Danish Ahmed Bombal: 1007733008

ECO225 Final Paper

Professor Nazanin Khazra

April 15th, 2023

Sentiment Analysis of Tweets for the 2020 US Presidential Election

Introduction

Social media platforms including Facebook, Instagram, and Twitter have altered how we share and communicate news and opinions regarding current affairs entirely. Twitter plays an integral part in the mainstream media and has allowed it to become a platform for electoral campaigning and the political discourse as a whole. Influential and powerful individuals and organizations including politicians, political parties, and news agencies have increased the use of Twitter to change public sentiment regarding current and public affairs (Matalon, 2021). Twitter has been used for building political narratives and has become a driving force since over two decades. Unfortunately, the nature of tweets is ambiguous since it represents emotions and so, we must identify and categorize them. To analyze whether these opinions are in support or against a party, sentiment analysis is a tool typically chosen to approach such a problem.

Sentiment Analysis has been previously applied on tweets from the US 2012 Presidential Election (Mohammad, 2014) as well as the 2016 Presidential Election (Joyce, 2018). Moreover, not only in the US but sentiment analysis has also been conducted on tweets on the political tweets of UK (He, 2012), Spain (Rodriguez-ibanez, 2021) and India (Ansari, 2019). The goal of sentiment analysis is to understand the opinions of people. It analyzes emotions, opinions, and attitudes towards issues and uses natural language processing and machine learning techniques to automatically identify subjective information from text data and then categorize it into positive, negative, or neutral sentiments (Coletto, 2015). Deep learning has been implemented in sentiment analysis of political tweets as well (Pota, 2018).

In this essay we aim to explore and quantify public engagement toward political parties on twitter to predict the election results. Throughout the research we observed how the likes, retweet counts, and user follower counts of tweets were positively related to the vote counts of most states in the United States via the number of tweets during the 2020 elections. We hypothesized that tweets with positive sentiments and emotions would have a positive correlation with vote counts in general.

Moreover, to complement our research and for a rigorous analysis, we incorporated external data that could potentially explain and predict the election outcomes alongside the twitter data. We unified the population, election results and the COVID-19 data sets in different stages of the research to explore relationships with different variables that could help us predict how and why Biden was victorious in the 2020 US presidential election.

Moreover, this paper takes a relevantly new approach to analyzing tweets in terms of time intervals. A time-series analysis of tweets that contains different sentiments for the respective candidates, Trump and Biden, was conducted. It indicated of how simple moving average models for different time windows of the sentiment scores can be related with real world events and how controversial incidents can impact electoral campaigns. Nevertheless, the results of the research are statistically consistent with the election results where Biden won in most states as predicted by the twitter data becoming the next President of the United States of America.

Data

During the 2020 US presidential election campaign, millions of tweets were posted by users around the world. The data used in this particular research consist of the two primary election candidates, Donald Trump, and Joe Biden with about 1.72 million tweets available. The tweet

data was accessed from Kaggle which has been collected since October 2020 for approximately one month (24 days to be precise) till the election month of November 2020 (Hui, 2020). In particular we kept all the tweets that were created before the election date, November 3rd, to make sure our research findings and prediction were based on the data prior to the actual elections. The tweets themselves were collected using the Twitter API. Each tweet possesses some common features that are included in the dataset. Some quantitative features of a tweet are number of retweets, number of likes and followers count on tweet creator. Moreover, every tweet contains the location that was parsed into different categories including its geographic location, state and city. There are many tools to measure the popularity of someone on platforms like Twitter. For instance, the number of tweets for each candidate could demonstrate their number of supporters. Similarly, the public engagement on these tweets like the number of retweets would indicate how well-known a particular tweet is and how often the message in that tweet is supported by others on the platform. Similarly, the number of likes on a particular tweet is another important variable to measure the popularity of a tweet and whether people favor a particular agenda or not. Moreover, the location origin of a tweet is of utmost importance since that would indicate whether the tweet count of a specific state within the US eventually dictates a relation between the real vote counts of the mentioned state and its ultimate result. We had tweets from around the world from multiple countries about the elections but we kept the tweets only originated from the United States to understand the actual election results within the US with respect to the sentiment of tweets. We had to remove some tweets that was missing information about states the tweet originated from since our analysis was based on state election results which would hinder our research. Moreover, heavy data wrangling was required to work with the data set since the tweets were far from clean. We removed any duplicate tweets that

contained both the candidates' names and were present in both the data sets to prevent any overlap. Similarly, we had to clean our tweet content. We removed mentions (@), hashtags (#), punctuations, website links, numbers, non-English tweets and removed any irrelevant words that didn't add any meaning to the tweet (stop words). We had to clean our data since we used a lexicon based sentimental analysis through VADER (Valence Aware Dictionary and Sentiment Reasoner) which only accepts cleaned English words and sentences and it automatically identifies the sentiment of a text based on words and slangs. The Sentiment Intensity Analyzer function is present in the natural language toolkit (NLTK) of python that processes our cleaned data and output the polarity scores given the content of the tweet. However, after some exploratory data analysis, we observed that majority of the tweets are regarding Joe Biden and that the number of tweets about Donald Trump are relatively low.

Summary Statistics

There were a few key characteristics that were unique to both the candidates respectively when it came to identifying the number of likes, user followers and retweets. After performing a sentiment analysis and assigning each tweet a sentiment of positive or negative we observe the total counts for these to make sure if there would be any signal over the reason why Biden won in the elections within the provided data.

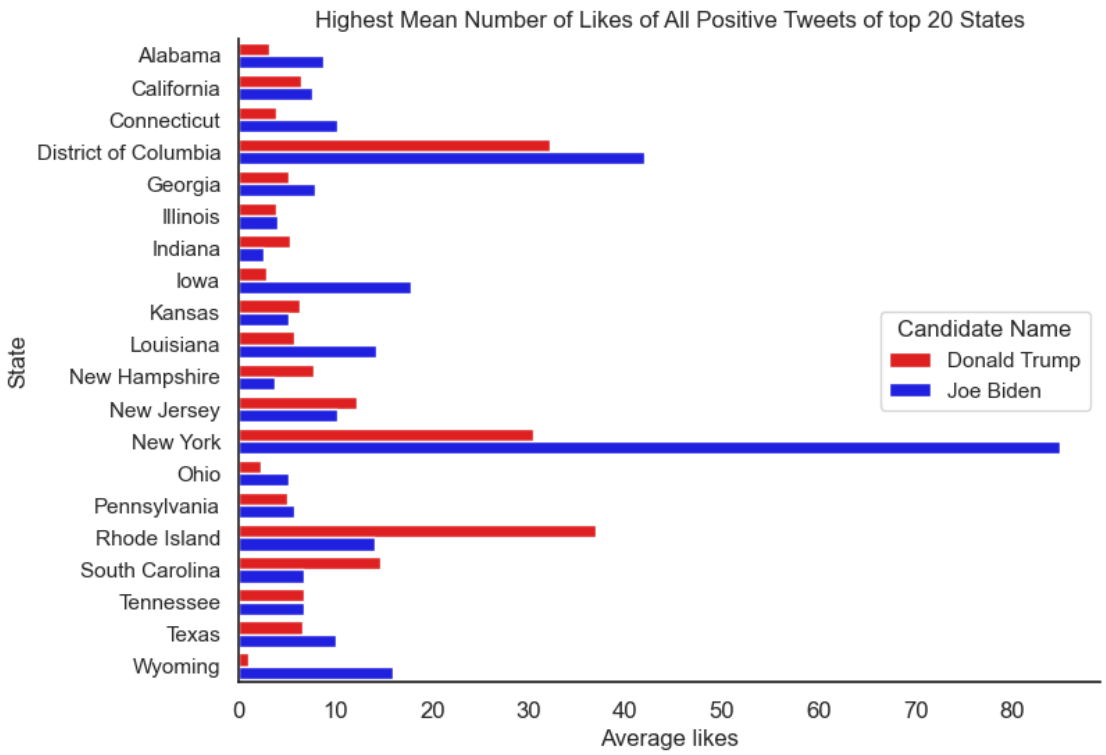
Number of Likes

The table below shows a higher number of tweets of Donald Trump than Joe Biden. Moreover, the total number of likes with Joe Biden tweets (777,191) is higher than Trump's total likes (1,035,446). This leads to the obvious conclusion that the average number of likes received by Biden tweets is higher than Trump's tweets as shown in the table. It is quite spectacular to

observe that the maximum number of likes on a single tweet of Biden is almost seven times more than any tweet that mentions Trump. This clearly indicates Biden's popularity in general.

Candidate Name	count	mean	std	min	25%	50%	75%	max
Donald Trump	104155.0	7.461869	173.643278	0.0	0.0	0.0	1.0	25987.0
Joe Biden	76632.0	13.511927	695.563695	0.0	0.0	0.0	1.0	165702.0

We also observed the highest mean number of likes of all positive tweets of top 20 states of each candidate. We observe that New York has the highest number of likes on average for Joe Biden positive tweets. It is not a coincidence that Joe Biden in fact did win in New York and the District of Columbia while Trump won in Indiana where he had a higher number of likes on positive tweets.

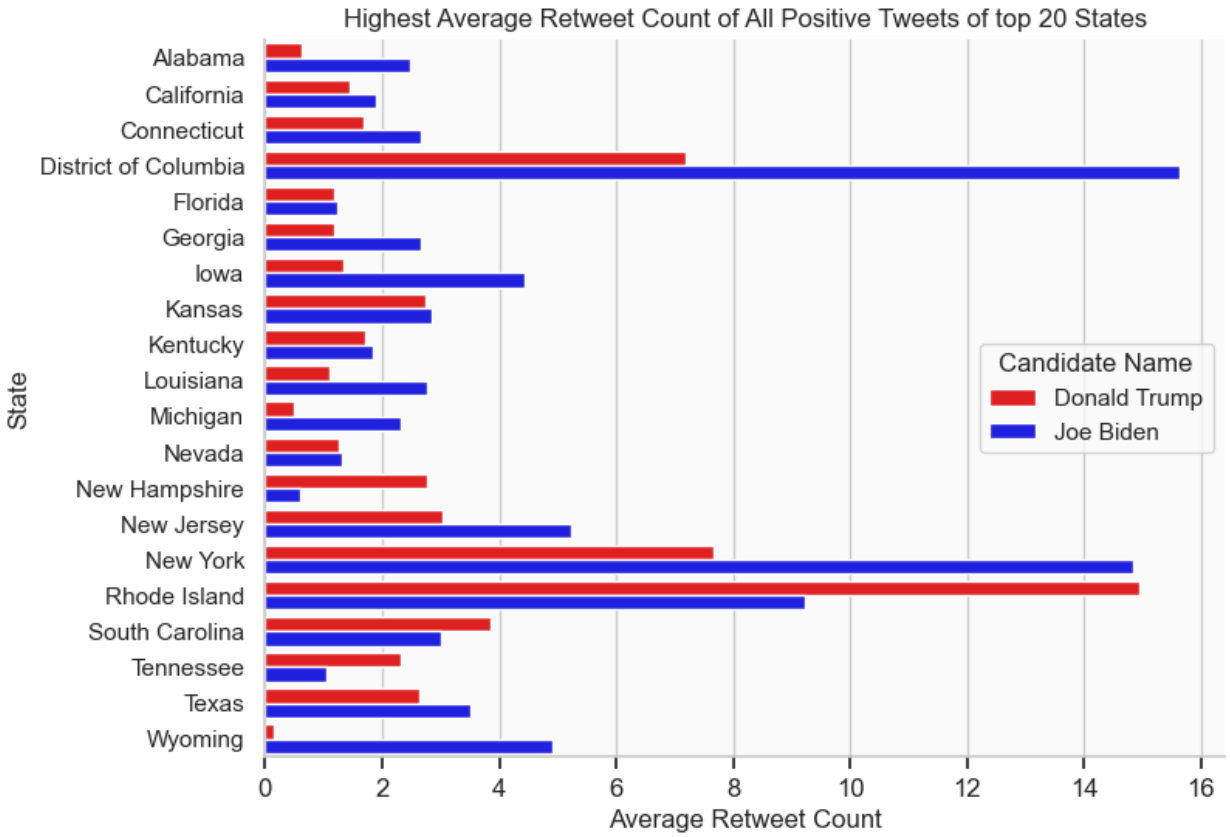


Number of Retweets

The following table shows information regarding retweets of the tweets that mentions the election candidates, Joe Biden and Donald Trump. Firstly, the number of observations for Biden is much lower than Trump. Yet, we see that the mean retweet for Joe Biden is higher than Donald Trump. This is because there are more retweets of the tweets with Biden (267,398) mentioned compared to Trump (212,546). Moreover, the maximum number of retweets on a single tweet relating to Biden is 3 times larger than of the maximum from the tweets which mentions Trump. This significance further shows to be true since the standard deviation of the retweets of Biden is double of that of Trump's, meaning that the retweets are more spread out and much far from the mean when compared to Trump's retweet data. This shows high variability of retweet counts in tweets relating to Biden. It is evident through the retweets that Biden's message and campaign is circulating more.

Candidate Name	count	mean	std	min	25%	50%	75%	max
Donald Trump	104155.0	2.040670	42.554416	0.0	0.0	0.0	0.0	5986.0
Joe Biden	76632.0	3.489378	85.341246	0.0	0.0	0.0	0.0	17652.0

We observe that New York and District of Columbia has the highest number of average retweet count for Joe Biden's positive tweets from the graph below. It is not a coincidence that Joe Biden in fact did win in New York and the District of Columbia. Similarly, Trump won in South Carolina where he had a higher average number of retweets on his tweets that showed a positive sentiment.

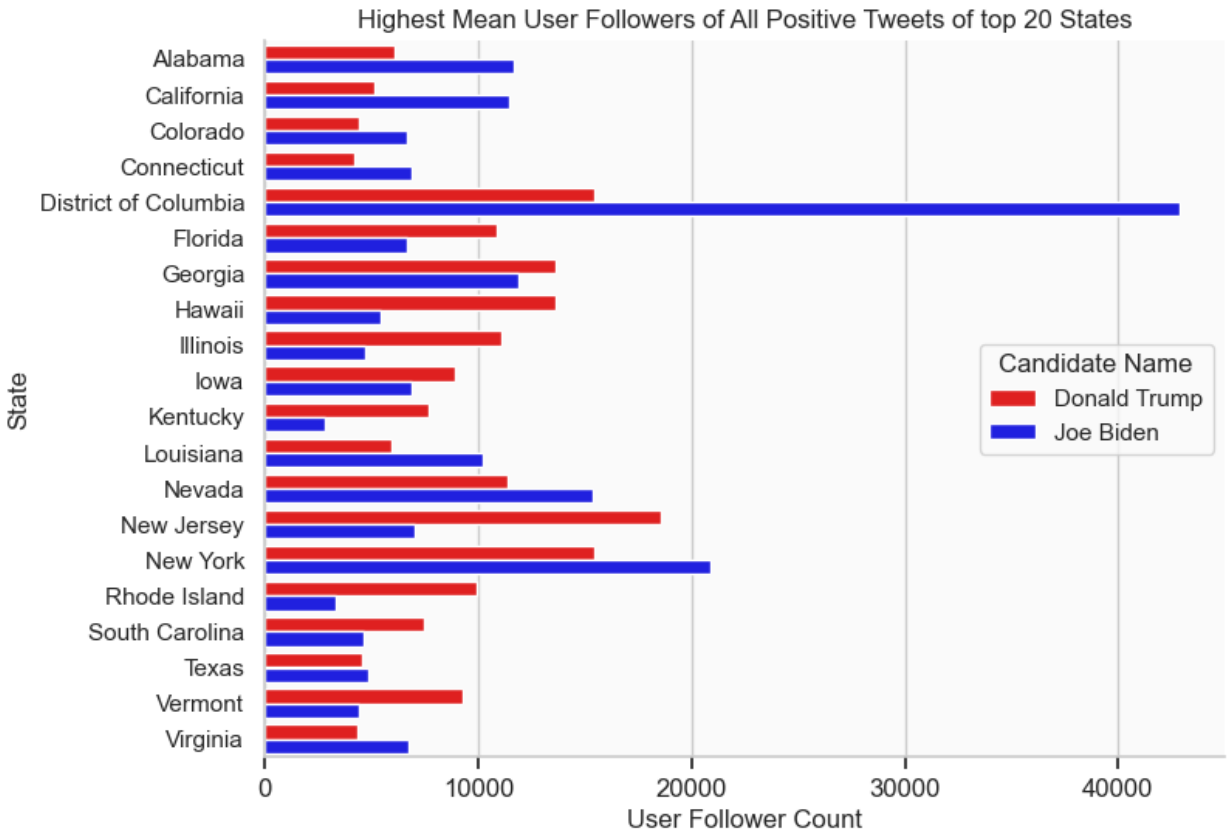


Number of User Followers

People who tweeted about Biden have a higher average following on Twitter than the users who tweeted about Trump which can be observed from the table below. Although user with Trump tweets has a higher number of total followers, there is one specific user who has the highest follower count that tweeted about Biden that was higher than any person who tweeted about Trump with about 5.7 million followers. The inter-quartile range (IQR) for trump user follower is 2548 while Biden's IQR is 2869. Biden's higher IQR indicates the higher dispersion of followers count which can also be observed by a higher standard deviation of followers of Biden supporters.

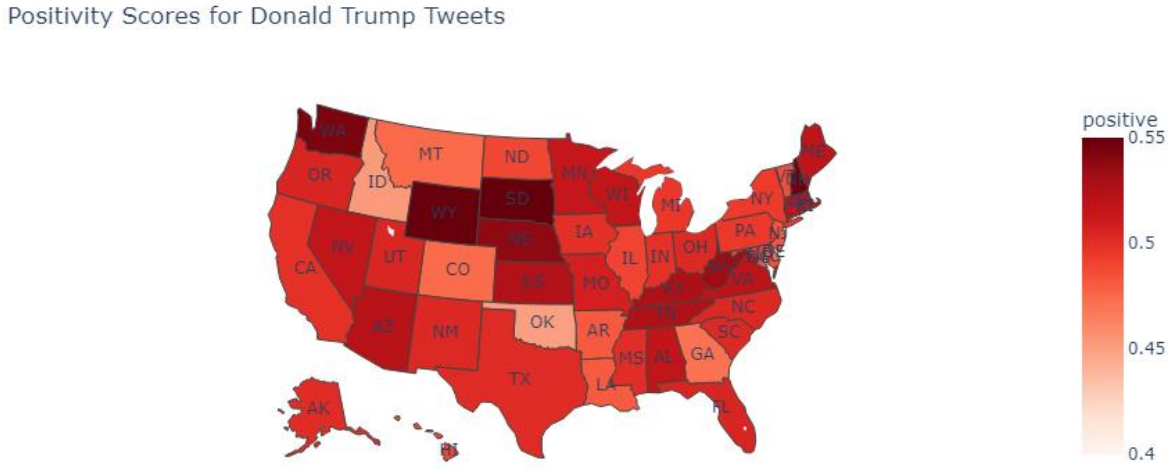
Candidate Name	count	mean	std	min	25%	50%	75%	max
Donald Trump	104155.0	8474.653132	81261.428768	0.0	138.0	694.0	2685.0	4163175.0
Joe Biden	76632.0	9929.758913	87360.647876	0.0	143.0	789.0	3020.0	5750841.0

We observe a similar trend from the plot below particularly regarding New York and District of Columbia which align with our previous observations. However, we see that New Jersey has a lot of user followers with positive tweets yet it was Biden who won the election in New Jersey. Note that there is an average of 40000 twitter user followers for Biden positive tweets in District of Columbia where he won.



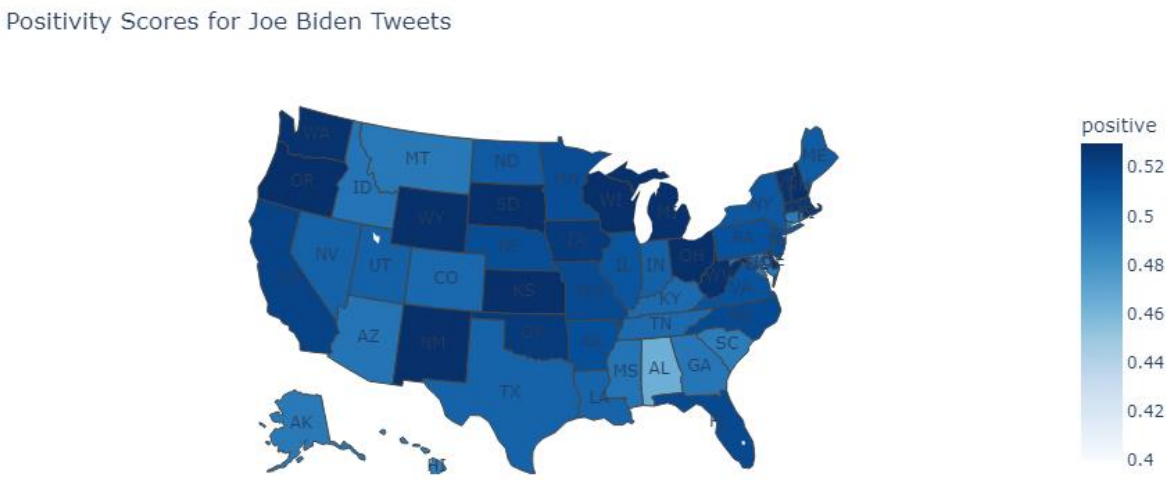
Positivity Scores for Donald Trump

The following map shows the mean positivity score distribution for all the positive tweets of Donald Trump in the states. We observe that the higher the positivity score, the darker the region is (dark red). In Trump's case, we see Texas, Nebraska, Kansas, South Dakota and Alaska has a high positivity score for Donald Trump (positivity score above 0.5). This clearly indicate that he should win in the aforementioned states which he actually did. There were a few outlier states which are contradictory to our results for example the state of Nevada, Wisconsin, Michigan and Virginia where Biden won, but Trump did receive positive sentiment tweets from these states. Note that the outlier states were battle ground states meaning both the candidates had an equal chance of winning in that state and the competition was not one-sided. Similarly, as we would expect a low positivity score for Trump would essentially indicate that he should lose in those states for example we see Georgia, Washington and New York with low positive scores where he indeed lost.



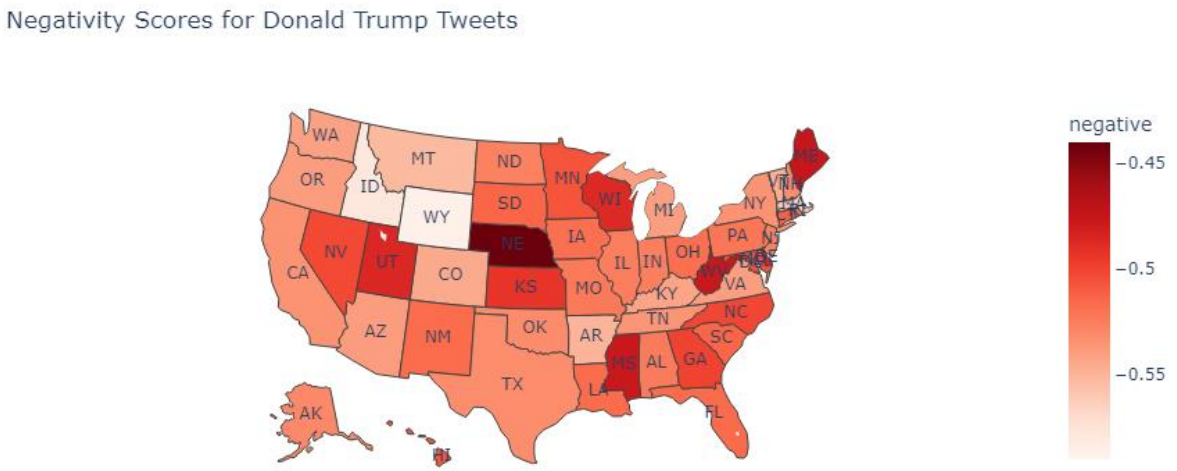
Positivity Scores of Joe Biden

The following map below represents the mean positivity score distribution for all the positive tweets of Joe Biden in the states. We observe that the higher the positivity score, the darker the region is (dark blue). In Biden's case, we see Washington, Colorado, Wisconsin and New Jersey has a high positivity score for Joe Biden where he indeed was victorious (positivity score ranging between 0.48 and 0.52). There are a few outlier states that are contradictory to the actual outcomes for example the state of Iowa and Wyoming where Trump won, Biden received high positive sentiment tweets in these states. Similarly, as we would expect a low positivity score for Biden would essentially indicate that he should lose in those states for example we see Florida, Utah and Ohio with low positive scores where he lost. There are a few contradictions as well like in Nevada and Virginia where Biden received very few positive sentiments, yet he won there.



Negativity Scores of Donald Trump

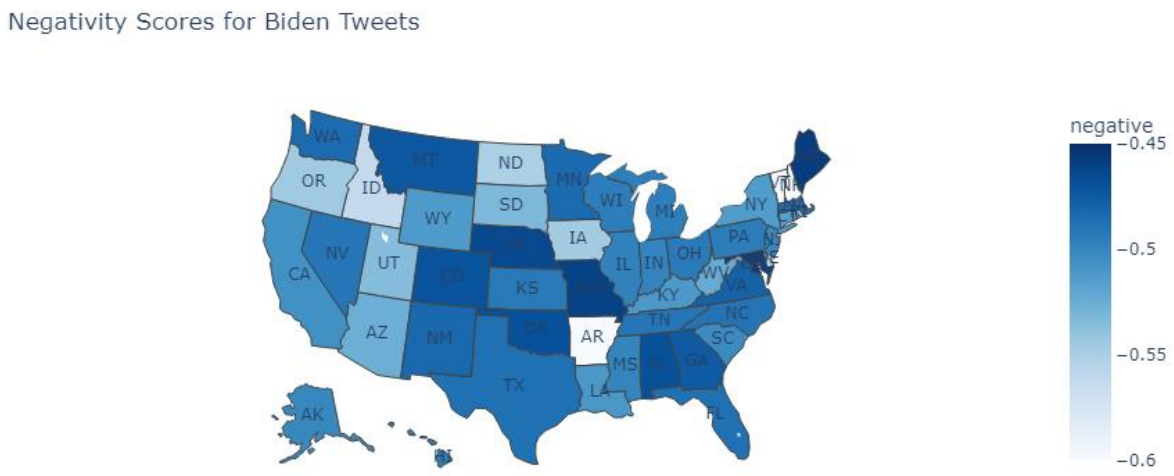
The map below shows the distribution of the mean negative score for all negative tweets of Donald Trump by state. The negativity scale can be interpreted as the more the negative the lighter the region would be, light red in this case (i.e., a score of -0.52 would be considered strong negative sentiment). The states which received low negative sentiment by Trump were Kansas, Kentucky and West Virginia which is consistent with the election results since Trump won in all the said states. There are a few contradictions like Texas and Mississippi where Trump received strong negative sentiment tweets yet he won there. Nevertheless, a low negative score can indicate whether people disliked that candidate and so if they could win there or not in the respective states.



Negativity Scores of Joe Biden

The following code outputs the map of the mean distribution of the negativity scores for all the negative tweets received by Biden by state. Biden received strong negative sentiment tweets in Idaho, North Dakota and Arkansas where we would expect him to lose where he indeed lost. Similarly, he received very low negative sentiment scores in Washington and Oregon which

explains why he won there. There are a few contradictions like Kansas where Biden received low negative sentiments yet, he lost. This is consistent with our previous map where Trump received high positivity scores in Kansas in particular. Thus, once we compare the graphs accordingly, we can see the relative scores between each candidate has a major impact on who would win in that state.



Population and Votes: The Truth of Voter Turnout

Alongside the twitter data set that reflects the tweets of the two primary candidates, there is a possibility to explore a new relation between the proportion of people who tweeted about the respective candidate on Twitter and what was the proportion who actually voted on the election day. To analyze this relationship, we would require the population data of each state. For that purpose, we scraped and incorporated the data set that contained the population size of each state. Although we had millions of tweets in our datasets, there was a void to answer over how much the tweets were representative of the votes received by each candidate by the people.

Moreover, an analysis with the population parameter would put further emphasis on how data from Twitter could be impactful and act as a reliable source to predict the presidential election

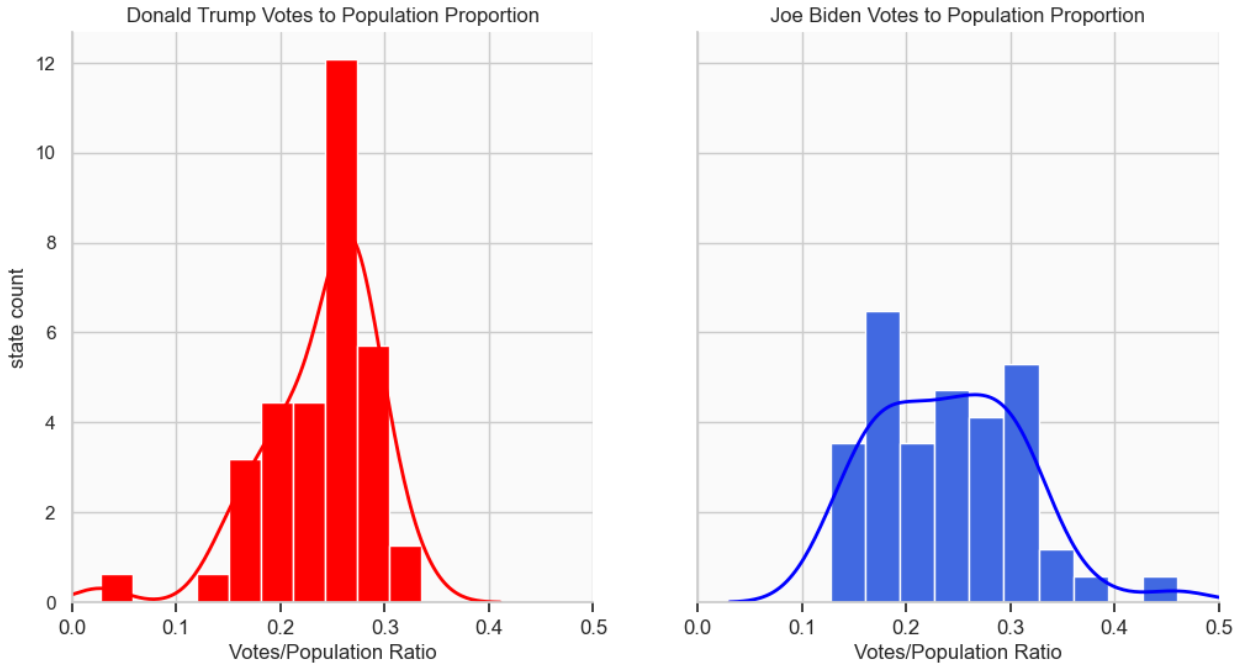
outcomes. If only, we could observe a similar ratio of the people who tweeted in a state for a candidate also voted for their favored candidate, we could conclude with concrete evidence that the twitter data indeed is reliable for predicting election results. The main idea of these calculations would be to demonstrate the public engagement on twitter and compare it to the engagement people show by moving forward in their ideas by voting their preferred candidate on the election day.

The table below indicates another factor why Biden won in the 2020 presidential election. The mean proportion of the vote to population ratio, also known as the voter turnout rate, of Biden is greater than Trump which indicates that Biden received more votes for each person in a state. Similarly, the standard deviation of Trump is lower compared to Biden's which means that Trump only got high number of votes in specific states while Biden had a higher proportion of votes with respect to population in the majority of the states. The maximum proportion of Biden is also higher than Trump.

Candidate Name	count	mean	std	min	25%	50%	75%	max
Donald Trump	51.0	0.239557	0.055079	0.026954	0.20420	0.258043	0.270110	0.335544
Joe Biden	51.0	0.240775	0.070610	0.127266	0.18263	0.236887	0.290132	0.460192

The histograms below visualize the proportions of each Candidate. It reaffirms the idea that Trump received a similar proportion in a lot of states while Biden received a varied proportion in the states. Since, Biden received a higher proportion (more than 0.3) in a lot of states, he therefore had an advantage of higher number of votes ending up winning the elections. It is important to realize through these graphs that more than half of the population of the USA do not

vote in the elections. Although the population data contains the people below 18 years of age, it is still a very low proportion of people who voted in general in the states.



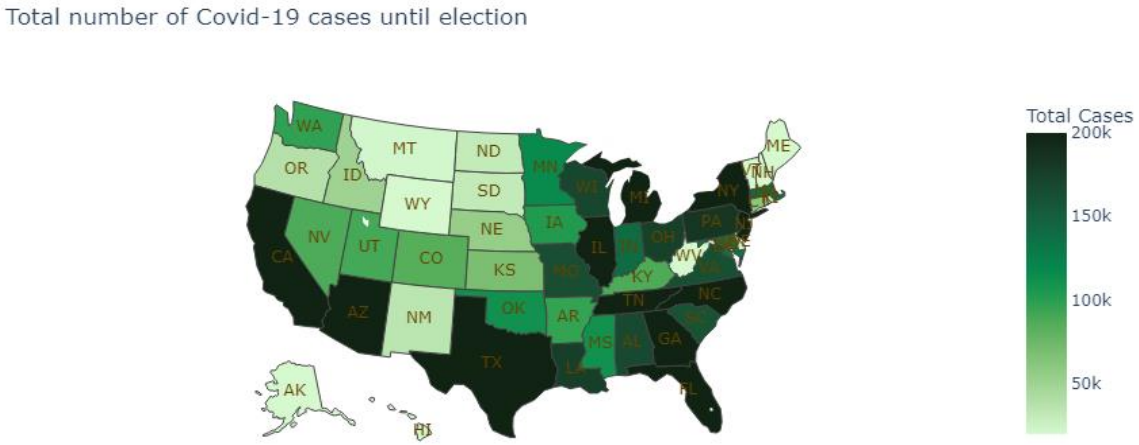
COVID-19 and the 2020 Presidential Elections

The Elections took place during the biggest global pandemic of the 21st century, covid-19. The pandemic reshaped the lives of every person in the world in some form or another. It was important to view the elections from a different lens and observe whether and if it affected the elections either in support of Biden or against Trump. After some external research regarding COVID-19 in the United States I saw that Biden had much favorable policies related to health (Khan, 2021). Biden promised better healthcare reforms compared to Trump so I hypothesized that the states that had a higher number of covid cases ended up supporting Biden. We found out

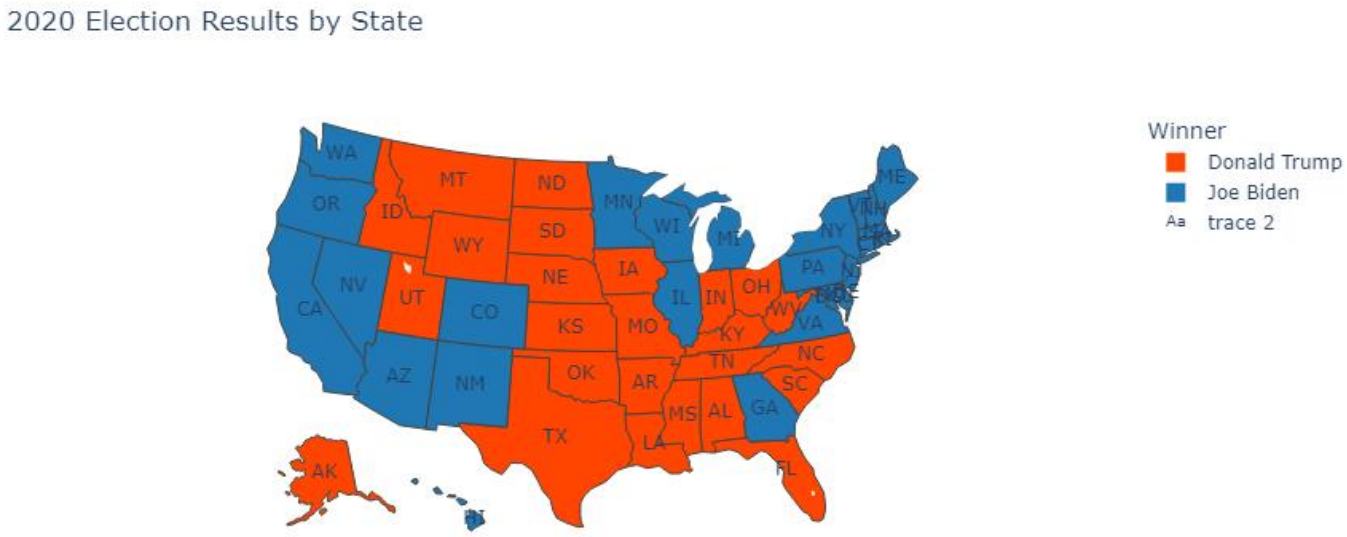
the data set that contains the covid-19 statistics of USA. We had access to a detailed data set that contained about all the information about covid related parameters including total number of cases and deaths etc. on state level which was updated every day till the election date.

The following code below shows a heat map with the covid cases of all states. We see that the states with high number of cases (dark green regions) Biden received a high average positive tweet sentiment as well as the votes population proportion. States like California, Pennsylvania and New York received high average positive sentiment as well as high total cases and so, Biden ended up winning in the aforementioned states.

The following map represents the number of cases in the USA from the first case confirmed on 21st January till the election date of 3rd November 2020. The states that had about more than 125,000 cases by the time of election, we could observe a pattern of Biden winning in those states like Colorado, Nevada, California and Georgia etc. It gives us the idea that Biden's favorable health policies were giving him an advantage in his campaign because of the presence of covid during the elections. Moreover, referring back to the positive score distribution map of Biden, it follows a much similar pattern with a few states overlapping in terms of high positive sentiment and high number of cases respectively.



For a quick reference to the states in which each candidate won and to confirm the previous conclusions we have the map for the election results below.

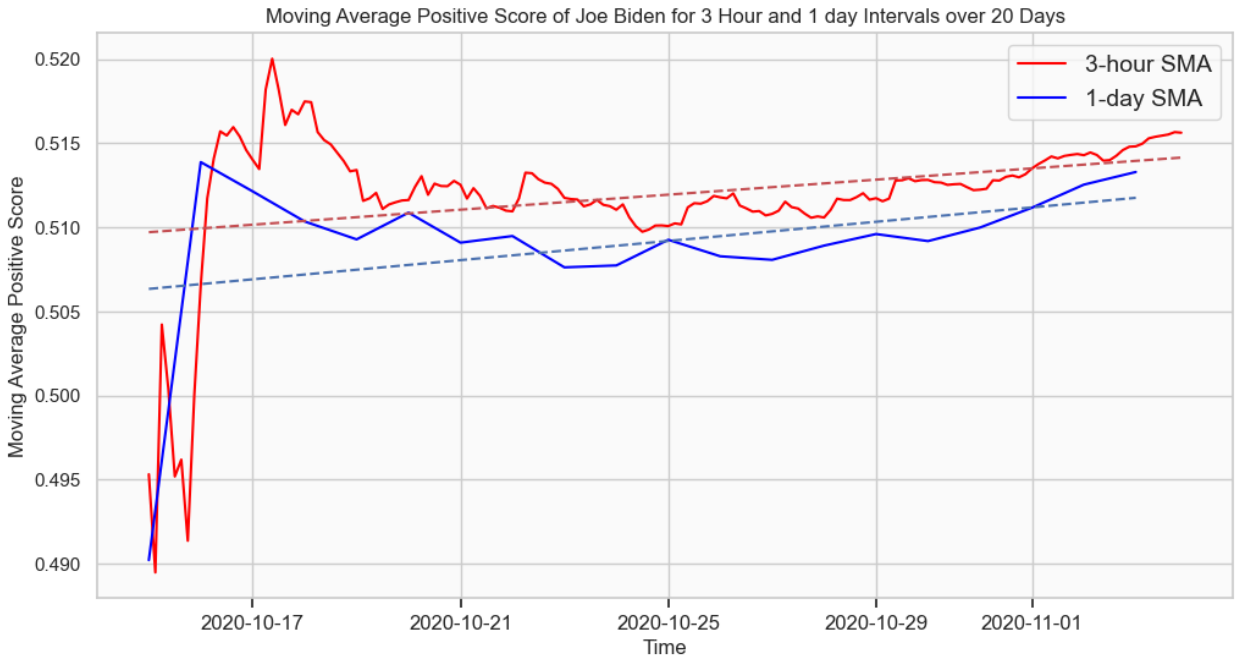


Time Series Analysis: The Impact of News

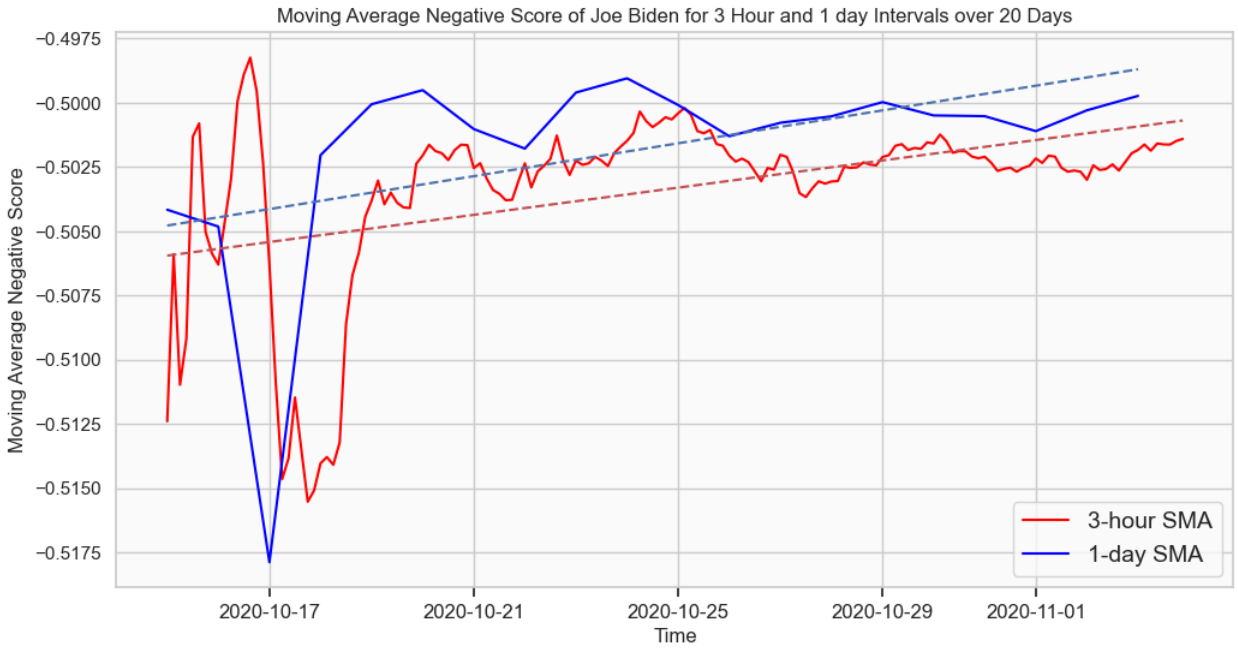
In order to observe how the average positive and negative sentiment changed throughout the election period for the candidates we used simple moving average time series graphs of the tweet sentiment score throughout the 20 days of the tweet data available from October 15th to November 3rd of 2020. The time-series graphs below are simple moving average of the positive and negative sentiment scores for Biden and average positive and negative sentiment score for Trump. The graphs contain two different time-windows, a 3 hour and a 24 hour one. A 3-hour time window means that the moving average is calculated based on the data from the previous 3 hours, while a 24-hour window means that the moving average is calculated based on the data from the previous 24 hours. Note that the averages are of the positive (or negative) sentiment score of the tweets of those previous hours. The main difference between using a 3-hour time window and a 24-hour window is the level of sensitivity and responsiveness to changes in the

data. A 3-hour window will respond more quickly to short-term fluctuations in the data, while a 24-hour window will be slower to respond to short-term fluctuations and will provide a smoother representation of the overall trend.

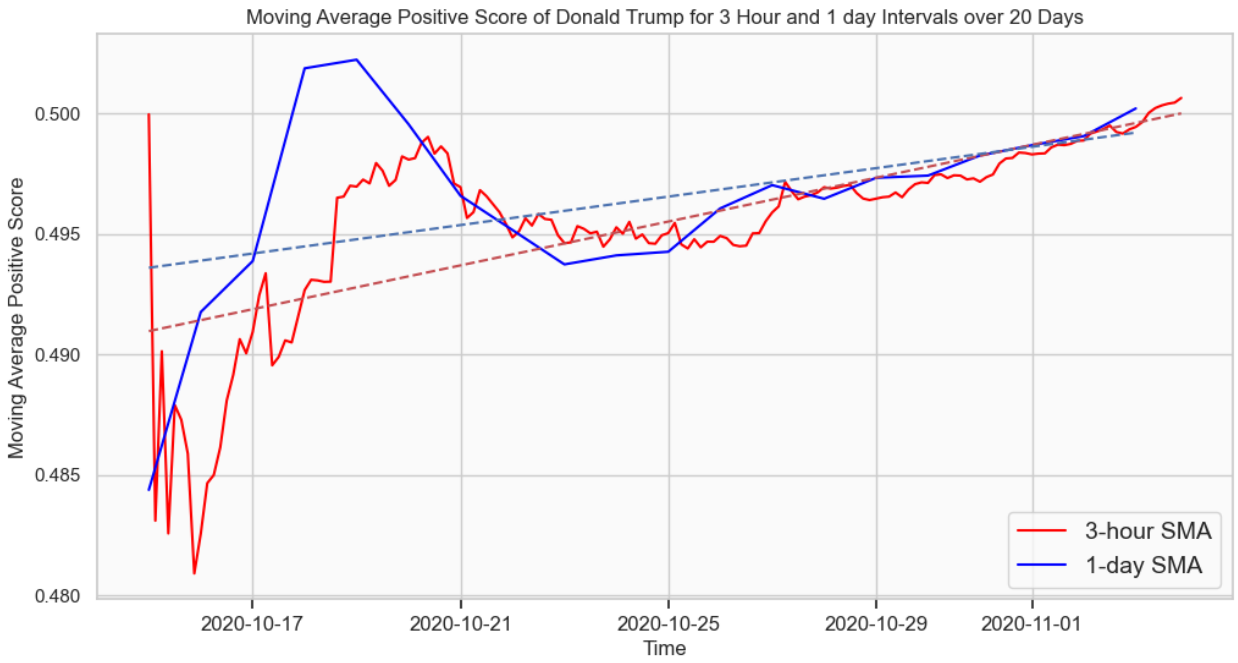
The graph below is the time-series graph that shows a distribution of the positive tweets of Biden throughout USA for the 20 days prior to the elections.



Note that the overall trend for the positive sentiment score up to the election date for Biden is increasing and positive for both time windows along with fluctuations. This is a good indicator of the overall sentiment change and how as the elections date was near people throughout USA were optimist about Biden being the next President. Similarly, we also saw the changes of the negative sentiment scores of Biden as seen below.

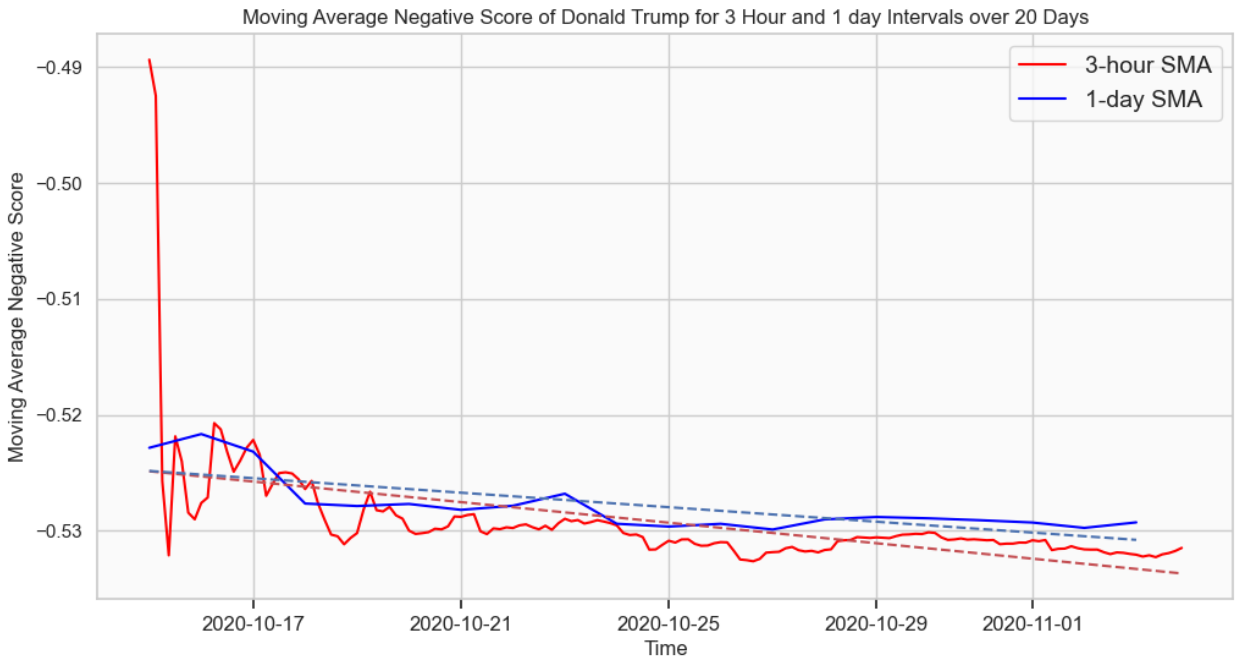


We observed in the 1-day moving average that there was a huge spike in the average negative score for Biden between October 15th to 17th. In particular, twitter put an end to blocking users who shared about an unconfirmed article from the New York Post about former Vice President Joe Biden's son Hunter on October 15th. The article discussed about serious alleged emails between a Ukrainian energy executive and Hunter Biden which became a controversy since, it was perceived against the interest of US (Morris, 2020). This also had an impact on Trump's campaign positively as seen below for his positive tweets.



The positive score for Trump during the said time period of 3 days also saw a huge positive spike when the controversy against Biden's campaign occurred. Unfortunately, Trump's tweets were not able to retain the positive score up to the election date.

In fact, the overall trend for the negative sentiment scores up to the elections for Trump is negative and increasing for both time windows which indicates that as the election came closer, the negative sentiment was increasing for Trump on Twitter. This makes sense as the election came closer Trump was becoming unpopular and the hate towards him was increasing given that the negative scores were increasing (negative slope as moving towards election). This aligns with the overall election results where he lost in the majority of the states.



Linear Regression Results

The implications of data from twitter have a far outreach when it comes to seeking relationships between how the real-world elections and votes are representative of the number of tweets each candidate received in the states. To explore such relationships, we first ran regressions to observe relationship between the number of tweets to the number of likes, retweets and user follower counts for each positive and negative sentiment for each candidate. The general multiple regression equation for the following regressions is the following;

tweet_count = $\beta 0$ + $\beta 1$ * retweet_count + $\beta 2$ * likes + $\beta 3$ * user_followers_count + ϵ

These regressions were calculated to observe how tweet count is affected by the engagement variables and then we finally interpret how impactful are the tweet counts with respect to the number of votes each candidate received.

The mean square error for the above OLS regression equation is;

$$\frac{1}{N}\sum_{i=1}^{N}\left((\text{tweet\_count}_i) - (\beta_0 + \beta_1 \text{retweet\_count}_i + \beta_2 \text{likes}_i) - \beta_3 \text{user\_followers\_count}_i)\right)^2$$

The objective is to estimate the values of the coefficients that minimize the sum of the squared errors between the actual and predicted values of the dependent variable. This can be expressed as the following function:

$$min_{\beta 0,\beta 1,\beta 2,\beta 3}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Positive tweets for Donald Trump

The table below shows the results of an Ordinary Least Squares (OLS) regression analysis with the dependent variable 'tweet_count' and three independent variables: 'likes', 'retweet_count' and 'user_followers_count'. It reports the coefficients, standard errors, and p-values of the estimated regression coefficients etc.

Model 1 includes the independent variables, constant term and retweet count, which has a statistically significant positive coefficient of 11.774 with a standard error of 0.198. This means that the predicted tweet count would be 11.774 when all other independent variables are equal to zero. the coefficient for 'retweet_count' is also positive and statistically significant at the 1% level, indicating that a one-unit increase in 'retweet_count' is associated with a 0.011 increase in the predicted tweet count.

In Model 2, 'likes' is added as independent variables along with 'retweet_count'. The coefficient for 'likes' is positive and statistically significant at the 1% level ($p<0.01$), which means that a one-unit increase in 'likes' is associated with a 0.005 increase in the predicted tweet count, holding all other independent variables constant.

In Model 3, 'user_followers_count' is added as another independent variable. The coefficient for 'user_followers_count' is positive and statistically significant at the 1% level, implying that a one-unit increase in 'user_followers_count' is associated with a 0.000 increase in the predicted tweet count, holding all other independent variables constant. Additionally, 'retweet_count' becomes statistically insignificant at the 1% level, while 'likes' remains significant at the 1% level.

The R-squared value measures how well the model fits the data. In Model 1, the R-squared value is 0.020, which means that the constant term explains only 2.0% of the variation in the dependent variable. In Model 2, the R-squared value increases to 0.034, indicating that 'likes' and 'retweet_count' explain an additional 1.4% of the variation. In Model 3, the R-squared value increases significantly to 0.102, indicating that the addition of 'user_followers_count' explains a further 6.8% of the variation.

In summary, the regression analysis suggests that 'likes', 'retweet_count', and 'user_followers_count' are significant predictors of the dependent variable 'tweet_count'. The model with all three independent variables has the highest predictive power, explaining 10.2% of the variation in 'tweet_count'.

The OLS regression line equation for the regression is;

trump_positive_tweet_count = 10.871 - 0.009 * retweet_count + 0.003 * likes + 0.000 * user_followers_count + ϵ

OLS Regression for Positive Tweets of Trump

	Dependent variable:tweet_count		
	(1)	(2)	(3)
const	11.774***	11.794***	10.871***
	(0.198)	(0.197)	(0.200)
likes		0.005***	0.003***
		(0.001)	(0.001)
retweet_count	0.011***	-0.010***	-0.009***
	(0.001)	(0.004)	(0.003)
user_followers_count			0.000***
			(0.000)
Observations	2,880	2,880	2,880
R^2	0.020	0.034	0.102
Adjusted R^2	0.020	0.033	0.101
Residual Std. Error	10.414 (df=2878)	10.343 (df=2877)	9.973 (df=2876)
F Statistic	59.693*** (df=1; 2878)	50.795*** (df=2; 2877)	109.099*** (df=3; 2876)
Note:			*p<0.1; **p<0.05; ***p<0.01

Negative Tweets for Donald Trump

The OLS regression results below are for a multiple linear regression model with tweet_count as the dependent variable and likes, retweet_count, and user_followers_count as the independent variables of the negative tweets of Donald Trump.

The coefficients of determination (R-squared) in the three models are 0.017, 0.019, and 0.090, respectively, which indicates that the models explain only a small proportion of the total variation in tweet_count. The regression coefficients for the constant term are statistically significant in all three models, with p-values less than 0.01, indicating that there is a statistically

significant intercept term even when all the independent variables are zero. The coefficient for likes is statistically significant at the 5% level ($p<0.05$) in model 2, but not in model 3. The coefficient is positive, which indicates that as likes increase, tweet_count increases as well in model 2.

The coefficient for retweet_count is statistically significant at the 1% level ($p<0.01$) in model 1 and statistically significant at the 5% level ($p<0.05$) in model 3, but not in model 2. The coefficient is positive in all three models, indicating that as retweet_count increases, tweet_count increases as well. The coefficient for user_followers_count is statistically significant at the 1% level ($p<0.01$) in model 3, and the coefficient is zero, indicating that user_followers_count has no relationship with tweet count in this model. The F-statistic for each model is statistically significant at the 1% level ($p<0.01$), indicating that there is strong evidence that at least one of the independent variables is related to the dependent variable.

The OLS regression equation for this model of the multiple linear regression would be;

trump_negative_tweet_count = 12.079 + 0.005 * retweet_count - 0.001 * likes +
0.000 * user_followers_count + ϵ

OLS Regression for Negative Tweets of Trump

	Dependent variable:tweet_count		
	(1)	(2)	(3)
const	12.998***	12.999***	12.079***
	(0.213)	(0.212)	(0.214)
likes		0.001**	-0.001
		(0.001)	(0.001)
retweet_count	0.007***	0.002	0.005**
	(0.001)	(0.002)	(0.002)
user_followers_count			0.000***
			(0.000)
Observations	2,880	2,880	2,880
R^2	0.017	0.019	0.090
Adjusted R^2	0.017	0.019	0.090

We conducted similar regression for both positive and negative sentiment of tweets that mentioned Biden.

Positive Tweets of Joe Biden

This table below shows the results of the Ordinary Least Squares (OLS) regression for positive tweets of Biden. The dependent variable is the count of positive tweets of Biden, and the independent variables are likes, retweet_count, and user_followers_count as done before.

Model 1 includes the constant term and has a coefficient estimate of 9.297, indicating that the predicted value of the dependent variable is 9.297 when all independent variables are zero. It also contains the retweet count where the coefficient estimate is statistically significant at the 1% level and positive (0.003), which means that an increase in the number of retweets is associated with an increase in the count of positive tweets of Biden.

Model 2 includes likes and retweet_count as independent variables. The coefficient estimate for likes is statistically significant at the 1% level and negative (-0.002), which means that an increase in the number of likes is associated with a decrease in the count of positive tweets of Biden. The coefficient estimates for retweet_count is statistically significant at the 1% level and positive (0.016), which means that an increase in the number of retweets is associated with an increase in the count of positive tweets of Biden.

Model 3 includes all three independent variables, and the coefficient estimate for user_followers_count is statistically significant at the 1% level and positive (+ 0.000), indicating that an increase in the number of followers is associated with an increase in the count of positive tweets of Biden (although a very small one).

The R-squared values in each model indicate the proportion of variation in the dependent variable that is explained by the independent variables. Model 3 has the highest R-squared value (0.087), indicating that the independent variables in that model explain more of the variation in the dependent variable than the independent variables in the other two models. However, the R-squared values are all relatively low, indicating that there may be other important factors that are not captured by the independent variables in these models.

Thus, the results suggest that an increase in the number of likes is associated with a decrease in the count of positive tweets of Biden, while an increase in the number of retweets and followers is associated with an increase in the count of positive tweets of Biden. However, the explanatory power of the models is relatively low, indicating that other factors may also influence the count of positive tweets of Biden.

The equation for this regression would be as follows;

biden_positive_tweet_count = 8.371 + 0.004 * retweet_count - 0.000 * likes + 0.000 * user_followers_count + ϵ

OLS Regression for Positive Tweets of Biden

	Dependent variable:tweet_count		
	(1)	(2)	(3)
const	9.297***	9.062***	8.371***
	(0.205)	(0.208)	(0.207)
likes		-0.002***	-0.000*
		(0.000)	(0.000)
retweet_count	0.003***	0.016***	0.004*
	(0.001)	(0.002)	(0.002)
user_followers_count			0.000***
			(0.000)
Observations	2,880	2,880	2,880
R^2	0.011	0.023	0.087
Adjusted R^2	0.011	0.022	0.086

Negative Tweets for Joe Biden

The following OLS regression results table shows the relationship between negative tweets about Joe Biden (the dependent variable) and the three independent variables, the number of likes, the number of retweets, and the number of followers of the tweets in the specified time period.

The coefficient for the number of likes in Model 2 is negative and statistically significant at the 1% level. This suggests that, on average, for every additional like, there is a decrease in negative tweets about Biden by 0.002. This finding may be interpreted as negative feedback, which tends to discourage individuals from sharing negative tweets about Biden. Moreover, in Model 2, the coefficient for the number of retweets is positive and statistically significant at the 1% level. This

implies that for every additional retweet, there is a predicted increase in negative tweets about Biden by 0.017. This positive relationship is a concerning trend, as it suggests that negative tweets about Biden can spread quickly and widely on Twitter.

In Model 3, the coefficient for the number of followers is positive and statistically significant at the 1% level. This relationship is very weak but statistically significant, indicating that users with more followers may have a slightly stronger impact on the number of negative tweets about Biden.

The R-squared value for each model is relatively low, indicating that the independent variables explain only a small proportion of the variation in the dependent variable. The F-statistic for each model is statistically significant, which means that at least one of the independent variables has a significant effect on the dependent variable.

Overall, these results suggest that the number of retweets is the most important predictor of the number of negative tweets about Joe Biden, while the number of likes and followers have weaker and less consistent impact on the outcome variable. The equation for this OLS regression is as follows;

biden_negative_tweet_count = 8.371 - 0.003 * retweet_count - 0.000 * likes + 0.000 * user_followers_count + ϵ

OLS Regression for Negative Tweets of Biden

	Dependent variable:tweet_count		
	(1)	(2)	(3)
const	7.328***	7.141***	6.478***
	(0.193)	(0.195)	(0.189)
likes		-0.002***	-0.000
		(0.000)	(0.000)
retweet_count	0.004***	0.017***	-0.003
	(0.001)	(0.003)	(0.003)
user_followers_count			0.000***
			(0.000)
Observations	2,880	2,880	2,880
R^2	0.010	0.019	0.114
Adjusted R^2	0.010	0.018	0.114

Overall the regression results are statistically significant and show that the engagement variables such as number of likes, retweets and user follower counts have some sort of impact on the number of tweets. The number of negative and positive tweets of Biden and Trump do have some relationships with the amount of engagement.

Tweet Count and Total Votes

Now we conduct the regressions between the number of tweets and the number of votes for each candidate in all the states to observe if the people who actually tweeted about the candidate voted for them or not.

This OLS regression analysis below shows the relationship between the total number of tweets and the total votes in a state during the election for the tweets relating to Trump. The coefficient estimate for the total tweet count variable is positive and statistically significant at the 1% level.

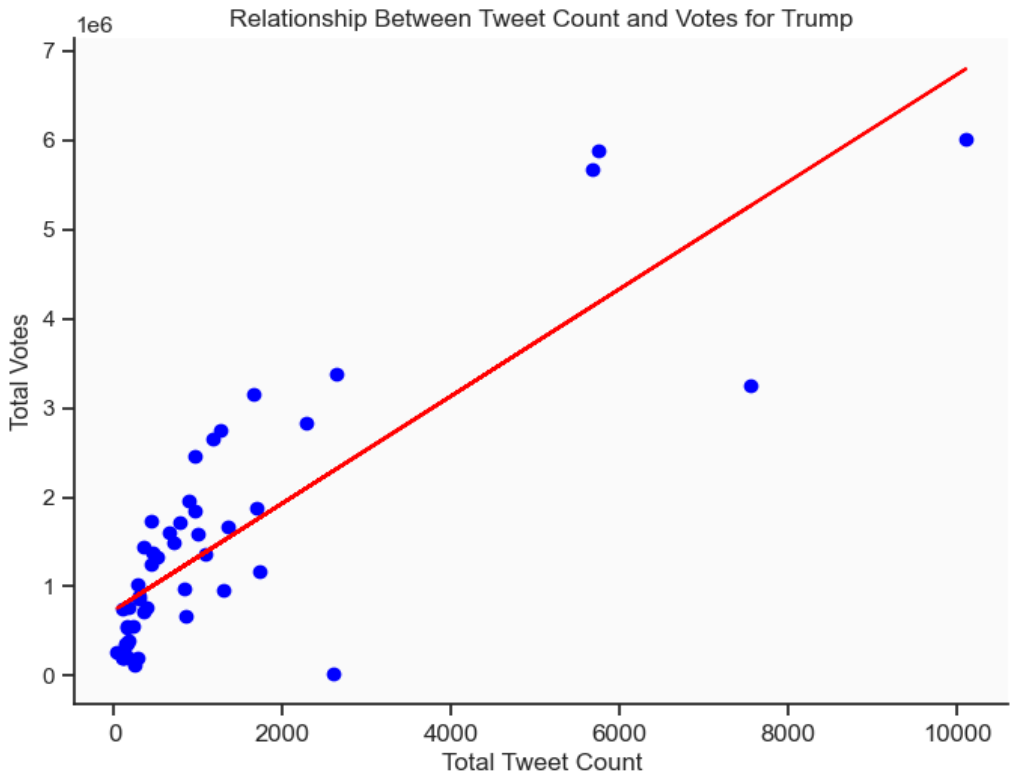
This implies that an increase in the total number of tweets about the election in a state is associated with an increase in the total number of votes cast in that state. The coefficient estimate suggests that a one-unit increase in the total tweet count is associated with an increase of 769.617 votes.

The R-squared value of 0.690 indicates that approximately 69% of the variation in the total number of votes can be explained by the total number of tweets. The adjusted R-squared value of 0.690 suggests that the model has a good fit with the data, even after adjusting for the number of independent variables. The high value of 0.690 suggests a strong positive association between the tweet count and the total votes for Trump in all the states. Overall, the results suggest that social media activity, as measured by the number of tweets related to the election, can be a significant predictor of voter turnout. Higher levels of social media activity do indicate greater engagement and enthusiasm among voters, leading to higher voter turnout.

OLS Regression between Total Votes and Tweets by State

	Dependent variable:total_votes
	Total Votes
	(1)
Total Tweet Count	600.148***
	(57.478)
const	734155.923***
	(131788.495)
Observations	51
R^2	0.690
Adjusted R^2	0.684
Residual Std. Error	798527.431 (df=49)
F Statistic	109.023*** (df=1; 49)
Note:	*p<0.1; **p<0.05; ***p<0.01

The regression plot indeed shows a positive correlation between the total votes and total tweet count for Trump.



A similar regression was conducted for the tweet and votes for Biden for each state.

The coefficient estimate for the total tweet count variable is positive and statistically significant at the 1% level as seen by the OLS regression table below. This implies that an increase in the total number of tweets about the election in a state is associated with an increase in the total number of votes. The coefficient estimate suggests that a one-unit increase in the total tweet count is associated with an increase of 908.367 votes. The intercept is also statistically significant at the 1% level, indicating that even in the absence of any tweets, there is still a significant number of votes for Biden. The coefficient estimate for the intercept term suggests

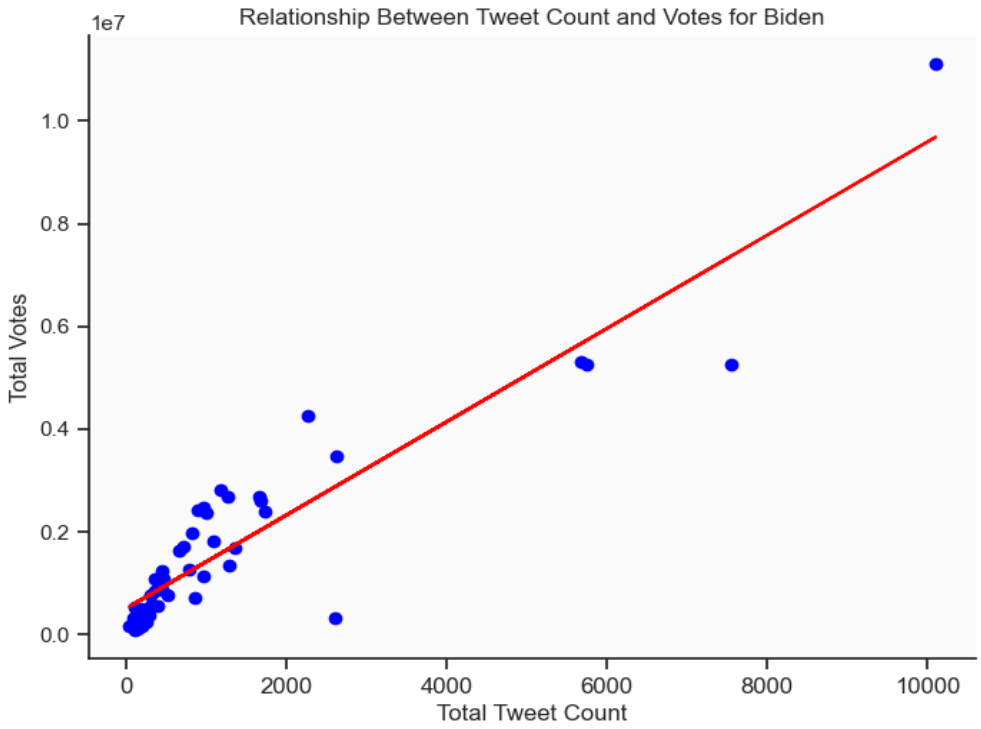
that in the absence of any tweets, the total number of votes in the states is expected to be around 506,405.

The R-squared value of 0.851 indicates that approximately 85% of the variation in the total number of votes can be explained by the total number of tweets. The adjusted R-squared value of 0.848 suggests that the model has a good fit with the data, even after adjusting for the number of independent variables. Note that Biden has a much stronger relation with a much higher R-squared. It cannot be a coincidence that in our data set the number of tweets for Biden for the last 20 days of the election is also very high as mentioned multiple times previously. Moreover, the residual standard error (RSE) of 754,344.517 indicates that the average difference between the actual and predicted number of votes is approximately 754,000.

OLS Regression between Total Votes and Tweets by State

	Dependent variable:total_votes
	Total Votes
	(1)
Total Tweet Count	908.367***
	(54.297)
const	506405.354***
	(124496.573)
Observations	51
R^2	0.851
Adjusted R^2	0.848
Residual Std. Error	754344.517 (df=49)
F Statistic	279.875*** (df=1; 49)
Note:	*p<0.1; **p<0.05; ***p<0.01

The same positive correlation can be observed by the plot for Biden as well.



Conclusion

Measuring the political economy through the lens of social media and computing which political party has a higher share in the online space is crucial in today's world. The research paper pursued building a relationship between different features of a social media post that demonstrates public engagement on social media platforms like tweets on Twitter. The activity of liking and retweeting can significantly reflect people's emotions at a certain time or in a certain region which in turn becomes a good indicator to predict the vote count and ultimately the election results.

Sentiment Analysis helped identify which states in the US were happy to support Biden or Trump. The Statistics tables indicated on average how many people were interested in the campaign and the number of likes and retweet counts each candidate received. Moreover, the

plots gave a visualization of which candidate was provoking more public engagement in certain states. In most of the States Biden was the one with more engagement on Twitter due to a high number of likes and retweets. The histogram provided us with a comparison of how the distribution of vote to population ratio differed between Trump and Biden among the states.

The maps demonstrated how the results of many states were accurately predicted just based on the intensity of the positive or negative sentiment each candidate received in every state. The higher the positivity score or the lower the negativity score the higher the chance for the candidate to win in that state and vice versa. One interesting pattern observed was how COVID-19 played as an external factor in favor of Biden due to his favorable health policies.

The time-series simple moving average models were pivotal to our research which demonstrated how as we came close to the election date the negative sentiment towards Trump and the positive sentiment towards Biden increased while the negative sentiment towards Biden and positive sentiment towards Biden decreased. It was noticeable how a real-life controversy of Biden's son during the election caused a huge upward spike in the negative sentiment towards Biden that caused the positive sentiment toward Trump rise.

Unfortunately, the regressions performed showed a relatively low relation between the number of tweets and the engagement variables like the number of retweets and likes. However, in each subsequent model for all the regressions, as we increased the number of independent variables, it improved our model representing that each engagement variable had some relation even if not strong.

Nonetheless, tweet count had a strong positive relationship with the vote count for each candidate which confirmed that the tweet count was affected by external factors that were

unknown within the data. This meant that further factors such as campaign funding, public rallies during the elections in the battleground states, political ideology, and other social and political factors for each candidate had to be taken into consideration and more data has to be incorporated for better results and predictions. Apart from additional data, more rigorous regression and machine learning models can be implemented to predict which candidate won in each state such as multinomial logistic regression, bootstrapping, random forest, and Naive Bayes classifiers. Such models would be much more appropriate in future research since they use categorical variables present in the Twitter data set. Although there were some contradictions when it came to predicting states for each model we created, we were able to predict the majority of the states correctly through different models and comparisons and show how public engagement of social media posts was a good approach to predicting the political leaning and the election results.

References

Ansari, M. Z. (2019). Analysis of Political Sentiment Orientations on Twitter. *Procedia Computer Science*, 1821–1828.

Coletto, M. (2015). *Electoral Predictions with Twitter: A Machine Learning Approach.* Venice: Research Gate.

He, Y. (2012). *Quantising Opinions for Political Tweets Analysis.* Istanbul: ACL Anthology.

Hui, M. (2020, November 8). *US Election 2020 Tweets*. Retrieved from Kaggle : https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

Joyce, B. (2018). *Sentiment Analysis of Tweets for the 2016 US Presidential Election.* Greensboro: IEEE Xplore.

Khan, A. (2021, August 12). *How the COVID-19 Pandemic Helped Biden Win the 2020 Presidential Primaries*. Retrieved from American Political Science Association: https://politicalsciencenow.com/how-the-covid-19-pandemic-helped-biden-win-the-2020-presidential-primaries/

Matalon, Y. (2021). Using sentiment analysis to predict opinion inversion in Tweets of Political communication. *Scientific Reports*.

Mohammad, S. M. (2014). *Sentiment, Emotion, Purpose, and Style in Electoral Tweets.* Ottawa: National Research Council Canada.

Morris, E.-J. (2020, October 20). *Smoking-gun email reveals how Hunter Biden introduced Ukrainian businessman to VP dad*. Retrieved from NewYork Post:

https://nypost.com/2020/10/14/email-reveals-how-hunter-biden-introduced-ukrainian-biz-man-to-dad/

Pota, M. (2018). A Subword-based Deep Learning Approach for Sentiment Analysis of Political Tweets . *International Conference on Advanced Information Networking and Applications* (pp. 651-656). Krakow: IEEE Xplorer.

Rodriguez-ibanez, M. (2021). *Sentiment Analysis of Political Tweets From the 2019 Spanish Elections.* Madrid: IEEE Access.