Spring 2021

Understanding the Factors that Influence Tweet Popularity

Kyle A. Jalbert
University of New Hampshire, Durham

Understanding the Factors that Influence Tweet Popularity

Kyle Jalbert

Thesis Advisor: Kholekile Gwebu

University of New Hampshire – Peter T. Paul College of Business and Economics

Table of Contents

Introduction

In today's world, social media has become a powerful source of influence and in many cases, the main news source for the public. During a presidential election, candidates are fighting for votes and it is clear that social media can serve as a powerful means to inform and sway voters. Postings that become popular on social media have the potential to influence large audiences, but how and why do some postings get greater reactions than others? This is an important question whose answer(s) can have implications on how future candidates, or influencers, can leverage social media to gain wider support.

This study investigates the impact that social media, specifically Twitter, can have on the public around the time of the 2020 United States Presidential Election. This study examines which variables make tweets more impactful than others, the sentiment these tweets reflect, and how users interact with each type of tweet. After collecting and cleaning a sample dataset of 570,713 tweets, 1,022 unique tweets were classified as the most impactful based upon the number of individuals who liked the tweets (aka "number of favorites"). A total of 91 variables describing each tweet were analyzed using R, Tableau, Excel and SPSS.

A Linear Regression model is used to examine the relationships between the independent variables describing each tweet and the dependent variable, the number of favorites that a tweet receives. Combining the findings from the regression model with findings from the sentiment analysis on the 1022 most impactful tweets provide many insights into which tweets have an impact, who reacts to these tweets and how they react.

Literature Review

The literature on misinformation is sparse, however, it is rapidly gaining interest given the potential negative ramifications associated with the spread of misinformation. With regards to the spread of misinformation in presidential elections, little research exists but there is a large amount of news being published, just not research on the effects of this news. There has previously been research on similar topics in elections, but social media is forever changing and will always produce different results which makes this research important and can also open insights on how consumers viewpoints have changed over time. Most of the similar research was based on the 2016 US Presidential Election fake news, which had different candidates and a different situation overall, especially regarding the fact that the world is facing a global pandemic during this 2020 election. There are also many studies regarding psychological effects from social media and how social media can influence the public based on psychology and social theories, which relates to the background, or subconscious, side of this research.

For example, Allcott and Gentzkow (2017) address the concern about the effects of false stories ("fake news"), circulated largely through social media around the 2016 US presidential election. They discuss the economics of fake news and present new data on its consumption before the election. Drawing on web browsing data, archives of fact-checking websites, and results from a new online survey, they found that:

I. “social media was an important but not dominant source of election news, with 14 percent of Americans calling social media their "most important" source”
II. “of the known false news stories that appeared in the three months before the election, those favoring Trump were shared a total of 30 million times on Facebook, while those favoring Clinton were shared 8 million times”
III. “the average American adult saw on the order of one or perhaps several fake news stories in the months around the election, with just over half of those who recalled seeing them believing them”
IV. “people are much more likely to believe stories that favor their preferred candidate, especially if they have ideologically segregated social media networks.”

Bovet and Makse (2019) used a dataset of 171 million tweets in the five months preceding the 2016 US presidential election day to identify 30 million tweets, from 2.2 million users containing a link to news outlets. Based on a classification of news outlets curated by www.opensources.co, this study found that 25% of these tweets spread either fake or extremely biased news. They were able to characterize the network of information flow to find the most influential spreaders of fake and traditional news and uncover how fake news influenced the presidential election. It was found that while top influencers spreading traditional center and left-leaning news largely influenced the activity of Clinton supporters, this is reversed for the fake news: the activity of Trump supporters influenced the dynamics of the top fake news spreaders.

Shu, et al. (2017) find that social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news from social media. On the other hand, it enables the widespread of fake news, i.e., low-quality news with intentionally false information. It is found that the extensive spread of this fake news has the potential for extremely negative impacts on individuals and society. Fake news detection on social media presents unique characteristics and challenges that make existing detection algorithms from traditional news media ineffective or not applicable:

I. Fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content. Because of this challenge, they included auxiliary information, such as user social engagements on social media, to help decide.
II. Exploiting this auxiliary information is challenging in and of itself as users' social engagements with fake news produce a large amount of data. Because the issue of fake news detection on social media is both challenging and relevant, they conducted this survey to further facilitate research on the problem.

In this survey, they presented a comprehensive review of detecting fake news on social media, including fake news characterizations on psychology and social theories, existing algorithms from a data mining perspective, evaluation metrics, and representative datasets. They also discuss related research areas, open problems, and future research directions for fake news detection on social media.

Huang, Jianyi, et al. look into the active period of popularity evolution indicates how long online content receives continuous attention from people. Although predicting popularity evolution has largely been explored, researches on predicting active period still remain open. If the duration of active period ahead of time, caching systems, online advertising, etc. are known, then they can run more effectively. Therefore, predicting active period is of great importance, but it is a non-trivial task because of the two major challenges:

I. Numerous factors can influence the duration of active period. To predict active period accurately, it's difficult to consider what factors and how to embed them in DNN model.

II. The triggering time to predict different active periods must be decided carefully, because the durations of active periods differed from one another.

This paper addresses these two challenges, focusing on Twitter hashtags as a case study. To deal with the first challenge, a DNN-based prediction framework is proposed, embedding dynamic and static factors by using LSTM and CNN respectively. To deal with the second challenge, an appropriate value of cumulative popularity is set to trigger predicting active period. Experimental and comparative results show the superiority of our prediction solution, comparing with spikeM and SVR.

There seems to be gaps in the literature regarding how users can maximize the impact of a tweet through what variables a user can incorporate, improve and what sentiment that a tweet should express. This study attempts to answer some of these reasons that some tweets are more impactful than others by constructing a regression model predicting the number of favorites a tweet should get and also by studying the sentiment expressed in the top tweets.

Hypothesis Development

There are many factors related to tweets and Twitter users that allows them to be impactful on Twitter. Out of the 91 variables describing each tweet, there were 7 that appeared likely to be important factors for making a tweet go viral and consequently have a large influence/impact. These variables include Retweet, Quote, Hash Tag, Followers, Verified, Media, Mentions and Creation. Table 1 provides a definition of each of the variables.

Table 1: Variable Definitions

Variable	Definitions
Retweet	# of retweets a tweet has received
Quote	Is the tweet a quote of another tweet?
Hash Tag	# of characters a tweet has in hashtags
Followers	# of followers the user that tweeted the tweet has
Verified	Is the user that tweeted this tweet verified?
Media	Does the tweet contain media (picture, video)?
Mentions	Does the tweet mention (@) another twitter user?
Creation	What time and date was this tweet created at

For the first variable, Retweet, it appears likely that if a tweet has more retweets, then it will also have more favorites. This is due to it having the potential to reach a wider audience more quickly. When another user retweets someone's tweet, then it directly places the tweet into all of their followers feed and this tends to become a chain effect which is why I expect that it will be a crucial factor in tweets becoming viral and more impactful. Thus, the first hypothesis is that

H₁ – *The number of retweets will have a positive impact on the number of favorites*

Quoted tweets are likely to play a role in the number of favorites that a tweet receives because quoting another tweet involves more users and adopts the attributes from that quoted tweet. Thus, the second hypothesis is that

H₂ – *Quoted tweets have a positive impact on the number of favorites*

Hashtags are also likely to play a role in the number of favorites. Longer hashtags are harder to read/decipher since there are no spaces in between words, which makes them difficult for users to understand. Longer hashtags are also not as commonly used which will reduce the visibility compared to using a hashtag that is more commonly used. Thus, it is expected that lengthy hashtags will not become a viral, widely used or searched and consequently they will not make a tweet as impactful. In other words, if a tweet has a very long hashtag, then it will have less of an impact compared to tweets with shorter hashtags. Accordingly, it is hypothesized that

H₃ – *The length of a hashtag has a negative impact on the number of favorites*

The fourth hypothesis considers the number of followers of the person posting the tweet. A tweet published a user who has more followers is likely to have a larger impact than a tweet posted by a user with fewer follows. This is because tweets are automatically displayed in a user's feed thus there are more eyeballs looking at the tweets of a user with more followers than a user with fewer followers. Consequently, if a user that published a tweet has more followers, then the tweet will get more favorites and be more impactful. It is therefore hypothesized that

H₄ – *There will be a significant positive relationship between the number of followers of a poster and the number of favorites*

As a verified Twitter account resembles an authentic, active user of public interest, users that are verified are likely to enjoy more attention. These verified accounts tend to be influential people or organizations that have a following outside of Twitter and already have a large fan base. Thus, if the user that published a tweet is verified, then the tweet is likely get more favorites and have a larger impact. It is thus hypothesized that

H_5 *– There is a significant positive relationship between verified accounts and the number of favorites.*

Tweets containing media, such as photos or videos, are likely to attract more attention than tweets only containing text because they are more interactive and attractive when scrolling through a Twitter feed. The more attention a tweet attracts the more favorites it is likely to attract. Thus, it is hypothesized that

H_6 *– There is a significant positive relationship between the existence of media on a tween and the number of favorites*

A tweet can involve/engage multiple users through the @ symbol, which can have the effect of increasing interactions. Consequently, the Tweet has the potential to reach a wider audience and in return, increase the number of favorites that a tweet will receive. If a tweet is a quote of another person's tweet, then it will have a larger impact. Thus, the final hypothesis is that

H_7 *– There is a positive impact between mentioning another user and the number of favorites.*

Methodology

Data and Cleaning

The data that has been used in this research is primary data directly drawn from Twitter. The data was collected between October 30th 2020 to November 11th 2020. This time period is particularly interesting because during the time directly before and following the election there was a lot of "chatter" on Twitter surrounding the Presidential Election. Some of the tweets posted during this time period became very popular (went viral) and consequently had a large impact on discussions surrounding the election in the popular press as well as across multiple social media websites.

The criteria for tweets being pulled were set by brainstorming a list of words that would be relevant to the US Presidential Election, as follows;

‘Trump’, ‘Biden’, ‘Republicans’, ‘Democrats’, ‘President’, ‘US Election’, ‘Fake news’, ‘vote’, ‘votes’, ‘ballots’, ‘White House’, ‘Election’, ‘Election Rumors’, ‘Electoral College’, ‘supreme court’, ‘Election Results’, ‘media’, ‘steal election’

These keywords are the criteria the R script used to pull all tweets containing any of these words over the last week from the time that the script was ran (see Appendix A for code). This script utilized the R package “Rtweet” and connected to the Twitter REST API. The script was ran daily and collected 91 variables (see Appendix A) for all tweets from October 30th to November 11th.

After collecting all of the “primary data” in a raw format, it was necessary to format and clean it before analysis. After formatting the data into a data frame, the approach to clean this data consisted of ensuring there was no duplicates and removing any duplicates by using the R function “duplicated()” in the tidyverse package. Before analyzing this data, it was important to ensure that the variables being used in the regression model were filled in for most of the records. Due to privacy settings on twitter some variables, such as location, are not available for every record in the data set. Because of this, only variables that had values for 80% of all records or more were considered in the analysis of this data set. This resulted in a final dataset consisting of 573,681 tweets with 58 variables for each tweet.

Descriptive Statistics

Below are the descriptive statistics for the data set used in the analytics that follows.

Table 3: Descriptive Statistics

Descriptive Statistics of Variables Used in Regression Model			
	Mean	**Median**	**Standard Deviation**
Retweet	**1.02**	**0**	**22.09**
Quote (0 = FALSE, 1 = TRUE)	**0.09**	**0 (FALSE)**	**0.29**
Hashtag	**8.85**	**0**	**24.08**
Followers	**22972.65**	**253**	**512809.10**
Verified (0 = FALSE, 1 = TRUE)	**0.04**	**0 (FALSE)**	**0.19**
Media (0 = FALSE, 1 = TRUE)	**0.10**	**0 (FALSE)**	**0.42**

Mentions	**0.62**	**1 (TRUE)**	**0.49**
Creation (Control Variable)	**N/A**	**N/A**	**N/A**

This data set seems to have a low number of retweets which may be a sign that there are a lot of low-quality tweets in the data and only so few end up going viral. This low median of retweets could also be a result of the sentiment being mostly negative, politically oriented tweets so users may not want to upset their followers or get misjudged by publishing them to their public profile. Even with this mean being so low, it is a very significant variable in being a reason that tweets are more impactful.

Looking at the mean (~9) and median (0) of the length of Hashtags represents that at least half of the tweets did not use hashtags but the ones that did must have used a lot of hashtags/long hashtags. Using long hashtags had a negative impact on a tweet going viral so it would be interesting to look father into this analysis and see if it was the low impact tweets using hashtags or if it was a mix.

Below are the number of tweets that were pulled on each date during data collection and were then analyzed.

Table 4: Sample Size

Number of tweets collected each day	
11/6/2020	**95613**
11/7/2020	**98423**
11/8/2020	**96827**
11/9/2020	**92519**
11/10/2020	**94672**
11/11/2020	**95627**

Analysis

A regression model was constructed to test the hypothesized relationships between the variables of interest and the dependent variable (number of favorites). Since the data being analyzed was collected over 13 days, the newer tweets were potentially captured before they had a chance to go viral or reach their maximum influence potential. Because of this, the regression model uses the "created_at" (date and time), relabeled "Creation", variable as a control variable to account for tweets that have had a longer time to be noticed than others.

Based on the above-mentioned hypothesis, the regression model was specified as follows:

$$Favorite = \alpha + \beta_1 Retweet + \beta_2 Quote + \beta_3 Hash\,Tag + \beta_4 Followers + \beta_5 Verified + \beta_6 Media + \beta_7 Mentions + \beta_8 Creation + \varepsilon$$

Findings

Table 3 summarizes the findings of the regression analysis. The model accounts for 93.5% of the variance.

Table 3: Regression Results

Variable	Estimate
Retweet	0.962**
Quote	0.002*
Hash Tag	-0.004**
Followers	0.033**
Verified	0.013**
Media	-0.008**
Mentions	0.001
Creation	-0.001
R^2	0.935
N	113254

a. Dependent Variable: Like

* and ** indicate significance at the 5% and 1% level, respectively

The first hypothesis suggested that the higher the number of retweets the higher the number of favorites. In other words, there should be a positive relationship between retweets and the number of favorites. The results from the regression model lend strong support to the hypothesis (β=0.962, p<0.05).

The second hypothesis argues that if a tweet is a quote of another tweet, then it will be more impactful. Although support was found for the hypothesis ($\beta=0.002$, $p<0.1$) the impact appears to be rather marginal and smaller than anticipated. The expectation for a larger impact was based on the notion that quoted tweets trigger a reply feed and are therefore more likely to be relevant to more users given that more users involved in the tweet.

The third hypothesis expected very long hashtags to not be as impactful as concise and shorter hashtags, which is accepted from the regression model. A significant negative relationship $\beta=-0.004$, $p<0.05$) is found between the length of the hashtag and the dependent variable. This implies support for H_3.

The fourth hypothesis suggests that if the user who published a tweet has more followers then the tweet will be more impactful. The results from the regression analysis reveal that the number of followers have a positive and significant ($\beta=0.013$, $p<0.05$) impact on the number of favorites thus lending support to H_4.

The fifth hypothesis deals with the relationship between verified users and the number of favorites. It implies that if the user who published a tweet is a verified user, then the tweet that they post will receive more favorites. The results from the regression model lend support for H_5 with the variable verified having a positive and significant impact ($\beta=0.033$, $p<0.05$) on the number of favorites.

The sixth hypothesis implies that if a tweet contains media then it will be more impactful. Interestingly, a negative and significant relationship $\beta=-0.008$, $p<0.05$) is found between "media" and number of favorites. This suggests that tweets with media receive fewer favorites than tweets without media, thus H_6 is not supported.

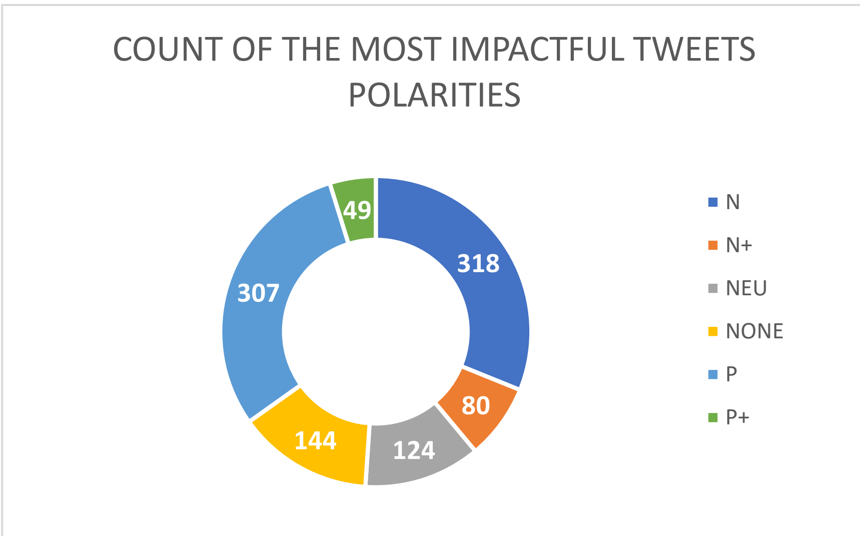
The final hypothesis indicates that tweets that mention other user will have a larger impact on the number of favorites. No significant relationship is found thus H_7 is not supported.

Additional Analyses

A sentiment analysis was conducted to gain additional insights into the nature of the most impactful tweets. The MeaningCloud API was used for sentiment analysis to assess whether positively oriented or negatively oriented tweets would get more reactions or produce a larger impact on Twitter. Looking at the 1,022 most impactful tweets, by number of favorites, the MeaningCloud API classified each tweet on a scale of Very Negative, Negative, Neutral, Positive, Very Positive, or None. This sentiment analysis helps provide answers for what type of information Twitter users are reacting to and how they react to it.

There are a similar number of total positive (356, 34.8%) and negative (398, 38.9%) tweets (See Figure 1) that are in the 1,022 most impactful tweets, but there is a larger amount of very negative (80) tweets compared to very positive (49) tweets.

Figure 1



Looking deeper into how users were reacting to the tweets based on sentiment classification, it was clear that very negative tweets had the most impact with an average of 70,245 favorites and 149 retweets per tweet (See Figure 2 & 3). The number of retweets does not vary much as tweets transition from being very negative towards being very positive, but there is a clear correlation of tweets becoming less impactful with the more positive sentiment they express (See Figure 2).

Figure 2

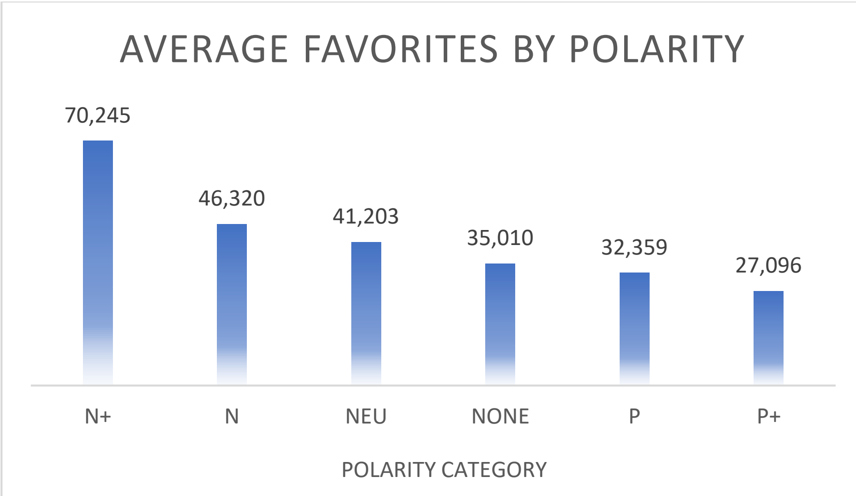
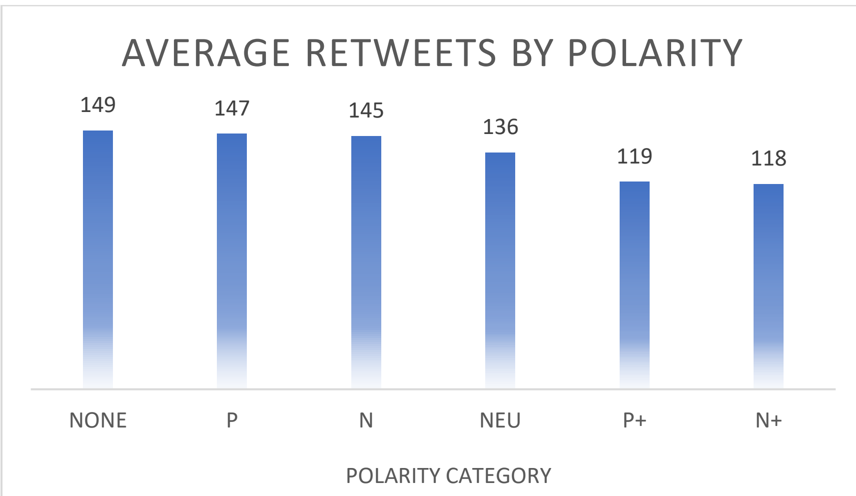
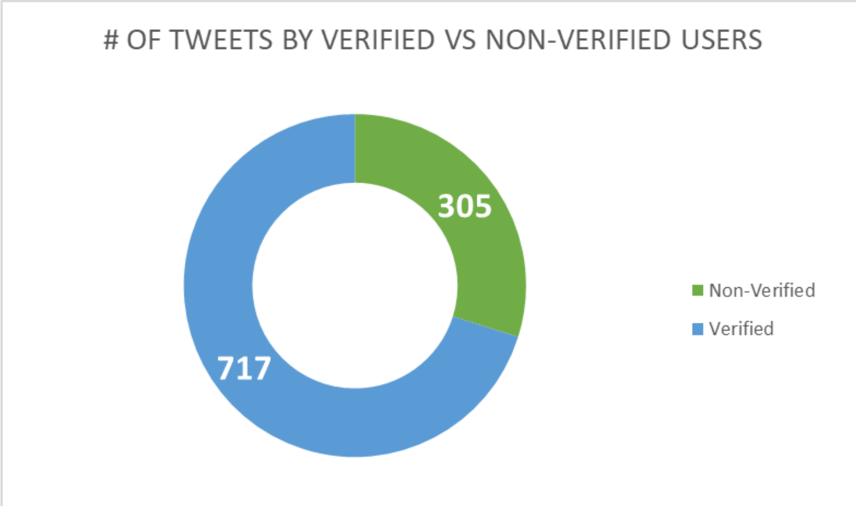


Figure 3



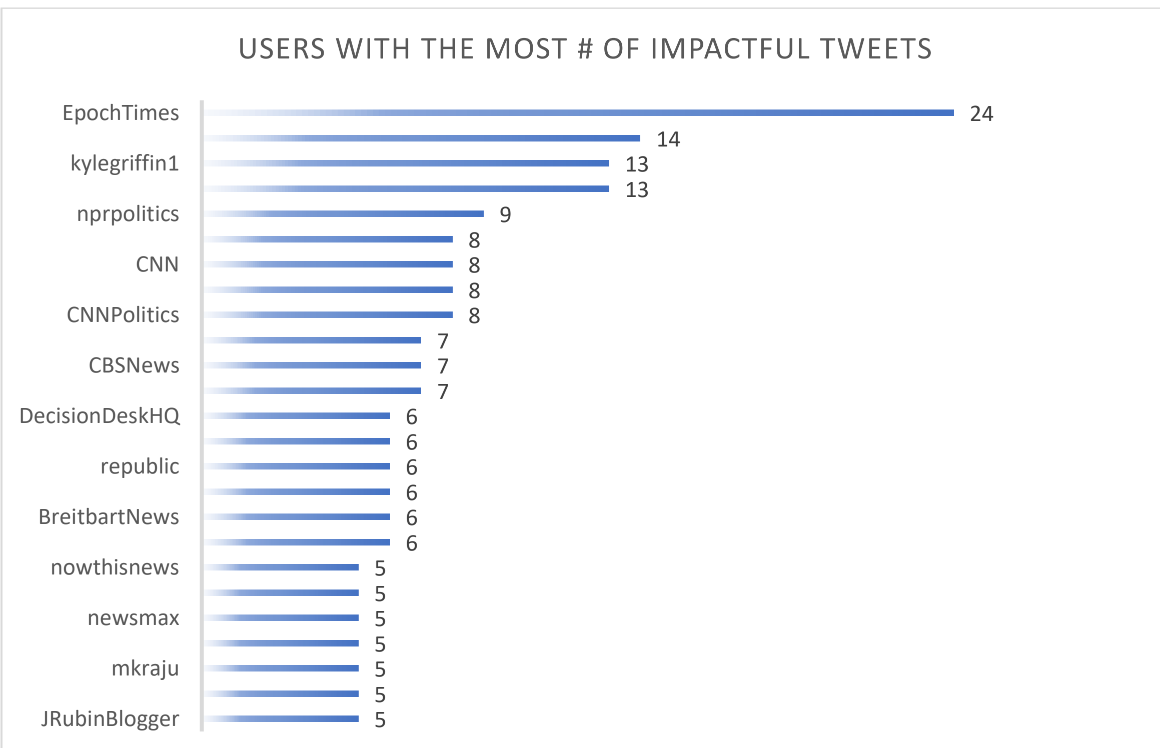
When analyzing the top tweets, there are a number of users that seemed to have a much more consistent impact on Twitter, by having numerous tweets in this top portion of the dataset. As verified users tend to have more followers on Twitter, their tweets are shown to also have a larger impact as 717 (70.16%) of the 1,022 most impactful tweets are published by verified users (See Figure 4).

Figure 4



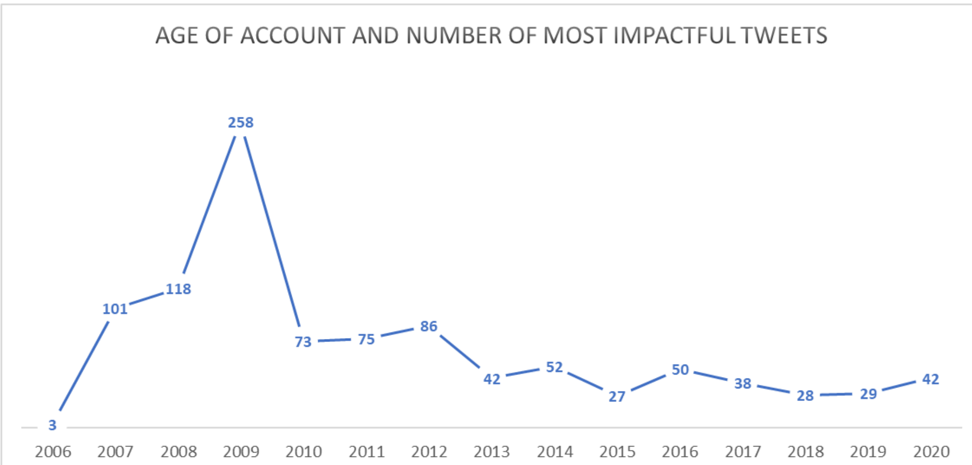
Out of the top 25 users that were able to get multiple tweets into the top 1022 tweets (See Figure 5), all of these users are verified users except for one, HowleyReporter, who managed to get 5 tweets into the top 1,022 most impactful tweets. This evidence suggests that verified users are able to publish more impactful tweets more often.

Figure 5



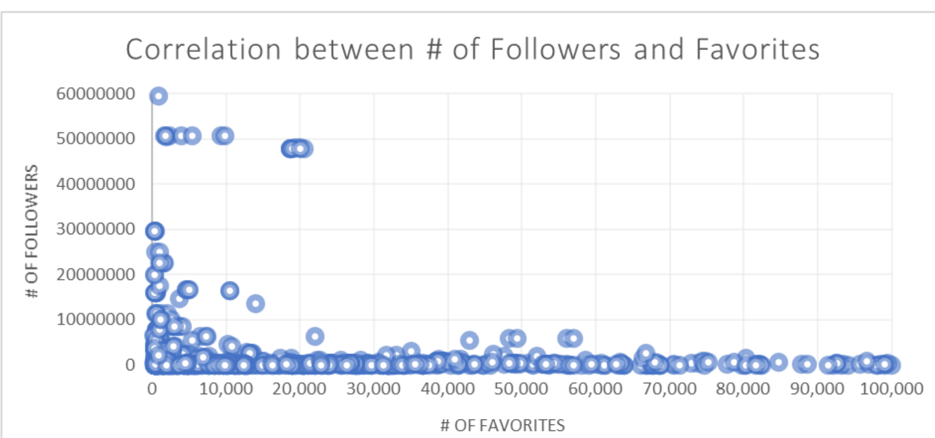
As we see that verified users more consistently have impactful tweets, it is also a potential because these accounts have been around for longer. As Twitter was launched and the first tweet ever was made March 21st 2006, there is a large increase, and peak, in the number of tweets of the top 1022 that were published by users who created their account in 2009 (See Figure 6).

Figure 6



When looking into more factors of these tweets becoming the most impactful, there seems to be no correlation of the # of followers of each user that published a tweet and the # of favorites that the tweet received for each tweet in the top 1022 (See Figure 7). Even though having a larger follower count does not mean that a user's Tweets will have a large impact, it does have the potential to do so since the tweet is automatically published to their followers feed.

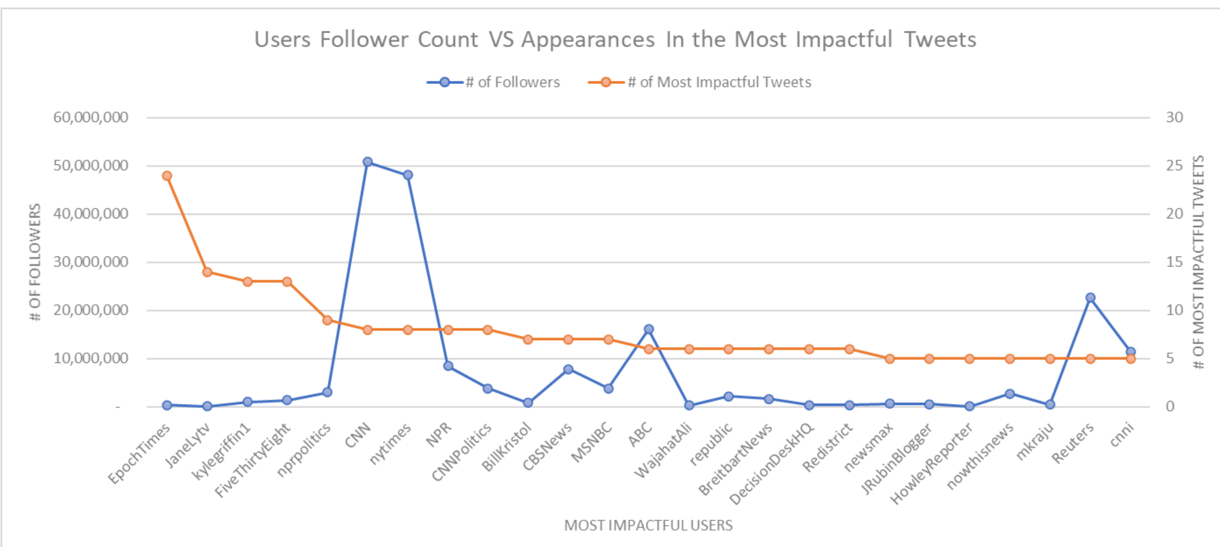
Figure 7



Even with this potential, it is clear that there is not a strong correlation, if one at all, between a user's follower count and their favorite count on their tweets. When analyzing these two variables for the top 1022 tweets, the correlation coefficient is -0.087 which shows the slightest negative correlation but is very close to zero, or no correlation. This correlation does not accurately describe the relationship between the number of followers an account has and the amount of likes that their tweets receive.

As users who have more followers have the potential to have more impactful tweets, this was not always the case when looking at the top 1022 tweets (See figure 8). The users that had the greatest number of tweets in the top tweets were not the ones with the most followers. As seen in Figure 8, CNN and NewYorkTimes has significantly more followers than the other users in this category but had the sixth and seventh the greatest number of Tweets in the top 1022 which suggests that there are other important factors in producing impactful tweets.

Figure 8



Conclusion and Implications

The results indicate that there is a clear difference between tweets and their level of impact on Twitter when regarding the US Presidential Election. There are a few factors that have great significance when influencing what tweets will become viral. Some of the most deciding factors that came from the regression model are the greater the number of retweets that a tweet receives, using shorter hashtags, the user who published the tweet having more followers, the user being a verified user, and not including media in the tweet. From a sentiment stand point, it appears that tweets reflecting a negative sentiment tend to become more impactful than any other type of sentiment. It is assumed that by combining all of these factors influencers and politicians can leverage their tweets to make a larger impact on their Twitter audience.

References

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*(1), 7. https://doi.org/10.1038/s41467-018-07761-2

Huang, Jianyi, et al. "Predicting the active period of popularity evolution: A case study on Twitter hashtags." Information Sciences 512 (2020): 315-326.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Appendix A

Sample of code for pulling data from R:

```
###Install and load required packages

## Install and load multiple packages at once


install_load <- function(pack){

  ## Statement to check if the package has been previously installed

  new_pack_load <- pack[!(pack %in% installed.packages()[,"Package"])]

  if (length(new_pack_load))
```

```
  install.packages(new_pack_load, dependencies = TRUE)
  sapply(pack, require, character.only = TRUE)}


package_load <- c("rtweet", "dplyr", "reactable", "glue", "httpuv",
                  "stringr", "purrr", "DT", "lubridate", "tidytext",
                  "wordcloud", "igraph", "ggraph","widyr", "ggmap",
                  "leaflet", "scales", "SchedulerR")
install_load(package_load)
## install remotes package if it's not already: https://github.com/ropensci/rtweet
if (!requireNamespace("remotes", quietly = TRUE)) {
install.packages("remotes")}
## install dev version of rtweet from github
remotes::install_github("ropensci/rtweet")
## load rtweet package
library(rtweet)


##Connect Via Twitter API
consumer_key <- 'xxxxxxxxxxxxxxxxx'
consumer_secret <- 'xxxxxxxxxxxxxxxxx'
access_token <- 'xxxxxxxxxxxxxxxxx'
access_secret <- 'xxxxxxxxxxxxxxxxx'


token <- create_token(
```

```
app = "rtweet_tokens19",

consumer_key = consumer_key,

consumer_secret = consumer_secret,

access_token = access_token,

access_secret = access_secret)

numberOfTweets <- 1000


US_ELECTION_Nov9 <- search_tweets('#US Election', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Fake_News_Nov9 <- search_tweets('#Fake news', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Election_Nov9 <- search_tweets('#Election', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Election_Rumors_Nov9 <- search_tweets('Election Rumors', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Trump_Nov9 <- search_tweets('Trump', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Biden_Nov9 <- search_tweets('Biden', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

President_Nov9 <- search_tweets('President', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Electoral_College_Nov9 <- search_tweets('Electoral College', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Election_Results_Nov9 <- search_tweets('Election results', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')
```

```
Democrats_Nov9 <- search_tweets('Democrats', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Republicans_Nov9 <- search_tweets('Republicans', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

White_House_Nov9 <- search_tweets('White House', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Media_Nov9 <- search_tweets('media', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Steal_Election_Nov9 <- search_tweets('steal election', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Vote_Nov9 <- search_tweets('vote', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

ballots_Nov9 <- search_tweets('ballots', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

Votes_Nov9 <- search_tweets('votes', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')

supreme_court_Nov9 <- search_tweets('supreme court', n = numberOfTweets,  include_rts = FALSE, `-filter` = "replies", lang='en')
```

Variables describing each tweet:

```
user_id = col_character(),

status_id = col_character(),

created_at = col_datetime(format = ""),

screen_name = col_character(),

text = col_character(),
```

source = col_character(),

display_text_width = col_double(),

reply_to_status_id = col_character(),

reply_to_user_id = col_character(),

reply_to_screen_name = col_character(),

is_quote = col_logical(),

is_retweet = col_logical(),

favorite_count = col_double(),

retweet_count = col_double(),

quote_count = col_logical(),

reply_count = col_logical(),

hashtags = col_character(),

symbols = col_character(),

urls_url = col_character(),

urls_t.co = col_character(),

urls_expanded_url = col_character(),

media_url = col_character(),

media_t.co = col_character(),

media_expanded_url = col_character(),

media_type = col_character(),

ext_media_url = col_character(),

ext_media_t.co = col_character(),

ext_media_expanded_url = col_character(),

```
ext_media_type = col_logical(),

mentions_user_id = col_character(),

mentions_screen_name = col_character(),

lang = col_character(),

quoted_status_id = col_character(),

quoted_text = col_character(),

quoted_created_at = col_datetime(format = ""),

quoted_source = col_character(),

quoted_favorite_count = col_double(),

quoted_retweet_count = col_double(),

quoted_user_id = col_character(),

quoted_screen_name = col_character(),

quoted_name = col_character(),

quoted_followers_count = col_double(),

quoted_friends_count = col_double(),

quoted_statuses_count = col_double(),

quoted_location = col_character(),

quoted_description = col_character(),

quoted_verified = col_logical(),

retweet_status_id = col_logical(),

retweet_text = col_logical(),

retweet_created_at = col_logical(),

retweet_source = col_logical(),
```

```
retweet_favorite_count = col_logical(),

retweet_retweet_count = col_logical(),

retweet_user_id = col_logical(),

retweet_screen_name = col_logical(),

retweet_name = col_logical(),

retweet_followers_count = col_logical(),

retweet_friends_count = col_logical(),

retweet_statuses_count = col_logical(),

retweet_location = col_logical(),

retweet_description = col_logical(),

retweet_verified = col_logical(),

place_url = col_character(),

place_name = col_character(),

place_full_name = col_character(),

place_type = col_character(),

country = col_character(),

country_code = col_character(),

geo_coords = col_character(),

coords_coords = col_character(),

bbox_coords = col_character(),

status_url = col_character(),

name = col_character(),

location = col_character(),
```

```
description = col_character(),

url = col_character(),

protected = col_logical(),

followers_count = col_double(),

friends_count = col_double(),

listed_count = col_double(),

statuses_count = col_double(),

favourites_count = col_double(),

account_created_at = col_datetime(format = ""),

verified = col_logical(),

profile_url = col_character(),

profile_expanded_url = col_character(),

account_lang = col_logical(),

profile_banner_url = col_character(),

profile_background_url = col_character(),

profile_image_url = col_character(),

dataset = col_character()
```