Twitter's Impact on Elections

Prem Pathuri
Zhi Chong Chris Lin

February 7, 2021

Abstract

The rise of social media has dominated every aspect of our daily lives. In the height of the 2020 presidential election and as COVID-19 rampaged throughout the world, it facilitated increased online discussion, as well as the spread of information and misinformation. This paper investigates the relationship that discussion on social media has with election outcomes. It finds that in comparing two distinct presidential elections, both of which took place as Twitter usage grew steadily, increased discussion levels were present in a Democratic win of the election.

1 Introduction

1.1 Significance

Social networks facilitate the rapid dissemination of information and content throughout the world as a result of the rise of internet technologies. These platforms allow a person to connect and dialogue with various people, of different creeds, colors, race, sexual orientation, and political ideology. Social media platforms also indirectly support the segregation of people with different viewpoints, as individuals with particular views tend to gather in a cluster and speak more with those that share the same standpoints. As information begins to spiral, it begins to have an influence on people's thoughts, beliefs, and ideologies.

This has become increasingly important in the recent era, most notably during 2016 and 2020 presidential elections. For example, in 2016, Russia was accused of interfering with the election through campaigns on social media. Popular social platforms such as Facebook and Twitter allowed information to spread. According to Donald Trump, these actions definitely contributed to his success in the presidential race. His statement was that without Twitter, he would likely have not won the election. Such a statement was refuted by Twitter's CEO Jack Dorsey, who claimed that Twitter does not have the ability to influence elections. Hence, our study looks into the extent to which election outcome could be influenced by opinions and discussions through online social media platforms.

1.2 Goals

Our goal in this paper is to investigate whether or not social media platforms such as Twitter can have an impact on large-scale national events such as the presidential election in the United States. To do so, we will be comparing activity and discussion levels between two distinct and controversial elections, the 2016 and 2020 elections. The differentiating factor between both of these elections, both of which caused dissension within the country, was the party that won. In 2016, Donald Trump of the Republican party won against Hillary Clinton of the Democratic party and in 2020, Joe Biden of the Democratic party won against Donald Trump of the Republican party who was running for re-election. Thus, if Twitter were to have the ability to potentially impact the outcomes of these elections, there would have been a difference in the amount of activity on Twitter in between the two elections – this investigation is the purpose of this paper. To achieve this outcome, we would produce an interactive visualization of sentiment and level of discussion across time towards the 2016 and 2020 presidential elections. Additionally, We would like to create a tweet-map of the United States for which the user would be able to look at aggregate statistics for states and cities.

2 Related Work

As this has come under scrutiny in recent years, there has been a lot of analysis done on Twitter's ability to have an impact on the outcome of presidential elections. Some studies have researched whether or not social media usage enables right-wing growth or stifle conservative speech. One such study delved into the 2016 election and showed that social media usage on Twitter decreased Donald Trump's vote share by 0.2 percentage points [3]. The study claimed that social media users were typically young, more educated, and live in areas with higher population density and as such, would favor the Democratic party more often than not. This is in contradiction to Donald Trump's own statement, attributing his victory to a substantial level of interaction with constituents via Twitter. On the other hand, another research paper analyzed sentiment on Twitter and showed that Twitter sentiment favored Donald Trump over Hillary Clinton in 2016, showing that perhaps sentiment is more indicative of the outcome [4]. These two studies contradict each other. The former doesn't address the fact that Donald Trump ended up winning in 2016 despite Twitter usage lowering his vote share while the latter claims that since sentiment was positive, Donald Trump, in retrospect, was the clear winner.

This discrepancy is what this paper shall be investigating. In doing so, we will be analyzing the discussion levels on Twitter in both the 2016 and 2020 elections. These elections were similar in that they both had high levels of chatter and discussion on social media platforms, negative stories for both parties, and debated content. We will attempt to see whether or not this discussion

favors the Democratic party or the Republican party in this investigation.

3 Methods

To gauge social media's influence on the outcome of elections, we defined a measure of the activity level on Twitter, which we will referred to as the "level of discussion". Since online interaction can be quantitatively evaluated with the number of likes and retweets for each original tweet, we will use the summary statistics from data collected with the Twitter API. To aggregate a total level of discussion, we will use the following formula to measure activity levels:

$$discussion = \ln(1 + likes + retweets)$$

This formula came from needing to deal with a skewed distribution. Daily discussion levels were incredibly right skewed, so we decided to log-scale them to eliminate skew in the distributions and make analysis easier. In doing so, we needed to add a 1 to each term to ensure that we were not taking the natural log of 0, on days which may have had zero discussion.

As there were numerous tweets per day, the number of likes and retweets were aggregated across each individual day over the election period in 2016 and 2020. These sums were then placed into the above formula and evaluated across time. In doing so, we are able to gauge the level of activity on each specific day and attribute such discussion to related key events throughout the election period.

4 Data

The data that we gathered were collected by two different sources. Our 2016 Presidential Election Tweet dataset originated from Harvard's Dataverse [2]. These tweet IDs were collected using candidate and key-election hashtags as the search query in the Twitter API via the Social Feed Manager feature.

The 2020 election dataset was gathered by graduate students at the University of Southern California [1]. This group recognized the importance of studying chatter on social media leading up to important events such as democratic elections. In order to ease the process of collecting the data for computational research, the group gathered Tweet IDs for each day within the election time frame using keywords and hashtags as their search query. It tracks political trends, events, and key figures within a one year time span. These tweet ID files are hosted on their GitHub repository, from which we downloaded the data using a script that ran cURL commands.

Both of these datasets are integral to our research and this investigation. The 2016 dataset only had tweets collected from July 13, 2016 and November 10,

2016 whereas the 2020 dataset had tweets originating from December 1, 2019 to January 29, 2021. In order to ensure that our investigation wasn't biased, we narrowed our time frame to that of the Harvard dataset so as to be investigating the same window of time in both elections.

4.1 EDA and Sentiment Analysis

To further investigate in our data, we prepossessed the tweets by cleaning non-text features, removing stopwords, and converting to lowercase. Next, we utilized a sentiment analysis to discover how the online conversations related to the election were engaged. We adopted the usage of VADER from the NLTK library and assigned a score on a scale of -1 to 1, with 1 being the most positive end. Finally, we trained this model with our set of original tweets for both the 2016 and 2020 data and the results are shown in fig 1.

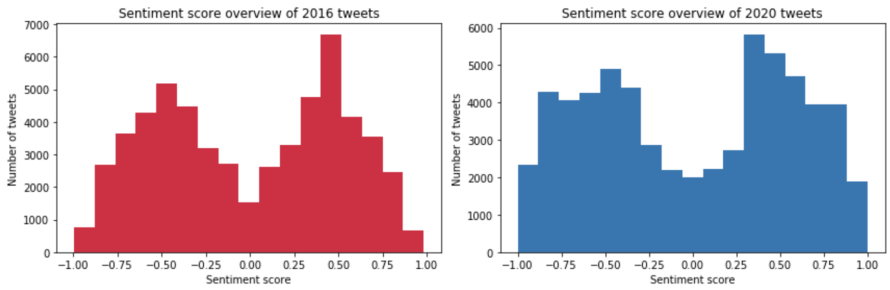

Figure 1: Comparison of Overall Sentiment

We observe a similar bi-modal distribution in both histograms as it is reasonable that the message from each tweet indicates either a positive sentiment or a negative sentiment. However, the distribution is rather normal during the 2016 election period, with the two peaks around -0.5 and 0.5. In the 2020 election period, we observe a more spread out distribution on tweets with a negative sentiment, and a right tailed distribution on tweets with a positive sentiment. This outcome leads to further analysis to determine how different the level of twitter sentiment was between the two election years and how has it created an impact on the election outcome.

Another step we took was to create word clouds that resembled the popular terms used throughout tweets. As shown in fig 2, the most common words are usually related to the candidates of the year - Donald Trump and Hilary Clinton in 2016 and Donald Trump and Joe Biden in 2020. The second tier of common words are generally election related, some of which include 'say', 'thank', 'poll', and 'people', most of which have neutral or positive connotations. Notably in 2020, 'covid' is one of the popular terms, however, the other general terms are similar across the two election years. Hence, our EDA suggests that there are minimal hidden variables that might impact the validity of our model.

Figure 2: Word Clouds

4.2 Daily Sentiment Investigations

After completing the initial EDA, we looked into a time-series sentiment analysis for each of the election periods. During the time-frame of July 13 to November 10 of each year, we plotted the average daily sentiment on a scale of -1 to 1.

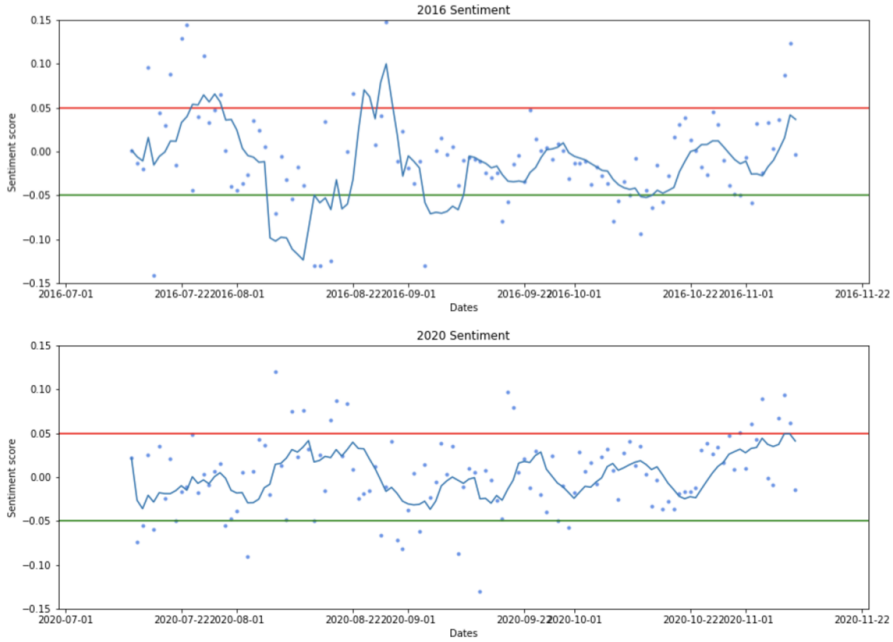

Figure 3: Comparison of Daily Sentiment

The daily sentiment comparison is based on original tweets posted every day. Due to sampling, the number of tweets everyday could vary across time, especially in the months before the election day. Hence, the sentiment score is fixed to a frame of -0.15 to 0.15, with a positive threshold set at 0.05 and a negative threshold at -0.05. A moving average of 7 days is also added to capture the sentiment trends as well as peaks and troughs directly.

In the 2016 plot (fig 3 top), the daily sentiment fluctuates severely through July and August, as discussions about the leaders heated up every once a while. Though fluctuations also exist in the 2020 plot (fig 3 bottom), the magnitude is smaller and generally lies between the positive and negative thresholds. As we approach the election day, the daily sentiment in 2016 dips twice, each corresponding to the presidential debates that occurred in early to mid-October. The days in October with a negative sentiment is likely due to higher levels of online conversation and topical discussion. In the sentiment trend for 2020, a slight dip occurred towards the end of September, matching the date of the first presidential debate in that election year.

As we approach November, the 2020 trend peaked twice, once on election day and again on the day the day President Biden was elected. This differed from the 2016 trend as the single peak occurred only after when the election results were finalized. Hence, the sentiment peaks and troughs throughout the election period has been interconnecting with key election events and the discussion levels are likely to be more concentrated during these dates.

5 Results and Analysis

The initial question was to identify a possible impact that Twitter may have had on the outcome of the presidential elections in 2016 and 2020. In order to to identify if this is true, we tracked daily discussion levels on Twitter during the time frame from July 13 to November 10. Here we plot these distributions.

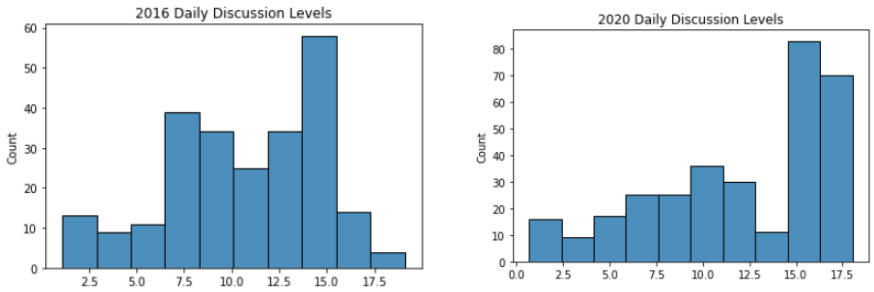

Figure 4: 2016 (left) and 2020 (right) Discussion Levels

As one can see, there is a major discrepancy between the two election periods.

The daily discussion levels in the 2016 election were much more normally distributed in comparison to the 2020 election discussion levels, which was heavily left skewed. This is reflected in their mean discussion levels as well, with the 2016 election having a mean discussion level of 10.68 whereas the 2020 election had a mean discussion level of 12.04.

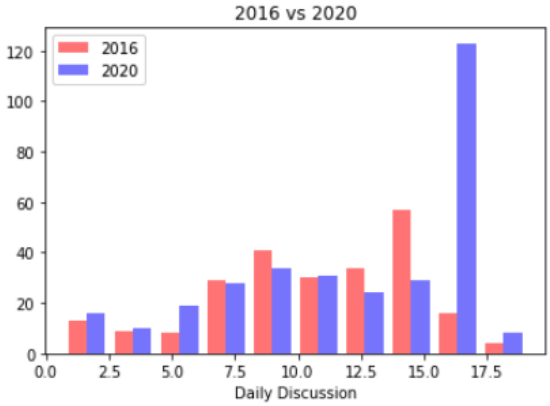

Figure 5: Comparison of Discussion Levels

5.1 Significance Test

As we have noted, the distributions of the discussion levels in 2016 and 2020 were quite different where the discussion levels in 2016 were much more normally distributed than in 2020. However, in order to quantify this difference and verify that it was significant, we conducted a Z test with a difference of means in two samples. To do so, we found the probability of detecting such a difference or greater using the formula:

$$Z = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The test resulted in a Z score of 3.9985 which gave us a p-value of 0.0001346. This passes all significance tests, at the 0.1, 0.05, and 0.05 levels, showing that the difference in distributions observed is extremely significant. As such, we can reject the hypothesis that discussion levels were similar in the two elections.

5.2 Discussion

This difference in the distributions as well as large difference in the means of discussion levels between the 2016 and 2020 elections lead us to believe the work of the Fujiwara et. al paper [3]. Their research showed that exposure to Twitter lowered the Republican vote share in the 2016 election, despite winning, which they claim is driven by moderate and and independent voters. Given our findings, we believe this to be true, as a drastic increase in discussion levels on Twitter as our investigation showed was present in the 2020 presidential election which resulted in a win for the Democratic candidate Joe Biden.

6 Conclusion

Due to the increase in online communication and the expansion of social media platforms, connecting with people and discussing about topical issues has been prevalent in the United States. Even the political candidates of each party use social media as a form of connection to the public and thus leading to online discussions about such national events.

Given our analysis results, it is clear that sentiment has been a factor that describes the level of discussion on twitter. Moreover, we formulated a level of discussion metric that summarizes the magnitude of discussion and compared such results between the election years 2016 and 2020. Our findings suggest that the online discussion has been more extensive in 2020 and the positive sentiment as we approached the election day also support the fact that online discussion has been supportive of the Democratic win in this election.

7 Appendix

Project Proposal: `https://docs.google.com/document/d/1Ch5_wLtWYhf2dhur8s25lKATeJYD4x3ERgV0u-OQ`
`edit?usp=sharing`

References

[1] Emily Chen, Ashok Deb, and Emilio Ferrara. election2020: The first public twitter dataset on the 2020 us presidential election, 2020.

[2] Daniel Kerchner Justin Littman, Laura Wrubel. 2016 united states presidential election tweet ids, 2016.

[3] Carlo Schwarz Thomas Fujiwara, Karsten Müller. How twitter affected the 2016 presidential election, 2020.

[4] Ankur Agrawal Tim Hamling. Sentiment analysis of tweets to gain insights into the 2016 us election, 2017.