. B., Morris, J. S., Otenyo, E. E., Powell, L., Smith, M.M., ...Waite, B. C. (2010). Communicator-in-chief: How Barack Obama used new media technology to win the White HouseEdited by John Allen Hendricks, and Robert E.Denton Jr. Lexington Books.
7. BBC. (, 2020). US election 2020: What is the electoral college?. Retrieved 13 January 2021, from https://www.bbc.com/news/world-us-canada-53558176
8. C. J. Hutto and E. E. Gilbert, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social

Media," presented at the Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI, 2015.

9. Chen, J. (2020). Important Instagram stats you need to know for 2020. Retrieved 18 January 2021, from https://sproutsocial.com/insights/instagram-stats/
10. Chetashri Bhadane, Hardi Dalal, Heenal Doshi, "Sentiment analysis: Measuring opinions," International Conference on Advanced Computing Technologies and Applications (ICACTA2015), Procedia Computer Science vol.no.45, pp. 808 – 814, 2015.
11. Clement, J. (2020). Internet users in the world 2020 | Statista. Retrieved 13 January 2021, from https://www.statista.com/statistics/617136/digital-population-worldwide/
12. Clement, J. (2020). Twitter: most users by country | Statista. Retrieved 18 January 2021, from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ Conference. Vol. 41. 20
13. D. Das and S. Bandyopadhyay, "Labeling emotion in Bengali blog corpus - a fine grained tagging at sentence level," Proceedings of the 8th Workshop on Asian Language Resources, pp. 47–55, Aug. 2010.
14. D. J. S. Oliveira, P. H. de Souza Bermejo, and P. A. dos Santos, "Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls," Journal of Information Technology & Politics, vol. 14, no. 1. pp.34–45, 2017, doi: 10.1080/19331681.2016.1214094.
15. Dean, B. (2020). How Many People Use Social Media in 2020? (65+ Statistics). Retrieved 13 January 2021, from https://backlinko.com/social-media-users

16. Facebook. (, 2020). Retrieved 18 January 2021, from https://about.fb.com/wp-content/uploads/2020/12/US-2020-Elections-Report.pdf
17. Forsey, C. (2020). Twitter, Facebook, or Instagram? Which Platform(s) You Should Be On. Retrieved 18 January 2021, from https://blog.hubspot.com/marketing/twitter-vs-facebook#:~:text=Ultimately%2C%20Facebook's%20purpose%20is%20to,time%20information%2C%20and%20trending%20news.&text=And%20then%2C%20Instagram%20is%20used%20to%20share%20photos%20and%20videos.
18. Glassman, M., Straus, J. R., & Shogan, C. J. (2009). Social networking and constituent communication: Member use of Twitter during a two-week period in the 111th Congress. Library of Congress: Congressional Research Service.
19. Golbeck, J., Grimes, J. M., & Rogers, A. (2010). Twitter use by the US congress. Journal of the American Society for Information Science and Technology, 61(8), 1612-1621.
20. Graff, G. (2020). 'There Are No Boundaries': Experts Imagine Trump's Post-Presidential Life if He Loses. Retrieved 13 January 2021, from https://www.politico.com/news/magazine/2020/10/30/imagining-post-president-trump-433704
21. Hutto, C. J., Yardi, S., & Gilbert, E. (2013). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text
22. Hwang, J. (2020). What Python package is best for getting data from Twitter? Comparing Tweepy and Twint. Retrieved 18 January 2021, from https://towardsdatascience.com/what-python-package-is-best-for-getting-data-from-twitter-comparing-tweepy-and-twint-f481005eccc9

23. Kim, S., Hovy, E. Determining the sentiment of opinions. International conference on Computational Linguistics (COLING'04). 2004.
24. Luciano Barbosa and Junlan Feng. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In Proceedings of the international conference on Computational Linguistics (COLING), 2010.
25. M. Eirinaki, S. Pisal, and J. Singh, "Feature-based opinion mining and ranking," Journal of Computer and System Sciences, vol. 78, no. 4. pp. 1175–1184, 2012, doi:10.1016/j.jcss.2011.10.007.
26. M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational. Linguististics. vol. 37, pp. 267-307, 2011.
27. Medhat, W., Hassan, A., & Korashy, H. Sentiment Analysis algorithms and applications: A survey. AinShams Engineering Journal, 5(4), 1093-1113. 2014.
28. Miao, H. (2020). 2020 election sees record-high turnout with at least 159.8 million votes projected. Retrieved 13 January 2021, from https://www.cnbc.com/2020/11/04/2020-election-sees-record-high-turnout-with-at-least-159point8-million-votes-projected.html
29. Moore, H., & Hinckle, M. (2020). Social Media's Impact on the 2020 Presidential Election: The Good, the Bad, and the Ugly. Retrieved 13 January 2021, from https://research.umd.edu/news/news_story.php?id=13541
30. Newsbreak. (, 2020). Florida Black Lives Matter Supporters: Our Protests Treated Differently than Mob that Stormed the Capitol | News Break. Retrieved 13 January 2021, from

https://www.newsbreak.com/florida/jacksonville/news/2142903720185/florida-black-lives-matter-supporters-our-protests-treated-differently-than-mob-that-stormed-the-capitol

31. NYtimes. (2020). Presidential Election Results: Biden Wins. Retrieved 17 February 2021, from https://www.nytimes.com/interactive/2020/11/03/us/elections/results-president.html
32. O"Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. FromTweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM
33. Parrott, W. G. "Emotions in social psychology: Volume overview." Emotions in social psychology: Essential readings. Ed. W. G. Parrott. Philadelphia: Psychology Press,2001: 1-19.
34. PRC. (, 2020). How Democrats and Republicans Use Twitter. Retrieved 13 January 2021, from https://www.pewresearch.org/politics/2020/10/15/differences-in-how-democrats-and-republicans-behave-on-twitter/
35. R. Jose and V. S. Chooralil, "Prediction of election result by enhanced sentiment analysis on Twitter data using Word Sense Disambiguation," Proc. Int'l Conf. on Control Communication & Computing India (ICCC), Trivandrum, 2015, pp. 638-641.
36. Roose, K. (2020). Trump Still Miles Ahead of Biden in Social Media Engagement. Retrieved 18 January 2021, from https://www.nytimes.com/2020/10/22/technology/trump-facebook.html
37. Stromer- Galley, J. (2020). Trump and Biden ads on Facebook and Instagram focus on rallying the base. Retrieved 18 January 2021, from https://theconversation.com/trump-and-biden-ads-on-facebook-and-instagram-focus-on-rallying-the-base-146904

38. Suciu, P. (2020). Social Media Could Determine The Outcome Of The 2020 Election. Retrieved 13 January 2021, from https://www.forbes.com/sites/petersuciu/2020/10/26/social-media-could-determine-the-outcome-of-the-2020-election/?sh=10532db226f6

39. Tumasjan, A.; Sprenger, T. O.; Sandner, P.; and Welpe, I. 2010. Predicting elections

40. Van Rossum, Guido. "Python Programming Language." USENIX Annual Technical

41. Webwise. (, 2018). Explained: What is Facebook? -. Retrieved 18 January 2021, from https://www.webwise.ie/parents/explained-what-is-facebook-2/

42. Steven Bird and Mark Liberman. A formal framework for linguistic annotation. Speech Communication, 33:23–60, 2001

43. De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. Proceedings of the International AAAI Conference on Web and Social Media, 7(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14432.

44. Human Rights and US Foreign Policy by Jan Hancock published in 2007.

45. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys by Dmitry Davidov, Oren Tsur and Ari Rappoport. ICNC/Institute of Computer Science The Hebrew University.

46. Freire, Manuel ; Serrano-Laguna, Angel ; Manero, Borja ; Martinez-Ortiz, Ivan ; Moreno-Ger, Pablo ; Fernandez-Manjon, Baltasar. / **Game Learning Analytics: Learning Analytics for Serious Games**. Learning, Design, and Technology. editor / Michael J. Spector ; Barbara B. Lockee ; Marcus D. Childress. Springer Nature Switzerland AG, 2016. pp. 1-29

47. The New York Times Journal results publication in exit Polls.

https://www.nytimes.com/interactive/2020/11/03/us/elections/exit-polls-president.html

Top of Form

Bottom of Form