University of Sheffield

# Can sentiment analysis of Twitter be used to predict the results of the 2020 US Presidential Election?

Calvin Karpenko

*Supervisor:* Dr. Chenghua Lin

A report submitted in fulfilment of the requirements
for the degree of MSc in Data Analytics

*in the*

Department of Computer Science

September 21, 2021

# Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: Calvin Karpenko

Signature:

Date: 21/09-21

# Abstract

Sentiment analysis is a process where pieces of text are mined for opinions and emotions. There are a large number of applications for this process, allowing companies to gauge customer feedback towards products and services, politicians to gauge opinions of new policies or their campaign, etc. As a result of more people using social media to post their views on a wide range of topics, researchers have been using sentiment analysis on social media postings (primarily Twitter) in order to gauge political sentiment. This has started a new field of election prediction that uses sentiment analysis on tweets in order to predict the outcome of an election. The aim of this project is to build a system where the sentiment towards each candidate of the 2020 US Presidential Election can be determined and tweets about the election are classified as positive, neutral or negative towards that candidate. Under this overarching aim is the objective to find the best model by comparing the performance. Of the models tested, the Linear Support Vector Classifier performed the best with an F1 score of around 0.75-0.76 for both candidates. Once analysed, and with caveats, Joe Biden is predicted to receive more votes than Trump, the results of which, when compared with the result of the election, gives an MAE of 1.545 and an $R^2$ of 0.536.

# Acknowledgements

I would like to acknowledge and thank my supervisor Dr. Chenghua Lin, who has been my guide throughout the project, giving his feedback on ideas and suggestions I had and enabling me to make this project my own. He provided a lot of advice when I sought it and asked questions that made me justify the decisions I made for the project. This made for a better written report.

I would also like to express my deep gratitude to my partner, who has been a large source of support throughout my whole academic experience. She has always been there whenever I needed her and words alone can't express my deep gratitude for her patience and support.

I would like to thank my sister who encouraged me to study as a mature student and then continue my studies to masters level. She has offered me advice throughout my studies and I want to thank her for inspiring me to challenge myself to do better and be better.

Finally I would like to thank the rest of my family who have been a source of support and encouragement, while also reminding me to take time for myself. It is easy to sometimes be so focused that you lose sight of your own wellbeing and I really appreciate the grounding that you have given me.

# Covid19 Impact Statement

Throughout the academic year there were restrictions imposed because of COVID-19 caused additional challenges for the completion of this project. The university switched to online delivery of all teaching, and university buildings were closed. All project meetings were shifted to email correspondence and video meetings.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

On 3rd November 2020 the 59th presidential election took place in the United States. Former Vice President Joe Biden and Senator Kamala Harris ran on the Democratic ticket and challenged the Republican incumbents President Donald J Trump and Vice President Mike Pence. Joe Biden and Kamala Harris were ultimately successful and became the next President and Vice President of the United States.

The election took place in the midst of the Covid-19 global pandemic and many states changed their voting laws to allow voters to vote by mail [1], enabling voters to take part in the election without putting their health at risk. This contributed to a record turnout with Biden receiving more votes than any presidential candidate has ever received before [2]. Conventional campaigning changed as both parties scaled down their national convention plans and held events remotely from different areas across the country. The Biden campaign called off door canvassing and emphasised more online campaigning and precautions were taken during the presidential debates with social distancing, screens and masks being worn.

As a result of the record turnout and the changes made as a result of Covid-19, some states were more prepared than others and this led to a longer wait than normal before the election winner was announced. This caused controversy as claims were made by the Trump campaign of election rigging and the weeks that followed resulted in a number of court cases brought by them in order to stop votes being counted and overturn the election results. The court cases did not stop the certification process and the Electoral College votes were counted on the 6th January 2021. This process was interrupted when pro-Trump supporters attacked the Capitol building, attempting to stop Biden from being declared the next President of the United States. While political polarisation in the US has been discussed in literature [3], this incident showed the scale of the polarisation and how it can lead to such incidents occurring.

Users of social media networks posted their opinions of the candidates, encouraged people to vote and gave commentary throughout the election campaign cycle as well. With such a polarised nature to their politics [4], Americans appear to be very divided and this makes sentiment analysis an attractive tool to investigate the sentiment towards both candidates in this election and see if it can be used as a method of predicting the outcome of the election. Such a tool with good performance would be very useful to political campaigns.

## 1.1 Aims and Objectives

The aim of this project is to ascertain whether the winner of the 2020 US Presidential Election can be determined from analysing the sentiment of tweets posted during the election campaign cycle. This will be achieved by developing a classification model that classifies tweets about each candidate into positive, neutral and negative sentiment. In doing so, the objective will be to test a number of different classification models with manually annotated training data in order to find the best performing model. The best performing model will then be used to classify the full dataset of tweets.

## 1.2 Overview of the Report

The project report is split into a number of chapters that are structured as follows:

- Chapter 1 consists of the introduction to the report and provides the context around the project and an explanation of its aims and objectives.

- Chapter 2 reviews relevant literature about sentiment analysis, sentiment analysis carried out on Twitter, election prediction and various challenges of using social media data. This chapter finishes by considering the machine learning techniques used in similar projects.

- Chapter 3 discusses the proposed project plan by walking through a number of design decisions and explaining the justification for design choices.

- Chapter 4 analyses the results of the project. It determines the best performing model and considers the use of that model on the full dataset of tweets, observing how the results compare to polling averages and the actual election results.

- Chapter 5 presents conclusions about the project and its main findings and proposes further work that can be carried out to improve the work done and further the field.

# Chapter 2

# Literature Survey

This chapter reviews the literature related to the prediction of elections using sentiment analysis on social media data. An introduction of sentiment analysis is presented and how that applies to social media posts. Literature on predicting elections using social media data is considered and the challenges that researchers face when attempting to predict elections using such data. Finally, an overview of machine learning approaches that have been used in the literature is presented.

## 2.1 Sentiment Analysis

Sentiment analysis is the use of natural language and text processing techniques to process and analyse emotions, opinions and attitudes of written text. It is used to determine the sentiment about a plethora of different elements ranging from politics, brands, products, services, people and more.

Such analysis has wide applications as many organisations are interested in finding out whether the public has a positive or negative view of their products, services, campaigns, policies, etc.

### 2.1.1 Types of sentiment analysis

In his book, Liu[5] describes three different levels of sentiment analysis that current research is focused on: *document level*, *sentence level*, and the *entity and aspect level.*

At the **document level**, the sentiment is classified of an entire document. This level makes the assumption that opinions are being expressed about a single entity and is therefore not useful for documents that discuss multiple entities.

At the **sentence level**, the sentiment of each sentence within the document is classified. This task is really two tasks, where the first task is to classify a sentence as objective or subjective and the second task is the sentiment classification. It is important to note that objective sentences can also imply the existence of opinions.

At the **entity and aspect level**, the sentiment is much more granular, focusing on the sentiment towards features of the entity and the entity itself, rather than the overall sentiment of the sentence and document. It considers not only the sentiment of the opinion, but the subject of the opinion as well. It allows the recognition of multiple opinions to be held about a particular entity. An example of this would be a piece of text where a viewer might review a movie(an entity) and say that the plot(an aspect) was terrible but the soundtrack(another aspect) was excellent. This sort of granularity just would not exist at the document or even sentence level and being able to recognise different aspects and entities allows an opinion to be determined at a more nuanced level.

These sorts of opinions are considered *regular opinions*, but *comparative opinions* can also be mined, such as "the plot of this movie sequel is worse than the first movie", or "this restaurant's service is better than that of another restaurant". Such opinions compare multiple entities based upon shared aspects.

### 2.1.2 Quintuplet Basis of Opinions

For tasks that involve sentiment analysis, Liu and Zhang[6] developed a basis that defines the constitution of an opinion. They defined an opinion as a quintuple $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ where:

- $e_i$ is the name of the entity.
- $a_{ij}$ is an aspect of the entity (or a feature of the entity).
- $oo_{ijkl}$ is the orientation of the opinion of the aspect $a_{ij}$ of entity $e_i$. This is essentially the sentiment towards the feature of the entity given.
- $h_k$ is the opinion holder.
- $t_l$ is the time when the opinion is expressed by the opinion holder.

Liu and Zhang considered that these five components are essential and that without any of them, it can be generally problematic to work with. An example of this is given in their work of the statement "The picture quality is great". It would be an opinion that is not useful as the object of the opinion is not known.

This basis can be used to essentially take unstructured data such as pieces of text and transform them into structured data. Liu and Zhang describe the objective of opinion mining as discovering all of the quintuplets in a collection of opinionated documents. They state that a series of tasks need to be conducted in order to achieve this objective. These tasks are described as follows:

1. **Entity extraction and grouping:** Extract all entity expressions in the document and group synonymous entities together.
2. **Aspect extraction and grouping:** Extract all aspect expressions in the document and group aspect expressions together into clusters.

3. **Opinion holder and time extraction:** Extract the opinion holder and the time that the opinion was expressed from the text.

4. **Aspect sentiment classification:** Determine the sentiment of each opinion about an aspect.

5. **Opinion quintuple generation:** Generate quintuples with the structured data that have resulted from the above tasks.

## 2.2 Sentiment analysis on social media posts

Websites in the early ages of the internet were more simpler than they are today. One would simply visit a web server and the web server would produce a static page created by the operator of the website. This delivery of purely static content (with no content directly generated by their users) has been referred to as ”Web 1.0” and the ”read-only” web [7]. Such content was largely edited and uploaded by the website operators to the web server.

The development of websites that allowed users to generate content led to a new era of the Web, referred to as ”Web 2.0” or the ”read-write web” [7]. These websites allowed users to submit content such as text, images and videos to the web server and many websites would display that content without prior moderation.

This led to the advent of social media websites such as Facebook, Twitter and Youtube which have allowed more people than ever before to share their views and opinions with others. These sorts of websites are now used by many people to discuss matters of interest with other users across the world. The development of Application Programming Interfaces (APIs) has made access to this content easier, allowing developers to create tools and enabling researchers to conduct research.

Twitter's relatively open API (when compared to its largest competitor Facebook) has resulted in it being extensively used as a data source for sentiment analysis in numerous research projects covering applications such as stock market predictions [8], football win predictions [9], and of course, election predictions (the literature of which will be explored later in this chapter).

Twitter can also be used commercially by brands and organisations to monitor public sentiment about their products and services. An example of this could be film studios measuring public sentiment about an upcoming movie that is due for release, or to get immediate feedback from those who have been to see movie that has recently been released. Another example could be political campaigns announcing a policy and measuring the public opinion on that policy.

In their 2015 systematic literature review, Jungherr found 127 studies that addressed the use of Twitter during election campaigns, describing Twitter as a ”pervasive tool in election campaigns”. They state that candidates, journalists and voters are increasingly using Twitter as a place to discuss politics and research reactions to policies [10].

## 2.3 Election Predictions

There has been a substantial amount of literature that has studied the use of Twitter to attempt to predict the outcome of an election. Studies have attempted to predict the outcome of elections in the US [11], UK [12][13], Germany [14], Indonesia [15], among others.

The first study to attempt to predict the outcome of an election using tweets was the study by Tumasjan et al. in 2010 [14], that used Twitter to predict the outcome of the 2009 German federal elections. They considered approximately 104,000 tweets in the run up to the election. Their conclusion was that the number of mentions of a party can be a reflection of the voter share and that it comes close to the election results. They used the Mean Absolute Error (MAE) for their metric which was 1.65%. However, Gayo-Avello et al in 2011 used the same methods (slightly modified to reflect the difference in the electoral system) in the 2010 US Senate special election in Massachusetts and the 2010 general Congressional elections [16]. They were unable to replicate the same success, finding a MAE value of 17.1% for the volume method and 7.6% for the sentiment analysis method.

Gayo-Avello's meta analysis[17] of the literature relating to using Twitter to make election predictions brought a lot of valid criticisms about the research that had been done, and made a number of recommendations. One of the primary criticisms of the studies that had been conducted up until that point was that none of them had actually made a prediction, with analysis taking place after the election result had been declared. There was also valid criticism that most of the studies carried out had only been carried out on single cases, making it difficult to see if there is consistency between elections. He identified six weaknesses in the literature at the time of his analysis.

- **None of the literature are forecasts.** The research is conducted and published after the outcome of the election.

- **Performance metrics need to be explored and standardised.** Commonly reported metrics are described as inadequate (for winner prediction) and using MAE as a metric makes it not comparable across different elections.

- **Rudimentary methods of sentiment analysis.** He claims that commonly used methods are only slightly better than random classifiers and they ”fail to catch the subtleties of political discourse”. No attempts are made to consider humour and sarcasm and ultimately methods need to be repeated for more than a single election to show consistency of the method.

- **Trustworthiness of tweets.** He states that all tweets are assumed to be trustworthy when this is not case. Tweets that are spam, misleading propaganda and astroturfing should be detected and filtered out, or at the very least, the method should be tolerant to the noise.

- **Demographic bias.** Such demographic bias is often ignored despite it being well known that there are demographic biases within the population of Twitter, and even within those who actively use Twitter to discuss politics. He states that researchers have not shown that the methods are tolerant to bias, and in fact treat it as noise.

- **Self-selection bias.** This is often ignored and no methods have been proposed to deal with this.

He made it clear that winner prediction and the number of correctly guessed races should be avoided in future studies. He recommended the reporting of MAE as a performance evaluation metric but also suggested analysing other measures and that the MAE of a model should be compared to the MAE of a reasonable baseline.

Skoric et al.'s meta analysis addressed the criticism of using social media data that is unrepresentative of the general population and stated that dismissing it misses capturing the dynamics of opinion formation [18].They expected machine learning techniques for sentiment analysis to be superior to dictionary-based sentiment classifiers because the implicit signals of preferences are captured rather than simply classifying words as positive and negative. This was not supported by the results when using $R^2$ but was supported when using MAE. Skoric et al. explained that while lexicon-based sentiments can capture noise in data, machine-learning based models have a higher precision at measuring the vote shares. They hypothesised that structural features of a model such as likes, friends, retweets etc would outperform sentiment features in predicting electoral outcomes. However they also hypothesised that a combination of the two in a model would outperform structural or sentiment features alone. This was found to be the case with the $R^2$ value being 0.621 for those models that combined structural and sentiment predictors together, comapred to a models with structural features alone returning an $R^2$ value of 0.605.

They discussed the diversity of data sources and identified a research gap in the literature where it is currently unknown whether one data source provides better predictions than the other, or even if a combination is better. They hypothesised that studies that use multiple data sources are more likely to yield accurate predictions when compared to single data sources, primarily because a broader cross-section of the electorate would be covered. Unfortunately due to the small number of studies that do use multiple data sources, there isn't enough to make any definitive conclusions. Skoric et al. found that the studies tend to use MAE (and other forms such as RMSE, absolute error, etc) and $R^2$ and emphasised Gayo-Avello's call for a standardised way of reporting performance metrics. They recommended that $R^2$ be the primary metric as it manifests "low variances and thus are more stable across different studies" [18].

Khan et al.[19] conducted a systematic mapping study in 2021 that identified 787 studies related to election prediction using Twitter. After the application of several criteria, they selected 98 studies, that spanned 28 countries, with USA and Indian elections being covered by more than half of the studies selected. They found that researchers used three approaches: sentiment analysis, volume-based analysis and social network analysis. They found that 64 studies (65%) of the studies they selected used the sentiment analysis approach only. If the political orientation sentiment analysis is taken into account as well, this number rises to 89%

of the studies. In terms of approaches, they found that 52% of the studies utilised supervised learning techniques to carry out the sentiment analysis.

Liu et al.[20] aimed to learn from some of the weaknesses in the existing literature and conducted a study around the prediction of the 2016 US Presidential election in Georgia. They integrated the sentiment of tweets with some economic variables (Unemployment Rate Growth Rate, GDP Growth Rate and the Per Capita Personal Income Growth Rate) and found 97% of counties in Georgia were classified correctly. The regression models were much less accurate with an average of 15% deviation. The estimated percentages of the votes was able to be aggregated to the state level to predict that Clinton would get less than 50% of the vote share.

## 2.4 Challenges of using social media for sentiment analysis

Sentiment analysis that uses social media postings presents a number of issues for researchers that must be considered when developing models. Some of these challenges have been raised in the meta analyses that have considered the literature in the field.

### 2.4.1 Fake and multiple accounts

Users of social networks are allowed to create multiple accounts. Indeed, Twitter specifically has the facility to allow users to link multiple accounts together and switch between them. This is useful from a user perspective, for example someone may have a personal Twitter account for their personal views and a separate account for their work. It is however a challenge for analysis as these multiple accounts can be used for a variety of different purposes. There is therefore a risk that a user can create multiple accounts to push a particular agenda or sentiment and make it appear as though they are more widely held views than they are in reality. Mustafaraj and Mextaxas found this to be an issue in a past Massachusetts Senate election where accounts were specifically created for the purpose of what they described as a ”Twitter Bomb”, designed to attack one of the candidates [21]. This wasn't unique to Twitter, as their later study showed that Facebook was also vulnerable to accounts being created for nefarious means [22].

While some of these accounts can be created for nefarious means, some of them are also created and used for satire. An example of this would be the creation of Twitter accounts satirising Republican Congressman Devin Nunes which led to Devin Nunes filing a lawsuit that he lost [23]. Regardless of the purpose of the accounts, they are not the unique account (or indeed views) of a real person and so this can have the effect of artificially amplifying certain agendas over others.

Users of social networks are also not required to provide any formal proof of their identity. While social networks do have verification schemes, the criteria for being considered a ”verified” user are very specific, which leads to verified status only being given to notable users such as celebrities, companies, public bodies etc. This does not help users or researchers separate a genuine account from a fake account or identify a bot, it only helps them distinguish between an impersonator of a notable user and the genuine user.

### 2.4.2 Demographic representation

Polling organisations take a great deal of care to ensure that their polling sample is reflective of the demographics of the population and when they get it wrong make changes to improve their polling sample, knowing that this is important when it comes to predicting election. The demographics of Twitter's user base has been considered in a few studies. In 2011, Mislove et al. studied the demographics of Twitter users in the US and found that users were not representative of the US population, with the user base being predominantly male and highly populated counties being overrepresented, amongst other things [24] .

In 2019 the Pew Research Center published their findings after surveying a sample of US Twitter users. They found that "Twitter users are younger, more likely to identify as Democrats, more highly educated and have higher incomes than US adults overall" [25]. They also found that a large majority of tweets come from a small minority of tweeters, with the median user only tweeting twice a month but a much smaller number of users tweet with much greater regularity. This behaviour was also found to exist by Mustafaraj et al. [26], and a similar sort of behaviour is described by Tumasjan et al. in their work, with them claiming that 4% of all users make 40% of the tweets [14]. This representation issue is compounded further because there is a self selection bias that is inherent when conducting sentiment analysis on tweets, as noted by Gayo-Avello who also made the point that "people tweet on a voluntary basis and, therefore, data are produced by those politically active" [17]. In essence, the subset of users tweeting about elections is a subset of those users who actively tweet. The subset of those who are active on Twitter is itself a subset of those registered on Twitter, and Twitter's demographics are considered unrepresentative of the population.

Further issues arise because Twitter is a global social network and its users are not located exclusively in the United States. US elections are followed in many other countries due to the influence that the US has across the world and the notability of the US President. In this particular election the views towards the candidates and their policies are very polarised and the outcome would affect other countries and their policies towards the US. As stated earlier, Twitter does not require proof of identity to access the website so one has to trust that someone stating that they are in Texas (for example) is actually in Texas and not based in Russia.

This leads to further complications in conducting sentiment analysis. More generally, it is simply not possible to know whether an individual is eligible to vote in an election. They could be ineligible for a number of reasons such as their age, their immigration status, or their voter registration status. This causes foreseeable challenges as a person's views and opinions on a political candidate may be counted which would increase the bias of any election prediction. Even if they don't have the right to vote in the election, they may express their opinion and that opinion may influence opinions of others who are eligible to vote.

### 2.4.3 Misinformation and bots

Misinformation is information that is false or inaccurate, spread with or without intention to deceive the reader. It has also become synonymous with the term ”fake news” that has become more widely used since the 2016 US Election [27]. Fake news was itself defined by Allcot and Gentzkow as ”news articles that are intentionally and verifiably false, and could mislead users” [28], however it could be expanded to cover information more generally, especially on social media where journalists post news articles or information directly to the public.

It has become a problem for social media platforms, with Bovet and Makse finding that 25% of the tweets they looked at that contained a link to a news outlet during the 2016 US Election cycle was spreading either fake or extremely biased news [29]. Investigations into some of the fake news articles that were posted in 2016 were found to have not only been produced in the US [30], but also from countries such as Macedonia [31].

More recently, Twitter has been seeking to tag tweets reported as misinformation, taking action against some of President Trump’s tweets during the 2020 election cycle by adding a warning label after he claimed that mail in ballots would be fraudulent [32]. Elections have been targeted with misinformation campaigns but information relating to the Covid-19 pandemic has also been affected, with misinformation being described as ”hindering the practice of healthy behaviors (such as handwashing and social distancing) and promoting erroneous practices that increase the spread of the virus and ultimately result in poor physical and mental health outcomes” [33] .

Much of this misinformation is spread by the use of automated accounts more commonly referred to as ”bots”. Bessi and Ferrara found that bots had a potentially distorting effect on the 2016 US election, estimating that 400,000 bots were engaged in the discussion about the election, being responsible for roughly 3.8 million tweets of the 20 million tweets that they studied [34].

This is an important factor to consider as the spread of election misinformation is designed to negatively affect a candidate and this would impact the sentiment towards that candidate. With 400,000 bots estimated to being involved in the election discussion in 2016, it could have happened again in 2020 and so the sentiment towards a candidate might not necessarily match that of the US public.

## 2.5 Machine Learning Techniques

As discussed earlier, many studies have used machine learning approaches to predicting elections via sentiment analysis. These tend to generally involve training a supervised model on a training dataset of tweets that are labelled as positive or negative towards a candidate.

### 2.5.1 Naive Bayes

Naive Bayes is a supervised model that applies Bayes' Theorem which describes the probability of an event based on the prior knowledge of conditions related to that event [35]. Bayes' Theorem is given as:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{2.1}$$

In this equation, $y$ is a label and $x$ is a feature. $P(x)$ is the probability of the feature in the data, $P(y)$ is the probability of the label in the data, $P(x|y)$ is the probability of the feature being $x$, if the label is $y$, and $P(y|x)$ is the probability of the label being $y$, knowing that the feature is fixed at $x$.

Naive Bayes is so-called because it naively assumes all the features are independent of eachother, given the label. This is called conditional independence and as Zhang observes, this assumption is rarely true in the real world [35]. It does however simplify the calculation process to incorporate all of the features as:

$$p(x_i|y) = p(x_1, x_2, x_3, ..., x_n|y) = \prod_{i=1}^{n} p(x_i|y) \tag{2.2}$$

This essentially allows each probability of a feature based on the class to be multiplied together to produce the probability of the features given the class. Naive Bayes uses training data to learn a model and when it predicts a label for a new feature set (for example a new sentence) it calculates the probabilities for each class and predicts the class to be the one with the highest probability.

### 2.5.2 Support Vector Machines

A Support Vector Machine is a model that is capable of performing linear or nonlinear classification, regression and outlier detection [36]. It works by finding a hyperplane that splits the data for each class by a maximum margin. Using sentiment analysis as an example, on one side of the hyperplane there would be positive texts and on the other side there would be negative texts. This is illustrated in Figure 2.1 where there are two distinct classes, represented by the filled circles and the unfilled circles. Hyperplane 1 (the line HP1) does not separate the classes at all, the line HP2 does separate the classes and HP3 also separates the classes. While HP2 separates the classes, the distance (or margin) between the two classes is much smaller than the margin between the classes in HP3 and so HP3 is a much better hyperplane as the margin between the two classes is much larger. The data points closest to a particular hyperplane are called its support vectors.

In a binary classification if the training data is given as $(x_i, y_i), ..., (x_l, y_l)$ where $x_i \in \mathbb{R}^n$ and where $y_i \in \{+1, -1\}$ the decision function is given as

$$g(x) = sgn(f(x)) \tag{2.3}$$

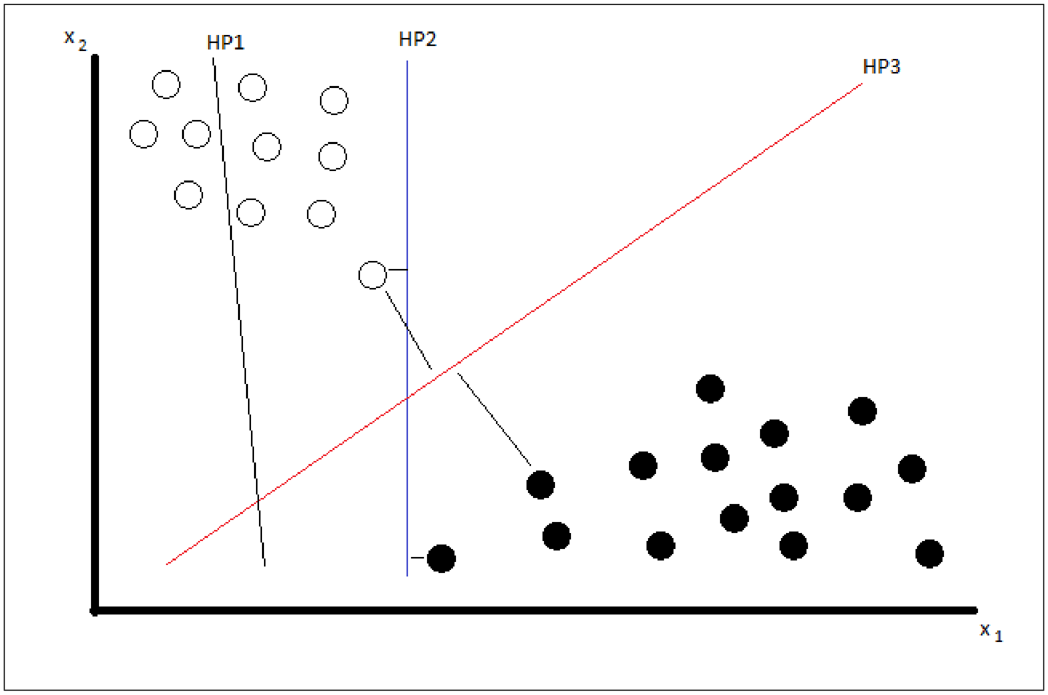

Figure 2.1: Linear Support Vector Machine classification of two classes and lines representing a hyperplane that does not separate the two classes (HP1), a hyperplane that does separate them (HP2), and a hyperplane that separates the two classes with a maximum distance between the two classes (HP3). HP3 is a more optimal hyperplane than HP2.

$$f(x) = \sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b \tag{2.4}$$

where K is a kernel function, $b \in \mathbb{R}$ is a threshold and $\alpha_i$ are weights. that satisfy the constraints[37]:

$$\forall i : 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.5}$$

Equation 2.4 can be rewritten as

$$f(x) = w \cdot x + b \tag{2.6}$$

where $w = \sum_{i=1}^{l} y_i \alpha_i x_i$. Training an SVM model is essentially optimising the following problem:

$$maximise \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.7}$$

making sure to satisfy the constraints in Equation 2.5. Solving this gives the most optimal hyperplane and decision boundary between the two classes [37].

So far only binary classification models have been discussed. Multi-class systems are not so easy as the outputs are not calibrated in a way that a binary system can be. The difficulty arises ultimately because Support Vector Machines do not measure their uncertainty with probabilities. This makes it harder to compare outputs to eachother.

There are a few approaches to multi-class classification though. The one-vs-rest approach essentially turns the multi-class system into a series of binary systems, where one class is treated as positive and the rest of the classes are treated as negative. This approach can lead to ambiguous labels for some inputs though [38].

Another approach is the one-vs-all approach, in which a series of binary classifiers discriminate between all the pairs of classes. For example, for 3 classes, there would be three binary classifiers that would classify between the pairs of classes (C1,C2), (C2, C3), (C1, C3). The issue with this is that it can also create ambiguities as well as taking longer to train and test each data point [38].

### 2.5.3 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a state of the art machine learning technique that was developed by Jacob Devlin and others at Google and published in 2018. It is used for natural language processing tasks and was designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on the left and the right [39].

The model was pretrained on a combination of the English Wikipedia and BooksCorpus dataset. Using a document-level corpus is critical in order to extract long contiguous sentences. Wikipedia is a free-to-use, user-written encyclopedia, and the Wikipedia dataset consisted of 2,500 million words. The BooksCorpus dataset is a large collection of free novel books which consisted of 800 million words.

The model is pretrained using two unsupervised tasks. The first of these tasks is referred to as Masked Language Modelling (MLM), where a percentage of tokens (15%) in the input is masked, and the aim is to predict what those masked tokens are. The final hidden vectors that correspond to the masked tokens are then fed into a softmax function applied over the vocabulary [39].

The second task is referred to as Next Sentence Prediction (NSP). This task aims to train the understanding of relationships between sentences. This is done by binarising a next sentence prediction task so that there is a dataset of sentence pairs, each pair consisting of two sentences (A and B), where in 50% of the examples the sentence B is the next sentence after A and is labelled as ”IsNext”. The other 50% is where B is a random sentence from the corpus and is labelled as ”NotNext”.

There have been studies on the 2020 US Presidential election where BERT has been used to conduct sentiment analysis. In their paper, Singh et al. were able to obtain F1-scores of higher than 0.90, outperforming Naive Bayes and SVM in classifying sentiment [11]. This model was not able to accurately predict the election outcome on a state-by-state basis however this was more likely as a result of the limitations of trying to use tweets with geolocation data that is discussed later.

Chandra and Singh also conducted sentiment analysis on the same election and were able to obtain F1-scores of 75.7 for a batch size of 64, reducing to 70.2 for a batch size of 128 [40]. The model incorrectly predicted that Biden would win states such as Texas, Ohio, Tennessee and Florida, when those states were eventually won by Trump. In the case of Tennessee, the state was won by Trump by a significant margin and no poll indicated that Tennessee was going to be won by Biden.

# Chapter 3

# Project Plan

In this chapter the project will be discussed, starting with a brief reiteration of the aims of the project. After this, there is an in-depth discussion about the methodology and then the choices that have been made for the project and the justifications for those decisions.

## 3.1 Project Aims

The primary aim of this project is to design and implement a classification model that can analyse political tweets made during the 2020 election and classify the sentiment of a tweet into positive, neutral or negative. The objective is to compare the performance of various models that have been used in the literature in order to determine the best performing model that can then be used to classify the full dataset.

## 3.2 Methodology

The method for predicting the outcome of the election will be to create two multi-label classifiers, one for sentiment relating to Trump, and one for sentiment relating to Biden. Each classifier will be able to classify the labels of positive sentiment for the candidate, neutral sentiment, and negative sentiment. The reason why there will be two classifiers and not just a single binary classifier that differentiates between Trump or Biden support is because a single classifier would make it difficult to categorise a tweet that expresses negative sentiment about both candidates. As such, the extra granularity that using a sentiment classifier per candidate provides makes it worth doing.

As much of the literature uses Naive Bayes and Support Vector Machines for their classifiers, this project will also consider the same. Scikit Learn provides a number of different types of Naive Bayes and SVM algorithms so the project will determine which is the best classifier. The models that will be considered under Naive Bayes are Multinomial Naive Bayes, Complement Naive Bayes and Gaussian Naive Bayes. For Support Vector Machines only the linear Support Vector Classification (SVC) and the polynomial SVC (with 3 degrees) algorithms will be considered. The SVM models will use the 'one-vs-rest' method to deal with multiclass classification.

For each candidate, the models will be applied and their performance measured by their F1 score. The best performing model for each candidate will then be applied to the full dataset of tweets. The number of tweets that were classified as showing positive sentiment for their candidates will then be compared. If a person is tweeting positively about a candidate or encouraging others to vote for that candidate, this can be regarded as enthusiasm. Negative sentiment about a candidate can be considered as anxiety towards them, as they wouldn't want a candidate to win if they have negative sentiment towards them. Research has shown that enthusiasm directly affects voting choice and reflects something close to the voting decision itself, whereas anxiety has no direct impact on a voter's choice [41].

Following this, the positive sentiment counts for each candidate will be summed together and the proportion of positive tweets for each candidate will be calculated. This would provide a measure of enthusiasm for a candidate.

## 3.3 Social Network

Twitter has a user base of over 300 million people and much of the content is publicly available to read. It will be used as the social network of choice for the data as it is used as a public forum for politics discussion, with most politicians having some form of presence on Twitter. It has an Application Programming Interface (API) which is very easy to use once registered and accepted. There are also a number of tools that can be used to download tweets and there are a number of datasets relating to the election that use Twitter as their source of data. It is widely used in the literature, more so than any other social network, no doubt due to the ease of access to the data that other social networks don't have.

## 3.4 Data

### 3.4.1 Collection

The dataset that will be used for this project is the public dataset provided by Chen et al. in their 2021 paper [42]. This is a multilingual dataset of over 1.2 billion tweets that covers the time period from December 2019 until June 2021 and is described in Chen's paper as the first public Twitter dataset on the 2020 US presidential election. In order to capture the tweets, they followed specific user mentions and accounts "that were and are tied to the official and personal accounts of candidates who ran for president", as well as relevant keywords [42]. The full list of accounts and keywords that were monitored can be found in Appendices A and B respectively.

The dataset consists of text files that are separated by hour, so 24 files cover the monitored tweets of a day. Each file contains only tweet IDs as it is not permitted to share datasets containing tweets under Twitter's API terms and conditions. The tweet IDs can be queried via the Twitter API to retrieve the metadata (such as author, tweet content, etc) in a process that is referred to as hydration. This process is carried out by using the Twarc Python command line interface.

For the purposes of this study only the tweets within a week prior of the election will be studied. Therefore the tweets between the dates 27th October 2020 and 2nd November 2020 inclusive. This is a similar timeframe that has been used in some of the literature [16]. It also reflects the sort of timeframes that polling companies use. Many studies also consider the day before the elections to be the date to finish collecting data, a point that was made in Gayo-Avello's meta analysis [17].

The dataset used to train the models consists of 2500 tweets that are manually annotated: 500 for positive sentiment about Biden, 500 for negative sentiment about Biden, 500 for positive sentiment about Trump and 500 for negative sentiment about Trump. The remaining 500 form the neutral sentiment. The dataset of manually annotated tweets is split into an 80% training data and 20% test data ratio. The entire dataset that the best performing model will be applied across consists of 6,473,973 tweets.

### 3.4.2 Limitations

Chen et al. identified a number of limitations with their dataset. Some of these limitations are that it is heavily skewed towards the English language, so therefore those who post tweets in other languages such as Spanish are unlikely to have their tweets considered. Another limitation is that tweets that have been removed by the user (or that were removed due to the user being suspended or banned) cannot be returned by the API. This is a limitation that could have an effect on the project as Donald Trump was permanently banned from using Twitter's services on 8th January 2021 [43]. Additionally, Twitter banned more than 70,000 accounts that were linked to the QAnon conspiracy theory after the 6th January attack on the Capitol building [44]. The QAnon conspiracy theory has been linked largely to supporters of Donald Trump and therefore the removal of so many users could likely distort the support for Donald Trump. Finally, the streaming API that the dataset curators used only produces 1% of the tweets in real time, so this is not a dataset of all the tweets that were posted during that time period [42].

An issue that some research has come across is in trying to predict the outcome of individual states. The usual approach has been to use geographical coordinates or the location presented in a user's profile metadata. However, it has been shown that there are significant demographic variations between those who opt into geo services and those who geotag their tweets and that Twitter users who publish their geographical information are not representative of the wider Twitter demographic [45]. It is also not difficult for a malicious user to spoof their location (either in their profile or in their tweets) which would have a larger impact on the overall sentiment for a state. This leads to outcomes such as those found in the study conducted by Singh et al. that suggested that Arkansas had more positive Democratic sentiment than almost everywhere else in the US [11], when the reality was that in that election, Trump won the state with 62.4% of the vote, Biden receiving 34.8%. Another point which is that the American voters do not exist in their own microcosm, isolated from everybody else. Their views and opinions are formed not only by what they read and see from within the country, but also from outside the country. Americans living outside of the country (for example those posted overseas on military duty) are also allowed to vote so the views of those outside of the country have some impact and should also be considered.

What this means though is that it is arguably very difficult to predict the election on a state-by-state basis with any consistent level of accuracy, making it not very useful for any practical work. Judging the national sentiment is something that polling companies do already and is something that can be measured.

### 3.4.3 Preprocessing

Tweets from users are not posted with consideration for machine learning models. Therefore data processing is required in order to transform the unstructured data into data that can be used by a model.

- All tweet content will be converted to lowercase so that words that contain uppercase letters are treated as equal to their lowercase versions.
- All URLs will be removed. This does mean that the context around people commenting on news articles about the election will be lost, but it is standard practice in the literature to do this.
- All the characters '@' and '#' will be removed from tweets. Tweets use '@' at the beginning of a string to denote a username and '#' to denote a topic. The username is important to get context (for example someone sending "You are a terrible candidate" to either candidate) and the string after the # is also important to get context: for example '#terriblepresident' has a clear sentiment.
- Emojis and other non-alphanumeric characters will be removed. There is no clear consensus on how they should be used for sentiment analysis within the field of election prediction.
- Each tweet will be tokenised into words and any of the words that are in the stopword list will be removed. that is used from the NLTK package.

It should be noted that spelling mistakes made by the authors of the tweets will be present in the dataset however there is no intention to will not be corrected.

### 3.4.4 TFIDF Vectorisation

The tweets need to be vectorised in order to be processed by the model. This is done by converting the dataset into a dataset of vectors, where each tweet is converted into a vector that has the length equal to the length of the vocabulary. The value of each point in the vector is the TFIDF which is computed as follows:

- $tf$ represents the term frequency, which is the number of times that a term in the vocabulary appears in the tweet.
- $df_w$ represents the number of documents that contain that particular term. In this project, a document is a tweet in the dataset.
- $idf_{w,D} = log\frac{|D|}{df_w}$ where $|D|$ is the total number of tweets in the dataset.
- TFIDF $= tf \cdot idf_{w,D}$

## 3.5 Metrics

The classification models will be evaluated by their F1 score. The F1 score is the weighted average of the precision and recall of a model, with both aspects contributing an equal amount. The F1 score ranges from 0 to 1, where 1 is best. The equation for an F1 score can be given as:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.1}$$

This can also be written as:

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN} \tag{3.2}$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of False negatives. In the multi-class case, the F1 score is the unweighted average taken across all of the classes.

The majority of the literature uses either Mean Average Error (MAE) or $R^2$ to compare their models to the election outcome. There is no consensus within the literature on the best metric to use to measure the performance and so for the sake of completeness both the MAE and the $R^2$ metrics will be reported.

## 3.6 Programming Language

Python will be used to perform the sentiment analysis as it allows the use of libraries such as Pandas to read in data files, NLTK to preprocess data and Scikit-Learn to perform the machine learning models chosen. Twarc, which interfaces with the Twitter API to collect tweets from tweet IDs, requires Python to be installed in order to use.

## 3.7 Relevant Libraries

### 3.7.1 Twarc

Twarc is a tool that can be used both as a Python library and on the command line. It interacts with the Twitter API and allows for the archiving of Twitter JSON data. It also handles Twitter's API rate limits, sleeping for a period of time to allow for the rate limits to reset. As Twitter's developer policies only allow public datasets of Tweet ID numbers to be published, a tool such as Twarc is required in order to query the API for each tweet and collect the relevant JSON data.

### 3.7.2 NLTK

NLTK, the Natural Language Toolkit, is a library for Python that is used in Natural Language Processing (NLP) tasks such as tokenising, stemming, lemmatisation and tagging. It is used in this project to provide the list of stopwords that are removed from each tweet.

### 3.7.3 Scikit Learn

Scikit Learn is a Python library that allows users to preprocess data and apply various predictive models to that data, such as classification, regression and cluster analysis. It is used in this project to preprocess data, fit models to the training data and to predict the class over the full dataset. It is also used to calculate performance metrics for predictive models.

# Chapter 4

# Results

In this chapter, the results of the model comparisons are presented after being trained on the training data. This allows a decision to be made as to which model is the best performing model. The model is then applied to the full dataset and the results of this analysis is also presented.

## 4.1 Model Training Results

Each model was trained to to determine the best classifier for sentiment analysis towards Biden, and also towards Trump. The F1 scores for each model results can be seen in Table 4.1 and Figure 4.1. It is clear that the Linear Support Vector Classification (SVC) model is the best performing classifier for both Biden and Trump sentiment analysis, with F1 scores of 0.75 for Biden sentiment and 0.76 for Trump sentiment. The Multinomial Naive Bayes (MNB) model was marginally better than the Complement Naive Bayes (CNB) model at classifying Trump sentiment, however the CNB model was also marginally better at classifying Biden sentiment. The F1 scores are within 0.03 of eachother so they are arguably very similar in performance.

| **Model** | **Candidate Classifier** | **Accuracy** | **F1 Score** |
|---|---|---|---|
| MNB | Biden | 0.66 | 0.66 |
| | Trump | 0.69 | 0.68 |
| CNB | Biden | 0.69 | 0.69 |
| | Trump | 0.67 | 0.67 |
| GNB | Biden | 0.62 | 0.62 |
| | Trump | 0.54 | 0.53 |
| SVC (Linear) | Biden | 0.76 | 0.75 |
| | Trump | 0.76 | 0.76 |
| SVC (Poly) | Biden | 0.47 | 0.46 |
| | Trump | 0.49 | 0.48 |

Table 4.1: Accuracy and F1 Scores for both sentiment classifiers across tested models

The Gaussian Naive Bayes (GNB) model performed worse with an F1 score of 0.53 for Trump sentiment classification and 0.62 for Biden. While the Biden sentiment performed closely to the MNB model, the Trump classification was worse than the other models that have been discussed so far. Finally, the SVC model with a polynomial kernel performed the worst out of all of the tested models, with F1 scores of 0.46 for Biden and 0.48 for Trump, both performing lower than 0.5.

It should be noted that each model achieved accuracy scores very close to their respective F1 scores. As with the F1 scores, the Linear SVC model performed best out of the models and the Polynomial SVC model performed worst. The accuracy scores can be compared to state of the art models such as the Bidirectional Encoder Representations from Transformers (BERT) model or the Long short-term memory (LSTM) model.

Chandra and Singh trained both the BERT and LSTM models on tweets relating to the 2020 US Election and although they obtained accuracy scores between 85% and 88% for the models they attempted, the F1 scores they achieved with a batch size of 64 were really no different to that of the Linear SVC model that was trained in this project, being about 75-75%. In the case of the 128-batch model they trained on BERT, the F1 score was worse than was achieved via the Linear SVC model, achieving an F1 score of 70.2. The LSTM model they used achieved an F1 score of 68.6 [40]. They used a different dataset of tweets in their study so the two aren't directly comparable, however it is interesting to note that the F1 scores are not much different between the results obtained in this provject and the results obtained from the BERT model in another study.

As the Linear SVC model performed best for both Trump and Biden sentiment, it will be used as the sole model that will be used for the Trump sentiment classifier and the Biden sentiment classifier when analysing the entire dataset.

## 4.2 Full dataset classification results

The full dataset was classified twice - first to determine the sentiment towards Trump and then to classify the sentiment towards Biden. The results of this can be seen in Figure 4.2 and Table 4.2.

Generally, a large proportion of approximately 40% of the tweets were classified by the Biden and Trump sentiment classifiers as showing neutral sentiment towards their respective candidate. Essentially what this means is that within those tweets, no sentiment towards the candidate was shown, so for example, with the Trump sentiment classifier, a tweet that expresses sentiment towards Biden would be treated as neutral, as it shows no sentiment towards Trump. That tweet would however be classified as positive or negative, depending on the content of the tweet, for the Biden classifier.

Another general comment that can be made is that the negative sentiment towards each candidate is stronger than the positive sentiment, with 32.741% of the tweets being considered negative towards Biden, and 40.247% of the tweets being considered negative towards Trump. This is in direct contrast to the positive sentiment, where 24.108% of the sentiment was

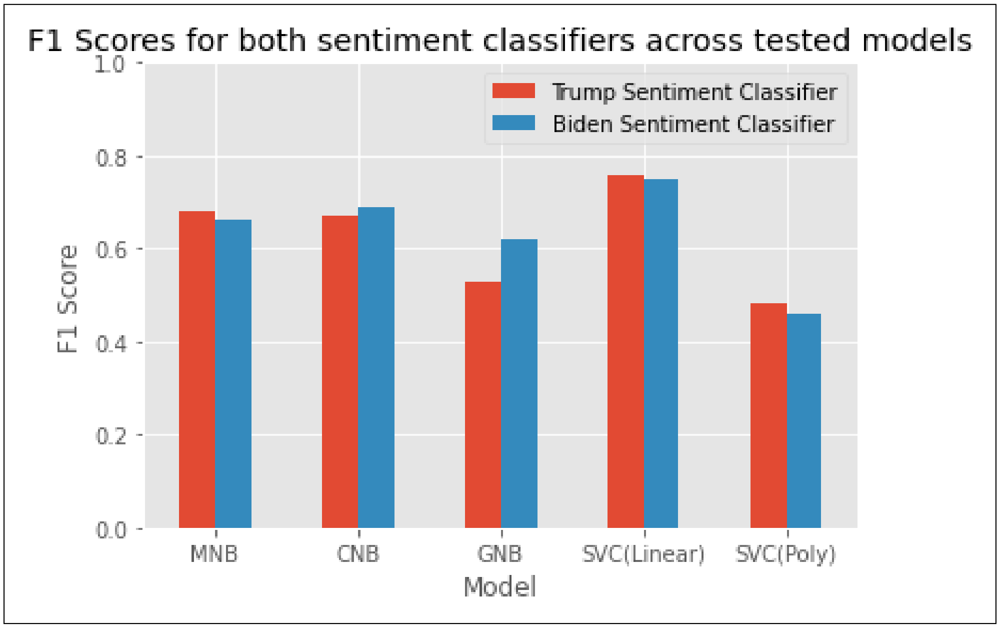

Figure 4.1: Bar plot presenting F1 scores for each model tested for each candidate sentiment classification.

| | Sentiment | | |
|---|---|---|---|
| Candidate | Positive (%) | Neutral (%) | Negative (%) |
| Biden | 24.108 | 43.151 | 32.741 |
| Trump | 20.691 | 39.062 | 40.247 |

Table 4.2: The proportion of tweets of the full dataset that were classified to each label by the Linear SVC model.

positive towards Biden and 20.691% was positive towards Trump.

If the ratio is taken of negative sentiment to positive sentiment, then Biden has 1.36 times more negative sentiment than positive sentiment, whereas Trump has close to twice as much negative sentiment as he does positive, at 1.95.

By looking at the proportion of the entire positive sentiment that a candidate received of the total positive sentiment, it is found that support (or enthusiasm) for Biden is at 53.81% and support for Trump is at 46.18%. Comparing this to the election outcome requires the results to be modified to only apply to two candidates as there were more than two candidates in the 2020 election. Therefore the votes of Biden and Trump are summed and the proportion of those summed totals are calculated. The modified 2-candidate percentage outcome for the election was 47.73% for Trump and 52.27% for Biden. Comparing the model outcome to the actual outcome presents an $R^2$ value of 0.536 and a Mean Average Error of 1.545.

It is also useful to compare these results to the polling averages taken around the same time. These polling averages are conducted by the political polling websites FiveThirtyEight, RealClearPolitics and 270ToWin, who aggregate political polls and produce an election

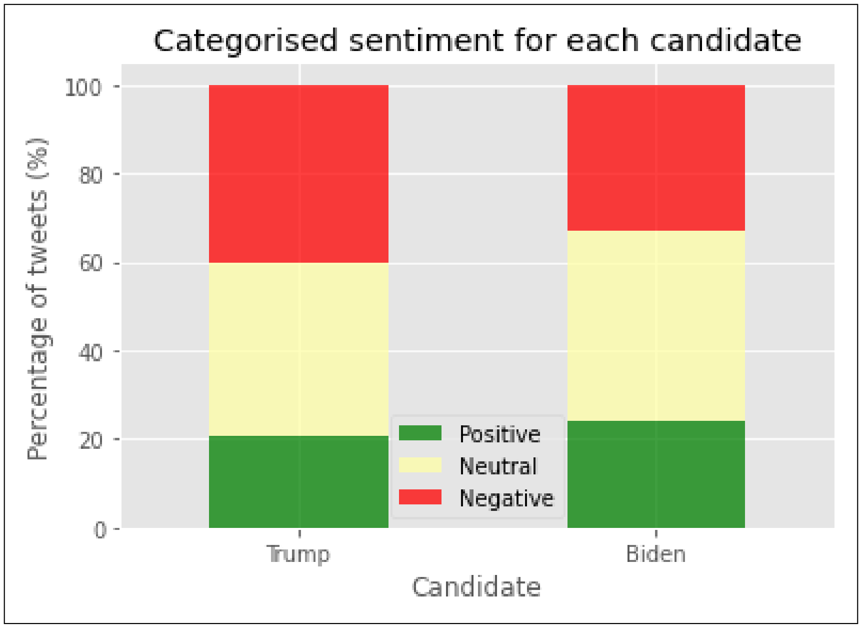

Figure 4.2: Stacked bar plot that shows the proportion of tweets of the full dataset that were classified by the Linear SVC model.

forecast from their statistical analysis. Their election predictions are not just of the percentage outcome but also the winners of each state. Their predictions are presented alongside the election outcome in Table 4.3.

Although the model developed in this project has a lower MAE, its $R^2$ value is also smaller. This would suggest that it is a worse performer than the polling averages, however the model developed in this project only considers the percentage between Trump and Biden whereas the polling averages consider other candidates too. This would explain their higher MAE as they have three numbers to consider. If the model developed in this project also considered the minor candidates in the election it is likely that the $R^2$ value could be higher.

| | Trump (%) | Biden (%) | Other | $R^2$ | MAE |
|---|---|---|---|---|---|
| **2020 Results (2 candidates)** | **47.73** | **52.27** | **-** | **-** | **-** |
| Model results | 46.18 | 53.81 | - | 0.536 | 1.545 |
| **2020 Results (All candidates)** | **46.9** | **51.3** | **1.8** | **-** | **-** |
| FiveThirtyEight Poll Average | 43.4 | 51.8 | 4.8 | 0.986 | 2.333 |
| Real Clear Politics Poll Average | 44.0 | 51.2 | 4.8 | 0.988 | 2.000 |
| 270ToWin Poll Average | 43.1 | 51.1 | 5.8 | 0.980 | 2.700 |

Table 4.3: Comparison of model results with election results and polling aggregation averages

# Chapter 5

# Conclusion and Future Research

In this chapter, conclusions are drawn on the results of the project and caveats are set about them. Afterwards, consideration is given to future research that could be conducted in order to improve results and push the field further.

## 5.1 Conclusion

Social media websites such as Twitter are increasingly being used by politicians and the public to discuss politics and engage with political campaigns. Sentiment analysis of these posts allows politicians and their campaigns to find the general sentiment towards them by the users of social networks.

In this project it was established that a linear Support Vector Classifier (SVC) performed better than a polynomial SVC for sentiment analysis. The linear SVC also performed better than all of the Naive Bayes models tested, with an F1 score that was close to 0.8.

Twitter users posted more negative comments about the candidates than they did positive comments in the week prior to the election, with both Biden and Trump receiving more negative comments than they did positive. In Trump's case, there were twice as many negative tweets as there were positive tweets. Proportionally the percentage of positive tweets to each candidate reflected the percentage outcome of the electorate. The negative sentiment also closely reflected the outcome, if anti-Biden tweets are assumed to have been made by Trump supporters, and vice versa.

While the sentiment towards each candidate closely resembles the outcome of the election, it is not possible to determine if this is a coincidence or not. There are a number of studies in this field where a methodology that has worked for one election has not been repeatable and has not generalised to elections held in the same country under the same system, never mind elections held in a different country and different electoral systems. As this project does not attempt to predict another election with the same methodology, it is not known if this methodology can be generalised.

In the UK and the USA for example, their electoral system does not require a party to receive a national percentage majority in order to win the election, as their system is based on parliamentary seats (UK) and Electoral College votes (US). This means a percentage outcome is not necessarily going to determine the winner of an election. That being said, by using previous election results, psephologists are able to predict who wins seats based on their national percentage share.

While the results are interesting, they should also be treated with much caution. There is a significant bias in using Twitter data and Twitter's user demographics do not reflect the demographics of the USA. There is also a self selection bias in that the dataset contains only the tweets of those who have chosen to tweet about the election. Finally, the dataset itself is biased as the Twitter API does not allow the rehydration of tweets made by users who were banned meaning that a significant number of tweets that are highly likely to learn towards supporting Trump are not present in this dataset. In addition, none of Trump's tweets are also included in the dataset due to his permanent ban from Twitter.

Sentiment analysis of social media posts provides a useful indicator of sentiment (or enthusiasm) that appears to provide similar results to traditional polling. While it was able to seemingly predict the proportion of votes in the election, caution is advised due to issues with the dataset. Granular prediction of individual states relies upon the use of geolocation metadata which many users don't use and therefore if the decision is made to only use tweets with geolocations, there is a risk of losing so many tweets that it would have an impact on the analysis. The translation of a national percentage share to seats or states would need to be done because social media posts are not able to provide consistent or accurate results in all states (especially sparsely populated states) on a granular level.

In conclusion, the Linear SVC model was the best performing classification model out of all of the models tested. The sentiment analysis of the dataset appears to model the outcome of the election, with some stated caveats. It is not clear whether the methodology is one that can be repeated across different US elections or different electoral systems however it is a methodology that can easily be scaled up to account for multi-candidate elections.

## 5.2 Future Research

The field of predicting election results carrying out sentiment analysis with social media data is an interesting field that is still in its infancy. Within the field there is work to do regarding standardising metrics and methodologies and generalising a methodology that can work with different electoral systems and processes.

Due to technical difficulties and time constraints, it has not been possible to test any deep learning models in this project. As a state of the art model, further work should consider the use of BERT to conduct the same analysis and compare the results with the performance of the other models.

Additional further work that should be considered is investigating the use of emojis and URLS in sentiment analysis. Emojis are visual representations of some Unicode characters that are used by social media users. There are 3521 emojis in the Unicode standard as of September 2020 and they represent flags, animals, ”emoticons” amongst other things [46].

Some of these emojis can represent sentiment, for example the heart emoji can represent love, whereas an emoji with an angry face can be used to represent anger.

Social networks allow links to be embedded within posts and some posts may for example be a reaction a news article that is in a URL. As a URL is removed from the dataset, the context of that is lost and so it would be interesting to consider whether including the metadata about a linked page (such as the article headline) into the sentiment analysis (along with the emoji as mentioned earlier) could improve classification.

Further work can be completed on this particular project by using a larger training dataset, a few thousand for each particular training data (positive-negative for both Trump and Biden as well as a larger neutral set) should make the results of the classification more accurate. In addition to this, it would be preferable to use a separate neutral dataset for both the Trump and the Biden training data. This would help ensure that a statement like ”I hate Trump” in a Biden sentiment classifier would result as a neutral sentiment towards Biden, as it neither expresses a positive or a negative sentiment about him.

In addition, while US politics is dominated by the Democratic and Republican parties, the Libertarian and the Green parties also do receive notable amount of votes and it would be interesting to see whether the addition of those two parties would make the model more or less accurate. Finally an area of further work would be to apply the same methodology to other elections to ascertain if it can be generalised to predict percentage outcomes.

# Bibliography

[1] D. Desilver, "Mail-in voting became much more common in 2020 primaries as covid-19 spread," *Pew Research Center*, Oct. 2020.

[2] G. Graziosi, L. James, and L. Hall, "Biden breaks record for most votes in history for any presidential candidate," *The Independent*, Nov. 2020.

[3] J. M. Grumbach, "From backwaters to major policymakers: Policy polarization in the states, 1970–2014," *Perspectives on Politics*, vol. 16, no. 2, p. 416–435, 2018.

[4] M. Dimock, "America is exceptional in the nature of its political divide," *Pew Research Center*, Nov. 2020.

[5] B. Liu, *Sentiment analysis and opinion mining.* Morgan & Claypool, 2012.

[6] B. Liu and L. Zhang, *A Survey of Opinion Mining and Sentiment Analysis*, pp. 415–463. Boston, MA: Springer US, 2012.

[7] S. Aghaei, M. A. Nematbakhsh, and H. K. Farsani, "Evolution of the world wide web : From web 1.0 to web 4.0," *International journal of Web & Semantic Technology*, vol. 3, pp. 1–10, Jan. 2012.

[8] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market.," *Journal of Computational Science*, vol. 2, pp. 1–8, March 2011.

[9] R. P. Schumaker, A. T. Jarmoszko, and C. S. Labedz, "Predicting wins and spread in the premier league using a sentiment analysis of twitter," *Decision Support Systems*, vol. 88, pp. 76–84, August 2016.

[10] A. Jungherr, "Twitter use in election campaigns: A systematic literature review," *Journal of Information Technology & Politics*, vol. 13, pp. 72–91, Dec. 2015.

[11] A. Singh, A. kumar, N. Dua, V. K. Mishra, D. Singh, and A. Agrawal, "Predicting elections results using social media activity a case study: USA presidential election 2020," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2021.

[12] F. Franch, "(wisdom of the crowds): 2010 UK election prediction with social media," *Journal of Information Technology & Politics*, vol. 10, pp. 57–71, Jan. 2013.

[13] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, "140 characters to victory?: Using twitter to predict the UK 2015 general election," *Electoral Studies*, vol. 41, pp. 230–233, Mar. 2016.

[14] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, pp. 178–185, May 2010.

[15] W. Budiharto and M. Meiliana, "Prediction and analysis of indonesia presidential election from twitter using sentiment analysis," *Journal of Big Data*, vol. 5, Dec. 2018.

[16] D. Gayo-Avello, P. Metaxas, and E. Mustafaraj, "Limits of electoral predictions using twitter," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, Jan 2011.

[17] D. Gayo-Avello, "A meta-analysis of state-of-the-art electoral prediction from twitter data," *Social Science Computer Review*, vol. 31, p. 649–679, Aug 2013.

[18] M. M. Skoric, J. Liu, and K. Jaidka, "Electoral and public opinion forecasts with social media data: A meta-analysis," *Information*, vol. 11, p. 187, Mar. 2020.

[19] A. Khan, H. Zhang, N. Boudjellal, A. Ahmad, J. Shang, L. Dai, and B. Hayat, "Election prediction on twitter: A systematic mapping study," *Complexity*, vol. 2021, p. 1–27, Apr 2021.

[20] R. Liu, X. Yao, C. Guo, and X. Wei, "Can we forecast presidential election using twitter data? an integrative modelling approach," *Annals of GIS*, vol. 27, no. 1, pp. 43–56, 2021.

[21] P. Metaxas and E. Mustafaraj, "From obscurity to prominence in minutes: Political speech and real-time search," *WebSci10: Extending the Frontiers of Society On-Line*, jan 2010.

[22] P. T. Metaxas and E. Mustafaraj, "The fake news spreading plague," in *Proceedings of the 2017 ACM on Web Science Conference*, ACM, June 2017.

[23] V. Ho, "Goats, cows and devin nunes' mom: how a republican's twitter lawsuit backfired," *The Guardian*, Mar. 2019.

[24] A. Mislove, S. Jørgensen, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist, "Understanding the demographics of twitter users," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 554–557, AAAI Press, 2011.

[25] S. Wojkik and A. Hughes, "Sizing up twitter users," *Pew Research Center*, Apr. 2019.

[26] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas, "Vocal minority versus silent majority: Discovering the opionions of the long tail," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 103–110, 2011.

[27] F. Miró-Llinares and J. C. Aguerri, "Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat'," *European Journal of Criminology*, p. 147737082199405, Apr. 2021.

[28] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, pp. 211–236, May 2017.

[29] A. Bovet and H. A. Makse, "Influence of fake news in twitter during the 2016 US presidential election," *Nature Communications*, vol. 10, Jan. 2019.

[30] C. Dewey, "Facebook fake-news writer: 'i think donald trump is in the white house because of me'," *Washington Post*, Nov. 2016.

[31] S. Subramanian, "Inside the macedonian fake-news complex," *Wired*, Feb. 2015.

[32] J. C. Wong and S. Levine, "Twitter labels trump's false claims with warning for first time," *The Guardian*, May 2020.

[33] S. Tasnim, M. M. Hossain, and H. Mazumder, "Impact of rumors or misinformation on coronavirus disease (covid-19) in social media," *Journal of Preventive Medicine & Public Health*, p. 171–174, May 2020.

[34] A. Bessi and E. Ferrara, "Social bots distort the 2016 u.s. presidential election online discussion," *First Monday*, Nov. 2016.

[35] H. Zhang, "The optimality of naive bayes," in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, vol. 2, 01 2004.

[36] A. Géron, *Hands-On Machine Learning with Scikit-Learn & Tensorflow : concepts, tools, and techniques to build intelligent systems*, p. 145. Sebastopol, CA: O'Reilly Media, 2017.

[37] M. Sassano, "Virtual examples for text classification with support vector machines," in *Proceedings of the 2003 conference on Empirical methods in natural language processing* -, Association for Computational Linguistics, 2003.

[38] K. P. Murphy, *Machine learning [electronic resource] : a probabilistic perspective.* Adaptive computation and machine learning series, Cambridge, MA: MIT Press, 2012.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[40] R. Chandra and R. Saini, "Biden vs trump: Modelling us general elections using bert language model," *IEEE Access*, pp. 1–1, 2021.

[41] G. E. Marcus and M. B. MacKuen, "Anxiety, enthusiasm, and the vote: The emotional underpinnings of learning and involvement during presidential campaigns," *American Political Science Review*, vol. 87, no. 3, pp. 672–685, 1993.

[42] E. Chen, A. Deb, and E. Ferrara, "#election2020: the first public twitter dataset on the 2020 US presidential election," *Journal of Computational Social Science*, Apr. 2021.

[43] Twitter, "Permanent suspension of @realdonaldtrump," Jan 2021. Last accessed 10 September 2021.

[44] Guardian, "Twitter suspends 70,000 accounts sharing qanon content," Jan 2021. Last accessed 10 September 2021.

[45] L. Sloan and J. Morgan, "Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter," *PLOS ONE*, vol. 10, p. e0142209, Nov. 2015.

[46] Emojipedia, "Emojipedia faq." Last accessed 10 September 2021.

# Appendices

# Appendix A

# Dataset Accounts and Mentions

Note that the dates are using the American format of month-day-year.

Table A.1: A full list of the mentions and accounts that were actively tracked in the dataset.

| Beginning of Table | | | |
|---|---|---|---|
| Mentions | Started tracking | Stopped | Restarted |
| @realDonaldTrump | 05/20/19 | | |
| @GovBillWeld | 05/20/19 | | |
| @MarkSanford | 05/20/19 | 11/14/19 | 09/25/20 |
| @WalshFreedom | 05/20/19 | | |
| @MichaelBennet | 05/20/19 | | |
| @JoeBiden | 05/20/19 | | |
| @CoryBooker | 05/20/19 | 01/13/20 | 09/25/20 |
| @GovernorBullock | 05/20/19 | 12/02/19 | 09/25/20 |
| @PeteButtigieg | 05/20/19 | | |
| @JulianCastro | 05/20/19 | 01/02/20 | 09/25/20 |
| @BilldeBlasio | 05/20/19 | 11/14/19 | 09/25/20 |
| @JohnDelaney | 05/20/19 | | |
| @TulsiGabbard | 05/20/19 | | |
| @gillbrandny | 05/20/19 | 11/14/19 | 06/20/20 |
| @KamalaHarris | 05/20/19 | 11/14/19 | 06/20/20 |
| @SenKamalaHarris | 05/20/19 | 11/14/19 | 06/20/20 |
| @Hickenlooper | 05/20/19 | 11/14/19 | 09/25/20 |
| @JayInslee | 05/20/19 | 11/14/19 | 09/25/20 |
| @amyklobuchar | 05/20/19 | | |
| @SenAmyKlobuchar | 05/20/19 | 03/03/20 | 06/20/20 |
| @WayneMessam | 05/20/19 | 12/02/19 | 09/25/20 |
| @sethmoulton | 05/20/19 | 11/14/19 | 09/25/20 |
| @BetoORourke | 05/20/19 | 11/14/19 | 09/25/20 |

Continuation of Table A.1

| Mentions | Started tracking | Stopped | Restarted |
|---|---|---|---|
| @TimRyan | 05/20/19 | 11/14/19 | 09/25/20 |
| @BernieSanders | 05/20/19 | | |
| @ericswalwell | 05/20/19 | 11/14/19 | 09/25/20 |
| @ewarren | 05/20/19 | | |
| @SenWarren | 06/20/20 | | |
| @marwilliamson | 05/20/19 | | |
| @AndrewYang | 05/20/19 | | |
| @JoeSestak | 05/20/19 | 12/02/19 | 09/25/20 |
| @MikeGravel | 05/20/19 | 08/06/19 | 09/25/20 |
| @TomSteyer | 05/20/19 | | |
| @DevalPatrick | 05/20/19 | | |
| @MikeBloomberg | 05/20/19 | | |
| @staceyabrams | 06/20/20 | | |
| @SenDuckworth | 06/20/20 | | |
| @TammyforIL | 06/20/20 | | |
| @KeishaBottoms | 06/20/20 | | |
| @RepValDemings | 06/20/20 | | |
| @val_demings | 06/20/20 | | |
| @AmbassadorRice | 06/20/20 | | |
| @GovMLG | 06/20/20 | | |
| @Michelle4NM | 06/20/20 | | |
| @SenatorBaldwin | 06/20/20 | | |
| @tammybaldwin | 06/20/20 | | |
| @KarenBassTweets | 06/20/20 | | |
| @RepKarenBass | 06/20/20 | | |
| @Maggie_Hassan | 06/20/20 | | |
| @SenatorHassan | 06/20/20 | | |
| @GovRaimondo | 06/20/20 | | |
| @GinaRaimondo | 06/20/20 | | |
| @GovWhitmer | 06/20/20 | | |
| @gretchenwhitmer | 06/20/20 | | |

End of Table

# Appendix B

# Dataset keywords

Note that the dates are using the American format of month-day-year.

Table B.1: A full list of the keywords that were actively tracked in the dataset.

| Beginning of Table | |
|---|---|
| Keywords | Tracked since |
| ballot | 06/20/20 |
| mailin | 06/20/20 |
| mail-in | 06/20/20 |
| mail in | 06/20/20 |
| donaldtrump | 09/12/20 |
| donaldjtrump | 09/12/20 |
| donald j trump | 09/12/20 |
| donald trump | 09/12/20 |
| don trump | 09/12/20 |
| joe biden | 09/12/20 |
| joebiden | 09/12/20 |
| biden | 09/12/20 |
| mike pence | 09/12/20 |
| michael pence | 09/12/20 |
| mikepence | 09/12/20 |
| michaelpence | 09/12/20 |
| kamala harris | 09/12/20 |
| kamala | 09/12/20 |
| kamalaharris | 09/12/20 |
| trump | 09/13/20 |
| #DonaldTrump | 09/13/20 |
| PresidentTrump | 09/13/20 |
| MAGA | 09/13/20 |

Continuation of Table B.1

| Keywords | Tracked since |
|---|---|
| trump2020 | 09/13/20 |
| Sleepy Joe | 09/13/20 |
| Sleepyjoe | 09/13/20 |
| HidenBiden | 09/13/20 |
| CreepyJoeBiden | 09/13/20 |
| NeverBiden | 09/13/20 |
| BidenUkraineScandal | 09/13/20 |
| DumpTrump | 09/13/20 |
| NeverTrump | 09/13/20 |
| VoteRed | 09/13/20 |
| VoteBlue | 09/13/20 |
| RussiaHoax | 09/13/20 |
| presidential debate | 09/28/20 |
| #debates | 09/28/20 |
| presidentialdebate | 09/28/20 |
| electoral vote | 09/28/20 |
| debates2020 | 09/28/20 |
| elections2020 | 09/30/20 |
| ivoted | 09/30/20 |
| #vote | 09/30/20 |
| vpdebate | 10/06/20 |
| vp debate | 10/06/20 |
| sen. harris | 10/07/20 |
| sen harris | 10/07/20 |
| mr. vice president | 10/07/20 |
| debate2020 | 10/22/20 |
| presidentialdebate2020 | 10/22/20 |
| debatetonight | 10/22/20 |
| harris | 11/02/20 |
| bidenharris | 11/02/20 |
| electionday | 11/02/20 |
| electionnight | 11/03/20 |
| election day | 11/03/20 |
| election night | 11/03/20 |
| electionresult | 11/03/20 |
| election result | 11/03/20 |
| decision2020 | 11/04/20 |
| countallthevotes | 11/04/20 |
| counteveryvote | 11/04/20 |

Continuation of Table B.1

| Keywords | Tracked since |
|---|---:|
| postelection | 11/04/20 |
| electoral fraud | 11/04/20 |
| voter fraud | 11/04/20 |
| electoralfraud | 11/04/20 |
| voterfraud | 11/04/20 |
| exitpoll | 11/04/20 |
| exit poll | 11/04/20 |
| sharpiehoax | 11/04/20 |
| stopthecount | 11/04/20 |
| stop the count | 11/04/20 |
| countthevote | 11/04/20 |
| count the vote | 11/04/20 |
| count every vote | 11/05/20 |
| racecall | 11/06/20 |
| race call | 11/06/20 |
| 46th president | 11/07/20 |
| president-elect | 11/07/20 |
| president elect | 11/07/20 |
| presidentelect | 11/07/20 |
| presidentbiden | 11/07/20 |
| president biden | 11/07/20 |
| vpharris | 11/07/20 |
| vp harris | 11/07/20 |
| vpelect | 11/07/20 |
| vicepresidentelect | 11/07/20 |
| vicepresident-elect | 11/07/20 |
| vice president harris | 11/07/20 |
| vicepresident harris | 11/07/20 |
| vice-president harris | 11/07/20 |
| vice president-elect | 11/07/20 |
| FLOTUS | 11/07/20 |
| POTUS | 11/07/20 |
| first lady | 11/07/20 |
| second gentleman | 11/07/20 |
| vice president | 11/07/20 |
| 46thpresident | 11/07/20 |
| joe and kamala | 11/07/20 |
| kamala and joe | 11/07/20 |
| @transition46 | 11/07/20 |

Continuation of Table B.1

| Keywords | Tracked since |
|---|---|
| biden-harris | 11/08/20 |
| buildbackbetter | 11/08/20 |
| build back better | 11/08/20 |
| biden cabinet | 11/09/20 |
| stop the steal | 11/09/20 |
| stopthesteal | 11/09/20 |
| bidentransition | 11/09/20 |
| biden transition | 11/09/20 |
| trumpconcede | 11/09/20 |
| trump concede | 11/09/20 |
| dr. biden | 11/10/20 |
| drbiden | 11/10/20 |
| dr.biden | 11/10/20 |
| transitiontobiden | 11/23/20 |
| transition to biden | 11/23/20 |

End of Table