Faculty of Engineering & Information Technology
School of Computer Science

<u>42913 Social and Information Network Analysis</u>
Autumn 2020

Assignment 3

Andres Felipe Lagos, Student ID: 13092248
Carlos Mario Carvajal Moreno, Student ID: 13144148
Ernest Ilustre, Student ID: 12763239

Subject Coordinator:
Prof Ying Zhang

Abstract

The following study presents an analysis approach to predict the next US presidential elections, taking place in November 2020. This report presents a social network analysis by using tweeter data from the main two candidates, the actual president Donald John Trump and the former vice president Joseph Robinette Biden Jr. In addition, it has been considered the real-time tweet data of the general population of decisive states, which are commonly known as swing states and have been previously identified as states that determine the elections outcome. The methodology presented compiles the analysis of the sentiment captured in the tweets from tweets collected from the U.S. and the swing's states, and an analysis of the tweets and reactions obtained from the twitter account of the candidates. Moreover, regular expression methods were used to identify the candidate that every tweet was referring to. For this, this study proposes an assemble method of two Natural Language Processing algorithms for sentiment analysis, by averaging the sentiment score. Lastly, a comparison and integrations of the two components will be used to produce a prediction and create a discussion around the relevance of this analysis and the accuracy of the results observed.

1. Introduction

U.S. presidential elections have always been a hot topic, that drives the interest of many people and organisations around the world, because of the role this country has in the global economy and in the political configuration. It is also very common to observe interest from other countries as the outcome of this election influences the policy and relationship that the U.S has with them. This includes commercial and trade relations, as well as support relations where the U.S is presents itself as a funder or financial supporter. The current president, Donald J. Trump, has been very controversial, because his particular and different way to address certain issues; from the moment he announced his candidacy throughout the course of his presidential term. In addition, according to the different polls, analyses and predictions for the 2016 election the candidate Hillary Clinton was declared the winner by a large merging. For instance, Borchers (2016) highlighted predictions by popular media outlets in the U.S. at the time, which claimed that Hillary had a chance between 71% and 85% of beating Donald Trump. These wrongly predicted outcomes have increased the interest of many academics and organisations to apply different approaches for the upcoming election, because the lack of certainty have exacerbated.

Over the last few months several issues have influenced the public opinion in the U.S. regarding Trump's leadership. Early this year, Trump was impeached by the House of Representatives, which caused a negative impact for his candidacy; but then acquitted by the U.S. Senate. Subsequently, he faced the COVID-19 crisis, having America becoming the epicentre of the pandemic, with the highest rate of new cases and deaths. Several people have questioned the way he has managed this situation, such as the opposition, the Democratic party, and other public figures, which have had a degree of influence over his supporters and detractors. Furthermore, over the last couple of weeks, social unrest has been observed across the country for the death of a person from the black community by the police force, which has also caused the emergence of opinions in social media platforms around his actions. On the other hand, Trump's main opponent, Joe Biden, has addressed publicly the way he would handle these situations if he was to become the 46th president of the United States. All these dynamics have influenced the public opinion in different ways and degrees, with consequent reactions being reflected and expressed in social media platforms. Thus, this study considers that the appropriate way to assess the probabilities of the candidates of winning must come from the study of the sentiments and traffic manifested in these platforms.

Social network analysis is a tool that addresses the aim of this report. By studying the data originated from social media platforms like twitter, it is possible to perform an evaluation of the sentiment over the mentioned issues and dynamics, and also to shape the process of building a prediction of the outcome on the upcoming presidential election. For this reason, Twitter has been chosen as the platform where data is collected and analysed, as its main purpose is to provide a virtual space where people can express their ideas through comments. This, this data enables the analysis of what the stakeholders, subject of this study, are thinking and the relation they have with users and degree of influence. Therefore, the following tools will be used to extract the data, analyse it and present the results in a visual way to predict the winner: Python and the libraries Tweepy, Twython, Twitter API and PowerBI.

Hence, the following report will first present relevant applications developed by academics that have addressed the 2016 presidential election outcome and predictions. Secondly, it will describe the methodology applied, such as the extraction of the data and the description of its nature, as well as the steps taken to process and analyse the twitter data. Thirdly, the experiment settings and the analysis of the results are presented, along with the respective discussion. Finally, the report will close with the integration of the different results and an overall prediction of the 2020 presidential elections. It is important to highlight that the analysis and prediction is based on today's data, with the assumption that it reflects the voter's mind today and not in November. Thus, it is envisaged that there will be many other affairs that will arise between now and November (Date for the electoral poll), that will have an impact on the future voter's decisions. Therefore, the analysis and conclusions presented in this study may not be representative then and a new study with updated data will be required.

2. Related work

Oikonomou & Tjortjis (2018) developed an approach by extracting and analysing data from the swing states. They performed a sentiment analysis for every state and compared the results with the actual outcome of the presidential election in 2016. From the comparisons, the authors were able to conclude that their predictions matched the actual results in the key states, indicating the potential scalability of their work. Therefore, some of their methodologies will be used for this study, such as the evaluation of data on the key states, because of their highly relevant and assertive course of action.

In addition, the methodology applied by Chouhbi (2020) on Clinton and Trump tweets has been rolled out in this work, to Trump and Biden tweets. The sentiment analysis performed over the tweets of the two candidates, along with the analysis of the analysis of followers, will provide a picture that will be cross-examined with the country in general and the swing states results. This will increase the scope of Oikonomou & Tjortjis (2018) and will ensure that the predictions presented in this report are the representation of two important perspectives. Furthermore, the authors introduced a Naïve Bayes methodology along with a sentiment analysis provided by a Natural Language Processing pre-trained model, that demonstrated accurate results in comparison to the actual results of the 2016 election. Therefore, part of this methodology will be considered and applied in this study.

Other relevant methodology like the one proposed by Yaqub et al. (2017), were also considered for the construction of the methodology presented here. The data collection by the authors through the stream Tweeter API will be used in this report, along with some of the considerations and assumptions. These ones include few data cleaning approaches, as it is expected that the streaming will capture bots and promotional and irrelevant accounts; a sentiment analysis on candidates tweet, which presents an alternative approach to Chouhbi (2020); and an analysis of the sentiment of citizens to the words Trump and Biden, which complements the first outlined approach.

3. Methodology

The twitter API was used to stream, collect and analyse tweets made in real-time. First, developer accounts with Twitter was obtained by the authors of this report, in order to obtain the credentials to develop a python script that is able to stream and search for tweets. Second, several queries were created to send twitter the requests for specific information; this, through filtering and specific commands given to a developed application for the information required for this study. The Twitter API documentation (2020) was followed to construct the application that correctly retrieve the tweets that contained relevant information for the U.S. presidential elections. Subsequently, the Stream and Search API were used according to the tweet's nature and owner that was looked for; while for historical tweets the Search API was more appropriate, for real-time tweets the Stream API was used. Both APIs returned a JSON format that was used for the construction of a data-frame with the library Pandas, which served as a tool for data manipulation and analysis. In addition, each one of the APIs presents a group of limitations. While the Search API allows the filtering of relevant information, it is limited in terms of the number of tweets that can be extracted. On the other hand, the Stream API lacks some of the filtering features, but allows the streaming of a considerably greater number of tweets, than Search. Subsequently, the collected data is cleaned and transformed to a data frame format and applied two Natural Language Processing tools for sentiment analysis. The results are then analysed and visualised in Microsoft PowerBI.

Assumptions:

When making the decision of choosing between the Search and Stream APIs, it was taken in consideration the relevance of the latest events in the United States. It is assumed that these ones have had an effect on voters' minds, which may be different to what it was 6 months, 3 months and even a month ago. In addition, those with unchanged set of minds and unmodified political tuning are assumed to still express their affiliation, and even more in polarising times. Therefore, it has been considered that a historical analysis or streaming of past tweets does not respond to the aim of this study and has been discarded; only the latest tweets of the past week have been considered. Hence, the Stream API was used for tweets created by US citizens, while the Search API was applied only to the candidates Trump and Biden, because of their historical relevance.

As indicated by Oikonomou & Tjortjis (2018), the States of Florida, Ohio and North Carolina have been swinging states that do not have a fixed political affiliation and change over time. These states have been declared decisive states by the authors and were given greater importance in the different analysis in this report. Moreover, despite the weight of these states in the elections, it has been considered that the latest events have caused a major impact in other states; and hence have also been included in the analysis, but with less assigned weight.

In respect to the candidates' tweets, it has been considered that the past 6 months will provide good information to be included in the analysis. For this reason, the Search API was used to extract the candidates tweets over the past 3 months, that were joined with the existing datasets created by Reese (2020) and by Rao (2020). This, was done because it has been assumed that the confidence of the candidates and its mind change overtime, represents a great value for the purposes of this study. This, as changes can be perceived by their followers and voters. Moreover, existing datasets were found with data up to April 2020, which were complemented with up to date streamed data from this study

Furthermore, it has been considered that even though retweets represent a certain level of endorsement, they do not follow or correspond entirely to the mind of the publisher (user performing a retweet) and in many occasions can cause a great level of noise. For this reason, for the streamed tweets from the Stream API, have been discarded and only the number of followers has been considered under a specific criterion. In terms of the candidates' tweets, obtained by the Search API, their retweets were considered for analysis, as it is less likely that a retweet performed by them has an inaccurate representation of their mind.

For a more accurate prediction, this study will need to be performed again close to the date of the election in November. This because it is envisaged that many more events that will change voters' minds will happen between now and November.

Data collection and query:

Tweets were streamed and collected in real-time from June 4th, 2020 to June 12th, 2020, achieving a dataset with of over 800 thousand tweets. This, presents an improvement in numbers of tweets collected and analysed by comparison on the set collected by Oikonomou & Tjortjis (2018). Although, the authors collected tweets over a period of 6 months prior the 2016 election date, they only achieved 277 thousand tweets. On the other hand, in this study the collected tweets were trimmed down to 432 thousand unique tweets of over 110 thousand unique users presenting an improvement in data gathering and processing. This aspect is very important as the U.S. population is 328 million, so the sample data must be big enough to be representative, but repeated users or bots must be excluded. It is envisaged that the continuity of this work until November 2020, will yield a greater accuracy in the prediction of the next president of the United States.

The data mined from tweeter was streamed and collected with several filters that support the aim of this study. First, key words were introduced in the Stream filter that contain both candidates' names and important words that have been representative for their campaigns, such as "Make America Great Again" or "Never Trump". These words are present in both text of the tweet as well as hashtags. In addition, to reduce bias in the collection, both negative and positive key words were balanced between the two candidates and are presented as follow:

- Trump
- Biden
- Maga
- Makeamericagreatagain
- Nevertrump
- Trump2020
- Biden2020
- Joe Biden
- 2020 elections
- Trump president
- Never Biden
- Biden president

As mentioned previously, the states Florida, Ohio and North Carolina were given a higher weight for the analysis, while retaining the rest of states and information from US citizens in other States. This filtering and weight assignment were not performed in the Streaming process, due to the limitations of the Stream API. Instead, these processes were applied directly to the columns in the Pandas data frame. Additionally, during streaming the language was filtered to only English tweets, as it is the national language of United States and for the simplicity of the analysis. It is important to highlight that United States is a multicultural nation with several different languages spoken, but the inclusion of different languages in the streaming can add noise from tweets written by non-American voters and introduce bias in the study.

For Trump and Biden tweets, their twitter id was utilised for the respective filters; being 25073877 for Trump and 939091for Biden. Moreover, as the Search API has limitations in the number of retrieved tweets, it was necessary to identify the tweet id of the tweets in their timeline in order to filter groups of 100 tweets for both candidates until completing the existing datasets with up to date tweets.

It is important to highlight that during the mining and extraction process, it was observed that many attributes are extracted from a tweet in a JSON format, from both the Stream and Search API's. As the list is very extensive, only the following relevant attributes were considered in this study:

- **Tweet id:** Id that Twitter gives to every tweet. (Unique)
- **User id:** Id assigned to users. (Unique)
- **User:** Name of the account. E.g. @joebiden
- **Date:** date when the tweet was created.
- **Text:** words in tweet.
- **Followers count:** Account followers.
- **Place:** Place where account was created.

To avoid catching tweets from fake accounts and bots, the filtering-out of accounts with no followers was introduced in the code and these accounts discarded. Moreover, for the attribute place it was discovered that many accounts belonged to other countries and therefore discarded as well. Lastly, it is important to highlight that when reviewing the dataset inappropriate use of the language can be found, however this report does not publish any specific tweet text or words. It only reports the result of a sentiment analysis performed over the tweet Text.

Data cleaning and processing pipeline:

A data pipeline was built to transform the streamed tweets in JSON format to an organised data frame in a csv format, that will later be analysed by the use of Microsoft PowerBI. This, because of the structure obtained by the process, allowing a filtering and visualisation application like PowerBI, to perform rapid analyses through a dashboard format. The final csv file containing all the processed data was constructed day by day, as tweets were collected in batches. This because both processes streaming and wrangling data in the pipeline were computationally expensive and required to be run simultaneously to achieve the desire quantity.

In terms of the pipeline, regular expression methods were used to extract meaningful information from the tweet. First, the library re in python allowed to search and match words in the tweet, in order to find the candidate, the user was referring to when writing a tweet. This not only allowed to identify if the tweets were referring to Trump, Biden or both, but also helped filtering out those tweets were neither of them were mentioned. Second, the pipeline through the re library also allowed to identify the location where the tweet was written. For this, a dataset of cities, towns and states in the United States was built to perform the search and match functions on the location given by the bio of the twitter users. It is important to mentioned that not every user has their location publicly available in their bio, which was required to be inputted by the authors of this report. For this, it was assumed that the likelihood of a user to be in the United States, when writing a tweet that refers to Trump and/or Biden in English language, is very high. Therefore, the missing values were replaced by 'United States', while the existing available locations where searched and matched over, to find the specific states where the account belonged.

Sentiment analysis:

The tweet text field was used to perform a sentiment analysis and conclude whether users have written a Positive, Negative or Neutral tweet. The Natural Language Processing libraries TextBlob and NLTK Vader were used with a sentiment analysis function applied to every tweet text. These functions yielded a sentiment polarity score and a compound score, respectively, both with values between 1 and -1, which was used to determine the connotation of the text.

This study proposes an assemble method to determine the sentiment of the tweet based on the vote of the two algorithms, which represents an improvement over Oikonomou & Tjortjis (2018) and Yaqub et al. (2017) who only used one analysis tool based on Naïve (Naïve Bayes model and SentiStrength respectively). However, TextBlob was also used for comparison and subjectivity by Oikonomou & Tjortjis (2018), but with results not assembled. Each one of the Natural Language Processing tools used for this study offers a different advantage and yield a different result according to their interpretational settings, which led the authors of this report to conclude that assembling represents a compiling and more accurate way to approach a sentiment analysis. Therefore, the sentiment scores yielded by both algorithms were averaged, and the result was used to classify the tweet as positive, neutral or negative. On one hand, TextBlob provides a function sentiment that returns a polarity and a subjectivity value and uses a ruled-base sentiment analysis library. This library is helpful because it calculates not only the sentiment but also how subjective or objective it is. If the polarity score is greater than zero the text is considered as positive, zero as neutral and less than zero as negative; and if the subjectivity (which is a score between 0 and 1) is closer to 0 the text is considered objective and closer to 1 subjective. On the other hand, NLTK Vader provides a function that classifies each component of the text into positive, neutral and negative and uses them to yield a compound or aggregated score that is normalised to be between -1 and 1. Similarly, if the compound score is less than zero the sentiment is negative, if zero it is neutral and if greater than zero it is positive. Hence, taking advantage of the two NLP tools, it has been decided to assemble both polarity and compound scores to produce a final sentiment score. This final score is then used to classify the text as positive if the score is greater than 0.05, neutral if the score is less than 0.05 or greater or equal than -0.05, and negative if the score is less than -0.05.

Prediction:

In contrast with the related works mentioned previously, this study considers both Positive and Negative tweets for the prediction of potential voters. While Oikonomou & Tjortjis (2018) only considered the positive tweets to account potential votes for a candidate, this study considers the positives as a vote for the mentioned candidate and the negatives to the candidate as votes for the opposite candidate. For instance, if a tweet mentioning Trump has been classified as Positive and a Tweet for Biden as Negative, then Trump will receive two votes. In order to reduce noise and the effect of bots and spams or accounts with hyperactivity, the user id field has been used to average the scores of a twitter account with multiple tweets. Thus, if an account is identified as hyperactive with several tweets made throughout the period of study, this one will only account for a single vote rather than being counted multiple times.

Subsequently, an analysis of the swinging states is performed separately from any other location, to determine the winner in that state. Florida, Ohio and North Carolina states represent those swing states and were assumed as the decisive states of the next election. In addition, following the Social Network Analysis Theory, influencers have been identified as those accounts (nodes) with high degree centrality (connections), and a proportion of their followers have been accounted as voters with the same sentiment of the influencer. This, by taking in consideration the theory of adopting behaviours. Hence, accounts with follower greater than 5,000 have been considered as influencers and a percentage of their followers are assumed as likely to adopt the same behaviour. Therefore, 0.05% of an account's followers is counted as voters with an adopted sentiment, with the assumption that at least 20% will likely embrace the sentiment. Additionally, it has been considered that there is likely that one account follows several influences, which has been included in the accountability of influential account's followers and the percentage of voters. Thus, the overall prediction considers both swinging states as well as influencers that help predicting the behaviour of a great part of the U.S. population. The estimation of the 0.05% was set in a conservative arbitrary way, taking in consideration the assumptions outlined. However, for a future study the authors of this study recommend the use of Bayesian Statistics to infer this number more accurately.

4. Experiment and analysis

4.1 Trump and Biden Activity in Twitter

The authors of this study collected tweets from both candidates from the start of the year, January 1st, 2020 till June 7th, 2020. As figure 1 shows Trump has made over 3000 tweets as compared to Biden tweeting just over 1000 times. From this, the team sees how active Trump is on Twitter as compared to Biden.

For the experimentation and analysis of tweets by Trump and Biden, Power BI was used to visualize the collected tweets. The columns used for the Power BI dataset were the username of both candidates, the tweets made, the number of favourites and retweets, as well as customised columns of both sentiment score and key phrases. The two customised columns were produced by the code available in Power BI that calculated the sentiment score of tweets as well as to filter out the key words of the tweet. The visualisations done on Power BI will consist of a word cloud for each candidate that consists of the key words, a pie graph that distinguishes the sentiment score of tweets, and bar graphs of tweets made, favourites, and retweets.

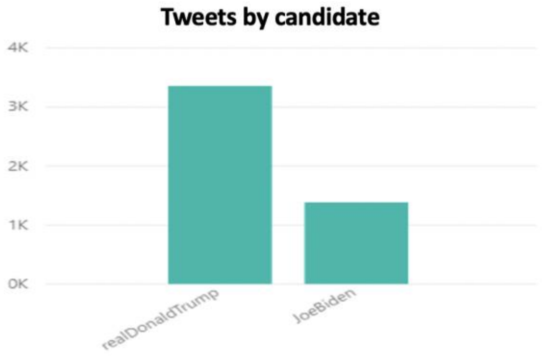

Figure 1. Bar graph of tweets made by Donald Trump and Joe Biden

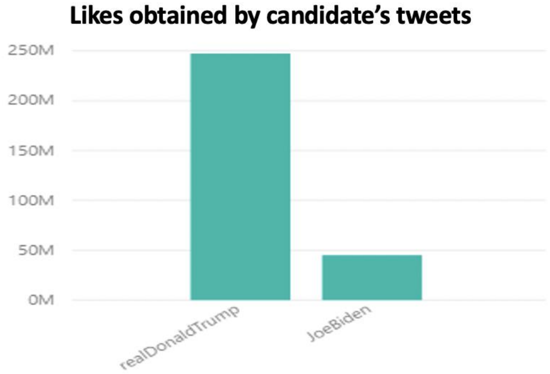

Figure 2. Bar graph showing the number of favourited tweets

Figure 2 represents the total number of favourites received by each candidate for the tweets they made from January 1, 2020 to July 7, 2020. Trump's totals close to 250 million favourites versus Biden's 50 million. This may be the case, due to the number of Biden's followers which is at 6 million, while Trump has 81.9 million. The number of supporters for Trump online based on favourited tweets clearly beats Biden. This aspect is taken in consideration when making the prediction, but by itself is not conclusive and required its integration with the population analysis.

Similar to Figure 2, Figure 3 clearly depicts that Trump beats Biden with the number of interactions made on Twitter. Trump has roughly around 70 million retweets, while Biden's tweets being retweeted is around 10 million. Again, it is deduced that this is due to the number of followers each candidate has, but might not correspond or match the voter's mind, entirely.

Figure 4 represents the tweets made by both candidates every month from January until June (as June is still in progress, this month's values correspond only to a third). The graph shows an insight on how activate each candidate is on Twitter, especially during these crucial moments happening in the United States at the moment. As displayed on Figure 4, it is clearly seen how active Trump is as compared to Biden on Twitter, especially during the month of May and June when Covid-19 and the Black Lives Matter movement have impacted US society widely. The biggest growth in number tweets by Trump is seen between April and May with an increase of 80%, while Biden had a small decrease in the number of tweets. This can be explained by the difference in strategy by both candidates, which from Trumps perspective it lays on Twitter engagement while from Biden perspective, his campaign may be looking at other channels rather than twitter.

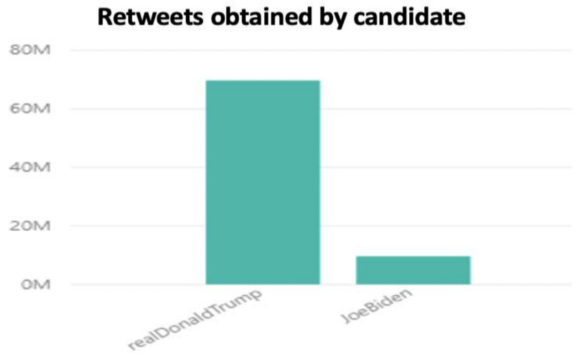

Figure 3. Bar graph showing the number of retweets

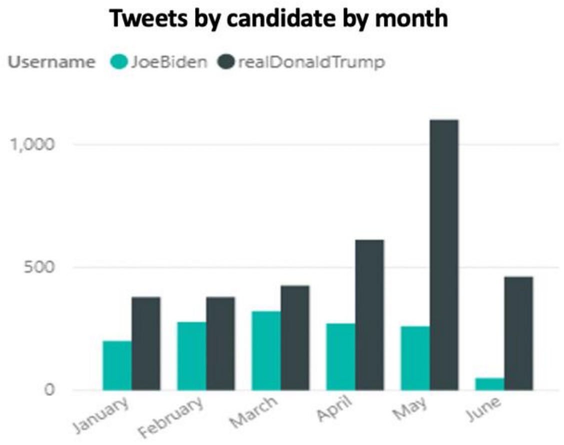

Figure 4. Tweets made by both candidates per

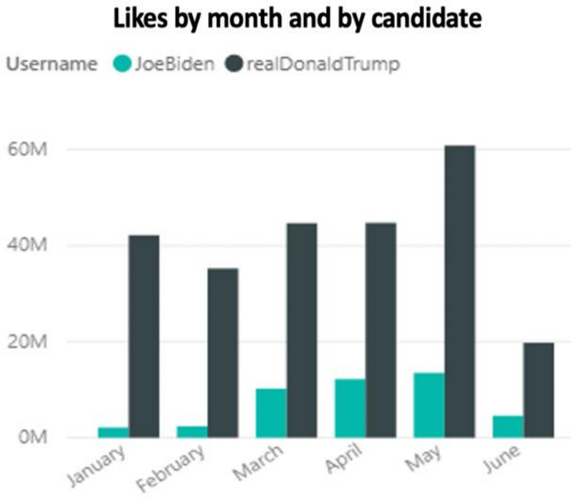

Figure 5. Favourites gained by each candidate per month

Figure 5 provides a clearer view of the favourites (likes in twitter language) received by each candidate per month as well as sees the gap in supporters of both candidates. Just in a span of a month, between April and May, Trump outperformed Biden with the growth rate of favourites (likes) received in tweets, with a 36% increase. In contrast, in the same period Biden presented a 2% increase in likes over the same period but, he had a decrease in the number of tweets, while Trump presented an increase. This gives an intuition that Trump has gained relatively more popularity over the last three months, than Biden, but it still lacks evidence for conclusion of who is more likely to win. It is also considered that the explanation to this phenomenon is given by the high level of engagement that Trump has with his followers on Twitter, through the frequency of his tweets, whereas it is observed that Biden might have a different engagement strategy.

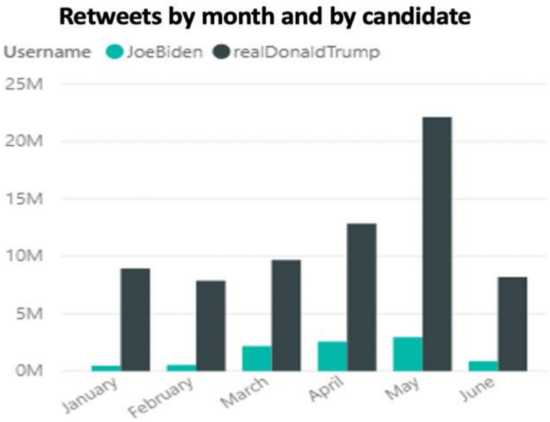

Figure 6. Bar graph showing the number of retweets

Lastly, the number of retweet interactions of both candidates follows the same pattern as the number of likes or favourites month by month. From figure 6, it is fair to conclude that Trump surpasses Biden in the absolute numbers of online followers and support. Trump presented an increase rate in retweets of 73% from April to May, while Biden had a 15% increase of retweets in the same period.

From this figure, it is possible to conclude that, if the analysis was based only on the number of followers and the engagement that the candidates have with their audience on Twitter, Trump will be predicted as the winner. However, these aspects cannot be studied in isolation and required the sentiment analysis study over the tweets of people in United States.

Trump and Biden tweets' Sentiment Analysis.

A word cloud visualization tool on Power BI was used to picture the top key words used by each candidate. The word cloud gives the team an idea on which key word on phrase the candidate uses the most on Twitter. As seen on figure 7, Biden's key words seem to address issues regarding Americans as well as current matters and affairs. This provides an insight of the impact and sense of the tweets being sent out by Biden. Moreover, it is possible to observe that a hot topic for Biden is his contestant Donald Trump, as the size of these words shows the frequency in which they have been mentioned in Biden's tweets. "people", "nation", "president" and "country" follow as second in Biden's interests.

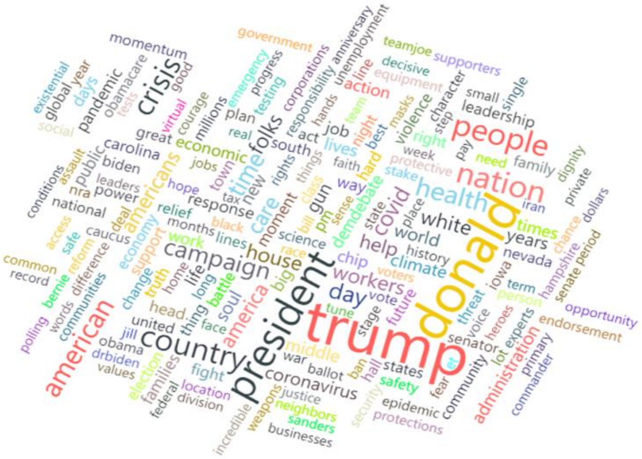

Figure 7. Biden's Word Cloud

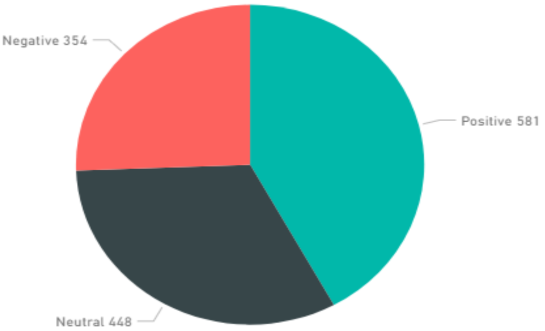

Figure 8. Sentiment Score of tweets made by Biden

Additionally, Figure 8 is a pie chart that has recorded the total sentiment score of the tweets made by Biden in terms of how positive, negative, or neutral the tweet is. 42% of Biden's tweets seem to be positive tweets, 32% of his tweets are neutral, and 26% are negative. This pie chart provided more information on the message that Biden is sending through Twitter and based on this chart, majority of his tweets give off a positive sentiment.

For the case of Trump's tweets, based from his word cloud, he seems to be using a lot of strong words in his tweets such as "hoax", "fake", "great", and such. As compared to Biden's word cloud, Trump's word cloud is more direct and impactful. In addition, it is possible to observed that Trump mentions his rival less than Biden mentions him, in relative terms. Which indicates that his strategy focuses on other matters, rather than on his contestant.

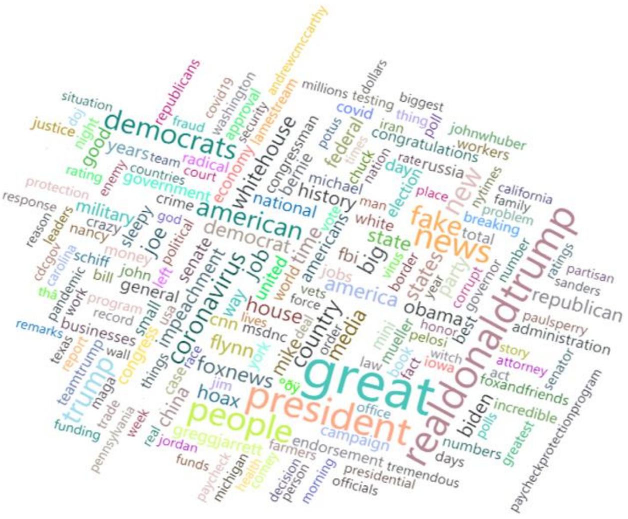

Figure 9. Trump's Word Cloud

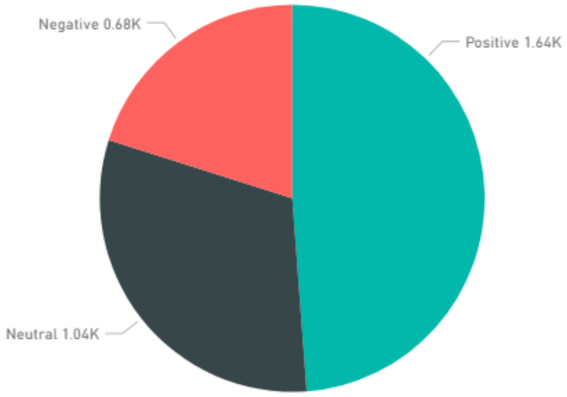

Figure 10. Sentiment Score of Tweets made by Trump

As presented on figure 10, it is possible to see that 48% of Trump's tweets have a sentiment score of positive and only 20% of Trump's tweets are negative. These results coincide with the previous bar graphs showing the number of likes and retweets Trump receives.

4.2 Twitter Users, analysis of swing states, country and general prediction:

This study evaluates the sentiment analysis of tweets made in United States to determine who will win the presidential elections in 2020. First, an overall evaluation of the whole country is presented to provide an overview of the general sentiment, considering all tweets. Subsequently, the results obtained in the 3 states are presented separately to construct the prediction. Lastly, a final inference is provided considering the Social Network Analysis theory and the impact of influencers on adopting behaviour; this to estimate voters for each of the candidates based on unique accounts and how influential they are.

Results:

United States: Considering the tweets from users around all states in the U.S. it is possible to observe that Trump has been mentioned more times that Biden in the last week. The following graph shows the overall picture of the sentiment towards both candidates.

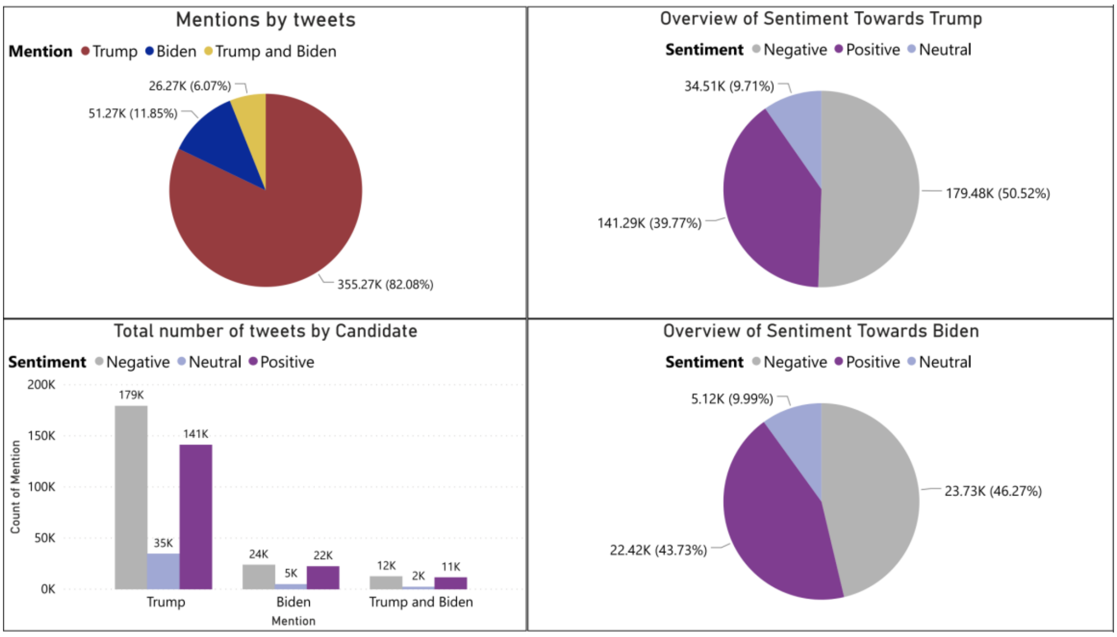

Figure 11. Sentiment Analysis of all collected tweets for all locations in U.S

From figure 11, it is possible to observe that the candidate Trump has been mentioned 82% of the total number of tweets collected in the last week, while Biden is only addressed 11.85% of the times. This indicates that the latest events have gotten people to talk more about Trump than Biden on Twitter.

Comparing the sentiment results, it is possible to observe that the proportion of negative tweets is greater for both candidates, indicating that people are more likely to comment something negative than something positive or neutral in tweeter. However, the comparing the two proportions in both candidates, it is possible to observe that Trump has gained a bigger proportion of negative tweets than Biden does. While 50% of the tweets addressing Trump are negative, only 46% of the tweets addressing Biden are negative. Alternatively, Biden has a 43.7% of positive tweets whereas Trump's positive tweets only account for 39.8% of the tweets addressing him.

It is important to highlight that even though Biden has considerably a smaller number of tweets addressing him, it is possible to consider that tweets addressing negatively one candidate can represent a potential vote for the opposite candidate. Thus, if this premise was to be true and the elections were held at the moment of this report, Biden would be the winner of the 2020 elections and the 46th president of the United States, by popular vote. On the other hand, as the number of tweets addressing both candidates only represents 6% of the total number of tweets collected, and its interpretation can be ambiguous due to the variation of the names and words distribution and location in the tweets, they were not considered for analysis.

Swinging states:

As it has been highlighted in this report, swinging states are considered to be decisive in the presidential elections. Their political affiliation swings between democrat and republican and as they are not married to one, they can determine the outcome of an election. For this reason, the same analysis has been applied to the states of Florida, Ohio and North Carolina.

Florida:

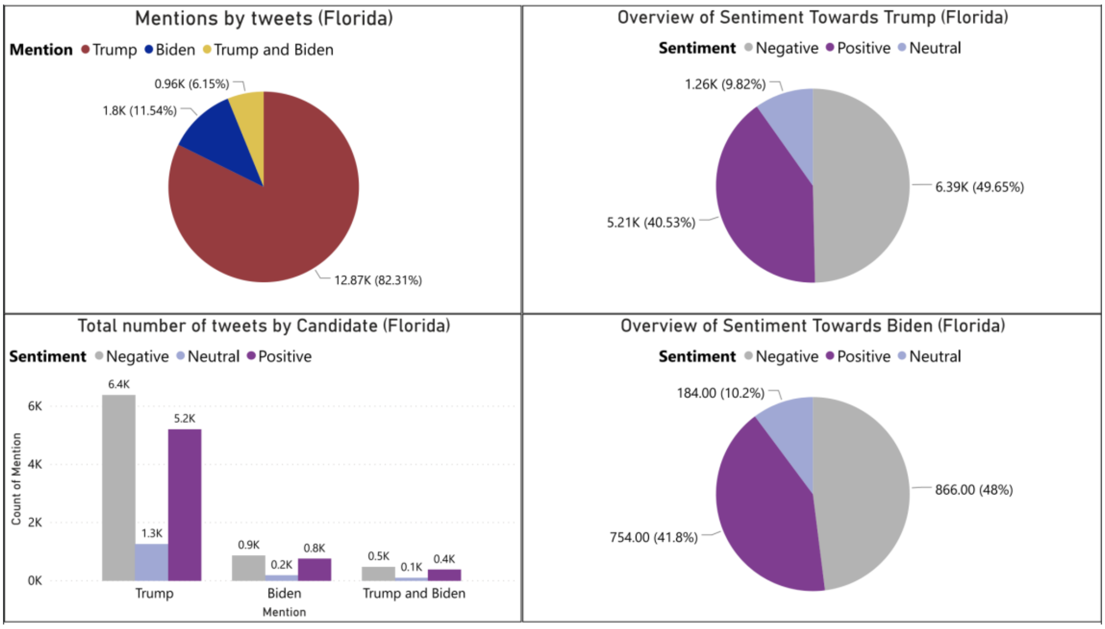

Figure 12. Sentiment Analysis of all collected tweets in Florida

From figure 12 it is possible to observe that the proportions in terms of mentions by tweet are similar to the entire country. However, Trump presents an improvement in the number of positive tweets as compared to the whole country, whereas Biden's negative proportion of tweets presented an increase, as compared to the whole country.

Therefore, if Florida was to decide the 46th president of the United States if elections were taking place at the moment, there would be a tie between the two candidates, leaving the decision to the rest of the country. In that case, Biden would be proclaimed the 46th president of the United States by the popular vote, as observed in the national analysis.

Ohio:

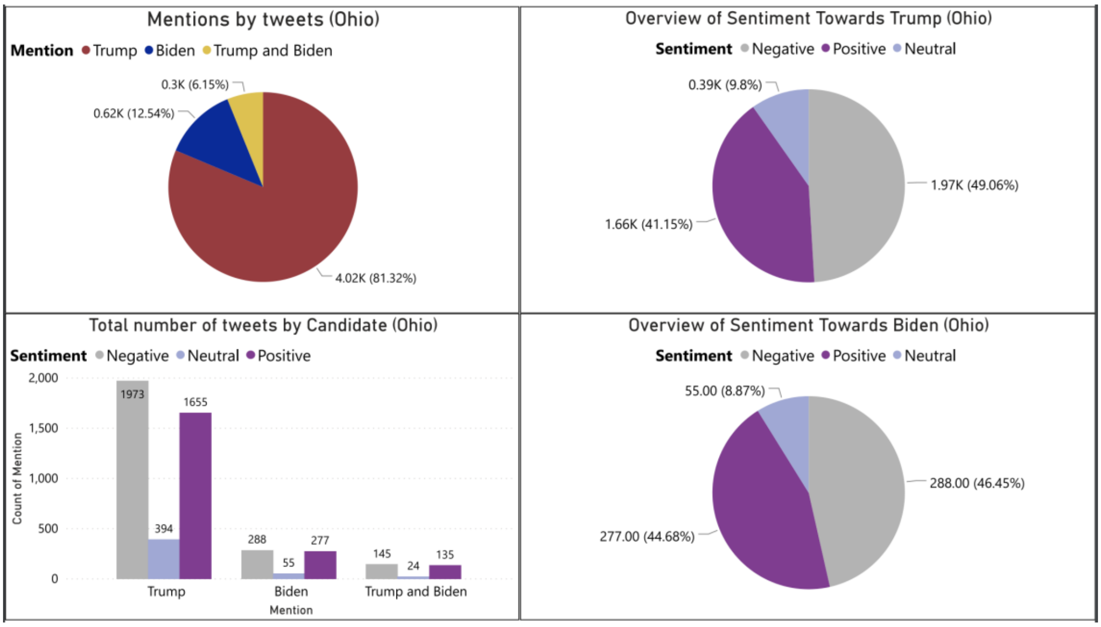

Figure 13. Sentiment Analysis of all collected tweets in Ohio

Figure 13 shows the sentiment analysis applied only to the state of Ohio. This state follows the national distribution and proportions with Biden being mentioned proportionally more than in the national analysis. There are improvements in the proportions of positive tweets for both candidates in comparison to the national results, with Trump winning 1.38 points and Biden 0.95 points. However, Trump's proportion of Negative tweets is bigger than Biden's, indicating that if Ohio's popular vote was to elect the next president of the United States, the winner will be Joe Biden. The margin between the two candidates is similar to the presidential results of 2016 where Hillary Clinton won the national popular vote by 2.1%.

Similarly, tweets that address both candidates were discarded, due to the high variability among the order and location of key words and names in the tweet, making the analysis difficult to scale.

North Carolina:

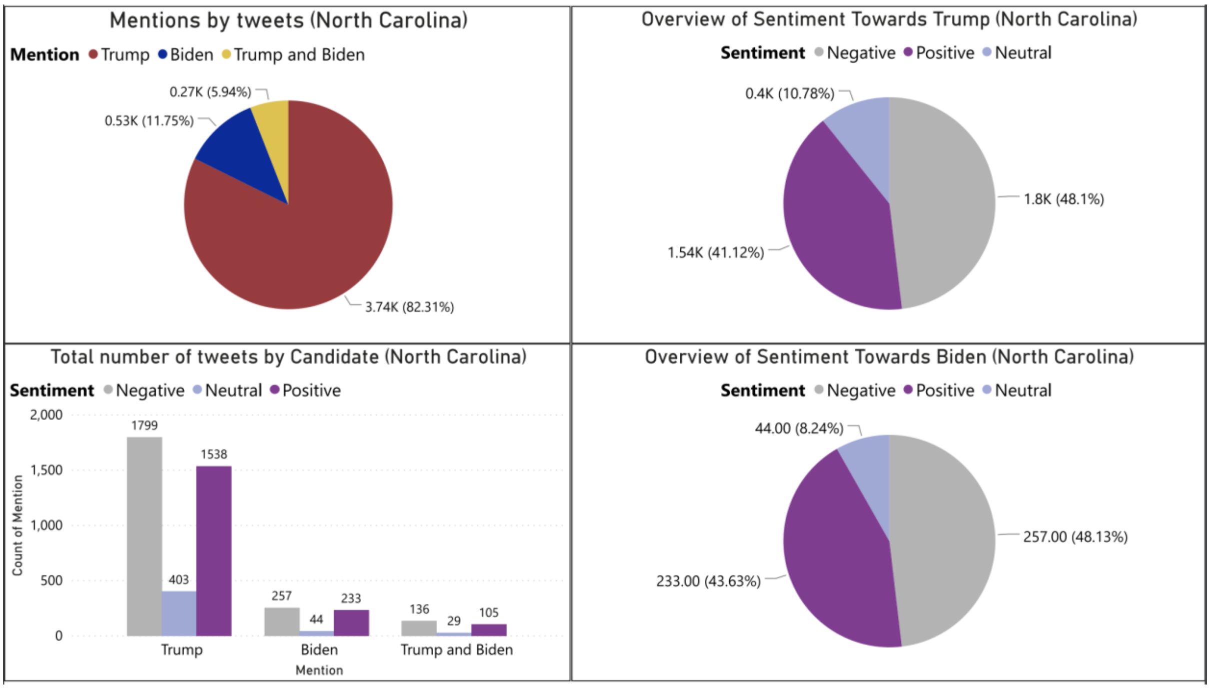

Figure 14. Sentiment Analysis of all collected tweets in North Carolina

Lastly, Figure 14 depicts the last analysis for the state of North Carolina. Similar to the national results Biden has a higher positive proportion in comparison to Trump, despite Trump having a greater number of mentions. On one hand, both candidates presented the same proportion of negative tweets, indicating that there can be a tie if the elections were based on negative sentiment. On the other hand, Trump having a greater number of negative tweets has a positive impact for Biden candidacy, because they represent a potential vote for Biden. Therefore, it is possible to observe that Biden will win this state, under these circumstances.

Thus, if the elections were determined by these three states, were taking place today and were following the assumptions made, Biden will be elected by popular vote as the next president of the U.S. Moreover, these results are comparable to Oikonomou & Tjortjis (2018), as they follow the same assumptions and process with an addition of an assembled sentiment analysis.

Influencers and adopting behaviour, overall prediction:

The previous results present a weakness that is addressed in this section because the methodology does not address the matter of existing influencers, their followers and adopting behaviour. It also omits the fact that several tweets made by a single account introduce noise and bias into the results, and therefore should not be considered as representative. In order to get as close to reality as possible, an analysis on the influencers and their followers is presented as a complementary approach to the study and represents an alternative to the previous analysis. Hence, in order to introduce these factors in the study, additional processing of the data was required. First, the data was further reduced by considering only unique accounts and their followers; so only the last tweet posted addressing a candidate was considered. Then the data was reduced from over 430 thousand tweets to 110 thousand unique accounts. In terms of Social Network Analysis and adopting behaviour this method presents an advantage by allowing the analysis over the number of followers of an account and its sentiment and by reducing further the noise created by hyperactive accounts. However, despite the methodology has the disadvantage of missing important information due to the consideration of only the last tweet made by the account, it presents a great opportunity for further research.

Secondly, a proportion of the number of followers of the influencer accounts was considered as adopters of the same behaviour and sentiment, and therefore as voters. This proportion was established based on the assumption that an account can follow multiple influencers with a possibility of clashing sentiments. Moreover, the likelihood of being chosen randomly and being a unique follower to an account, in a pool of multiple influencers, was also considered. This ensured that followers were accounted as unique voters once and not multiple time and considered that the larger the pool is, the smaller the probability of being unique follower and being chosen it gets. Therefore, it was established that the chances are 0.05%, which means that this percentage was taken as the proportion of unique followers of a particular account that embrace the sentiment and become unique voters. Lastly, an account with 5,000 followers or more was considered as influencer, and whose discussed proportion of followers would become voters. Accounts with less than 5,000 were not considered as influencers, their followers were not considered as voters and the account only accounted for one voter.

Hence, this analysis was performed based on influencers, a proportion of the followers that adopt the sentiment of the influencer and the sentiment and account snapshot of the last posted tweet.

Results:

United States: The initial results observed by accounting unique twitter accounts as unique voters indicate that Biden would win the popular vote by 49.04%, while Trump would only obtain 40.86% of the popular vote (Fig. 15 A). These results become even more favourable for Biden if the theory of influencer’s and the proportion of followers adopting the sentiment of the account are taken in consideration. It is possible to observe that by the results that influencer accounts that would vote for Biden have more followers than accounts that would vote for Trump. Hence, Biden would further beat Trump by 53.98%, while Trump would lose by 35.75% (Fig. 15 B).

In either case (A or B) Biden would beat Trump. However, the case where the votes are counted from unique accounts are very similar to the national average poll published by Real Clear Politics 'Polls' (2020) where Biden wins the 49.8% of the popular vote while Trump only gets

41.7%. This indicates that the distribution of the unique accounts results is similar to the distributions of the average of the randomly sampled polls performed by different media channels in the United States like CNN or Fox News. Moreover, according to Skelley (2020) Trump's last approval rating was 41.1% which is very aligned and consistent with the national results obtained in this study. Thus, figure 15 A and B depicts both unique account as voters and influencers followers as voters' results obtained from the sentiment analysis of tweets over the past week.

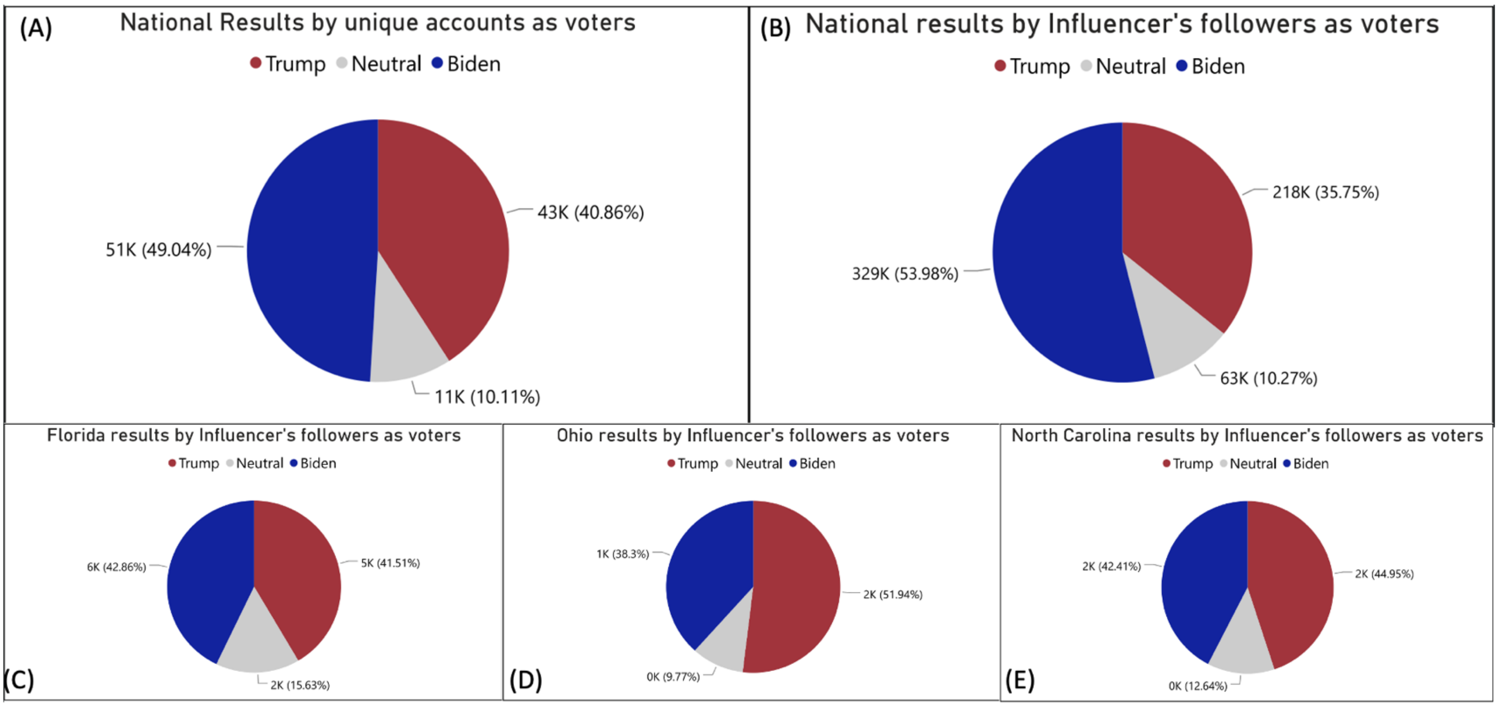

Figure 15. Influencer's analysis nationally and by state

However, despite Biden winning the popular vote nationally Trump still holds high chances of winning because of the results obtained in the key states. In comparison to the 2016 election, the candidate Hillary Clinton won the popular vote by 48.2% while Trump only got 46.1% of the votes and yet, Trump as elected by the electoral college as the 45th president of the United States. Furthermore, Clinton was predicted to be the winner by the vast majority of polls performed prior the election which resulted in a big surprise when the opposite occurred. Then Ohio, North Carolina and Florida turned red, aspect that gives explanation to the 2016's outcome. Hence, there are a lot of similarities in this study, the actual national polls and the 2016 national polls where the Democratic candidate was predicted to be the winner, and therefore it becomes necessary to discuss and study the key states.

Key States: Considering the proposed proportion of followers of influencers in the key states as voters, the results show a different behaviour from the national results. While Biden wins in Florida by 42.9% over 41.5% obtained by Trump, Trump wins in Ohio and North Carolina by 51.9% and 44.95% (Fig. 15 C, D and E) over Biden's 38.3% and 42.4%, respectively. Again, these results aligned with the results published by Real Clear Politics 'Polls' (2020) but are more pronounced in this study. Therefore, if the decision would fall under these three states Trump will win the elections, despite Biden winning the national popular vote by a larger margin; generating a very similar situation observed with the polls and the actual results of the 2016 election.

5. Conclusions

From the experimentation and analysis of the tweets made by Biden and Trump, the followers of Trump on Twitter appear to support and agree to Trump's point of view. However, his followers are not an accurate representation of U.S. voters, as the candidate may have many other non-us citizen followers. On the other hand, Biden presents less activity and less followers and reactions and also can equally be considered as not representative to predict the outcome of the election. The first part of the study shows the contrast in strategies by both candidates and the use of Twitter and Biden has a lower engagement, while Trump has great support and followers.

On the other hand, considering the entire pool of tweets with no account filtering, if the election outcome was based on the key 3 states, the winner was predicted to be Biden as the sentiment of the tweets in these states shows a more positive image towards this candidate. In addition, aggregating the results of the three states, it can be observed that the negative proportion of tweets obtained by Trump represents a voter opportunity for Biden, giving him an advantage for the outcome of the upcoming election. However, this methodology was found as biased because multiple tweets by a single account would inflate the results observed.

In contrast, the weakness found in the first study was addressed by only considering the last tweet produced by the collected accounts. Then, the analysis was performed over single accounts being accounted as unique voters followed by the consideration of the existence of influencers and their impact on the behaviour of a proportion of their followers. By doing so, it was possible to conclude that a similar situation to the 2016 election is being observed. The democratic candidate Biden wins the national popular vote by a larger proportion but is likely to lose the presidency again because of the results observed in the key states, where Trump wins 2 out of three. Moreover, the results observed in this study aligned with the results published by actual polls performed by several media channels in the United States, in addition to an observed consistency with the actual approval ratings. This provides confidence that the methodology applied was successful and consistent to reality.

Finally, the results presented in this study are early results and required the analysis of the upcoming events until November 2020, when the U.S. election takes place. However, it is considered that the methodology presents an improvement to similar studies that have been conducted previously on the 2016 elections. This, because of the extra considerations taken in the data processing and cleaning, and the Social Network Theory of Influencers and adopting behaviours, which compiles and complements the Social Network Analysis performed over real-time streamed tweets.

6. References

Borchers, C. 2016, 'The wrongest media predictions about Donald Trump', *The Washington Post*.

Chouhbi, K. 2020, 'Twitter Deep-Dive Analysis: US Presidential Election', viewed June 1st 2020, <https://medium.com/analytics-vidhya/twitter-deep-dive-analysis-us-presidential-election-d6b32acc449b>.

Developer, T. 2020, 'Search Tweets', © 2020 Twitter, Inc., May 30th 2020, <https://developer.twitter.com/en/docs/tweets/search/overview>.

Oikonomou, L. & Tjortjis, C. 2018, 'A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter', *2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM)*, pp. 1-8.

'Polls' 2020, viewed 13 June 2020, <https://www.realclearpolitics.com/epolls/2020/president/us/general_election_trump_vs_biden-6247.html>.

Rao, R. 2020, 'Joe Biden Tweets (2012 - 2020)', <https://www.kaggle.com/rohanrao/joe-biden-tweets/>.

Reese, A. 2020, 'Trump Tweets', <https://www.kaggle.com/austinreese/trump-tweets>.

Skelley, G. 2020, 'Trump’s Approval Rating Has Dropped. How Much Does That Matter?', *fivethirtyeight*.

Yaqub, U., Chun, S.A., Atluri, V. & Vaidya, J. 2017, 'Analysis of political discourse on twitter in the context of the 2016 US presidential elections', *Government Information Quarterly*, vol. 34, no. 4, pp. 613-26.