Political Popularity of Misinformation

Catherine Tao, Aaron Chan, Matthew Sao

University of California, San Diego

March 7, 2021

1 Abstract

For our research on Political Popularity of Misinformation, we want to research the influence politicians have on Twitter, a well known social media platform for users to voice their opinions to a wider audience. The information shared on Twitter that we are interested in will be grouped into scientific information or conspiracy information. Politicians can easily sway public opinion with a simple tweet, therefore we wanted to analyze how much they can influence other Twitter users with their tweets.

We gather ten politicians who we consider to spread scientific information on Twitter and ten politicians who we consider to spread misinformation on Twitter. These two groups will be analyzed to show how controversial a tweet appears. We analyze this by looking at ratios for tweet engagement as well as a rolling cumulative maximum metric to see growth over time.

For this overall project, we want to see the amount of engagement per tweet for a given politician. We calculate this with a mathematical ratio equation to calculate each tweet's engagement. The ratio equation incorporates the number of likes, retweets, and comments a given tweet holds. We conclude that a high ratio means more controversy surrounding the tweet and a low ratio meaning less controversy.

Another metric we use is the rolling and cumulative maximum metric. This metric is used as a measure to analyze a politician's growth/ popularity over time. We gather the tweets of a particular politician and we then find the cumulative maximum number of likes and retweets for the tweet. We also look at the rolling max over a span of 4 months of a politician's tweets.

The results of our investigation showed that politicians who spread misinformation have a higher ratio value on average and have less overall likes over their tweets. Our permutation test shows that our scientific group has been consistently growing and increasing in growth over time. In contrast, our misinformation group has been growing significantly, but only in the more recent years. The growth for both scientific and misinformation groups are similar only in 2017. Overall, our results show that a politician can experience the most growth through spreading non-controversial, scientific information.

2 Introduction

The rise of the internet and easily accessible and instantaneous information in the recent century has caused a significant change in the way that the public ingests their news. In a large part, this shift to instantaneous public information has allowed this generation to be the most informed that it has ever been, but also the most opinionated and misconstrued.

Social media has become the main source of easily accessible and digestible information for field experts and organizations to publicly spread news, but at the same time it has become a place where individuals can spread their beliefs as fact and influence others' opinions on subjects readers have yet to be informed about. As the internet is a place open for anyone to share information, the validity of information presented is not always guaranteed to be accurate or benevolent.

For our research on Political Popularity of Misinformation, we want to research the influence politicians have on Twitter, a well known social media platform for users to voice their opinions to a wider audience. The information shared on Twitter that we are interested in will be grouped into scientific information or conspiracy information. We have chosen ten politicians to represent our scientific group and another ten politicians to represent our misinformation group. Politicians can easily sway public opinion with a simple tweet, therefore we wanted to analyze how much they can influence other Twitter users with their tweets. This specific analysis is interesting because we are able to determine whether a politician's tweet has influence on the public based on tweet engagement and a politician's growth on Twitter.

For each politician, we gather our Twitter data by using Tweepy, a python library from Twitter's API. This is to collect Twitter accounts and individual tweets. Each tweet has a unique ID which we are able to rehydrate to gather the info from the specified tweet. Some of the relevant information as a result from rehydrating include the tweet text, date, number of likes, number of comments, and much more.

Throughout our investigation, we used mathematical methods in order to analyze engagement of the tweets and to compare our two groups. The ratio method is used to analyze engagement of a politician's tweets. This method takes in account the retweets, likes, and comments of a specific tweet. The cumulative maximum and rolling maximum methods are used to measure a politician's growth over time. Finally, we used a permutation test in order to compare our two sample groups to draw conclusions.

Data visualizations are shown to illustrate the technical findings into a visual representation where we can view trends and patterns. The graphs shown are a way to compare different groups of politicians.

[1] Many of the politician's tweet IDs were gathered from a third party source which stores all individuals holding office from the Senate and Congress. The starting tweets for each individual varies depending how long they have been actively tweeting on their specified Twitter account.

3 Data Collection

Our data consists of a collection of tweets for each individual politician, also known as their timeline. We obtain the tweet IDs that compose our politicians' timeline from George Washington University's TweetSets database. The TweetSets database has datasets consisting of tweets for research and archival purposes, covering a wide range of topics such as climate change, the 2018 Winter Olympics, the two most recent presidential elections as well as tweets made by politicians of the 115th and 116th Congress.

For our analysis, we chose to focus on politicians who served in the 116th United States Congress, which corresponds to two datasets, Congress: Representatives of the 116th Congress and Congress: Senators of the 116th Congress. We specifically chose the 116th Congress as it is the most recently concluded session at the time of writing. The two datasets combined contain 2,756,042 tweet IDs and were collected between January 27, 2019 and May 7, 2020 from Twitter's API using Social Feed Manager. [2] The earliest tweet in this dataset occurred on January 26, 2008 while the last tweet was made on May 5, 2020. It is worth noting that not all of the politicians have tweets spanning multiple years. This is a result of some politicians having just been recently elected to Congress, such as Alexandria Ocasio-Cortez whose first term was the 116th Congress.

To start our data collection process, we first identified twenty politicians, ten of which were known to spread misinformation during their time in office and ten which were known to spread scientific information. In order to classify a politician as someone who spreads misinformation we researched notable politicians and justified their classification through reports and news articles detailing their statements on topics ranging from the coronavirus to the most recent election. For example, Senator Joni Ernst, who falsely claimed that healthcare providers are inflating the number of coronavirus cases, or Representative Matt Gaetz, who falsely claimed that Antifa members were part of the riots on Capitol Hill. [3] [4]

After identifying our politicians, we gathered the user IDs for their Twitter accounts using Tweepy, which are then used to query the two Congress datasets. We use a politician's user ID as opposed to their username because a politician's username may change over time while their user

ID remains constant. The datasets also contain a file of the House and Senate members along with their user IDs which is an alternative way to obtain these IDs. To query the datasets, for each politician, we selected either the Representative or Senator dataset depending on their position and inputted their user ID in the "Contains any user id" box under the "Posted by" section. This process gives us a txt file of tweet IDs for each politician which we then rehydrate using Twarc. The output is a json file for each politician that contains tweet objects returned by Twitter's API. The average number of tweets for our scientific politicians is 4,563 while the average number of tweets for our misinformation politician is 5,446.4.

In order for us to calculate a tweet's ratio, we need to have information about the number of times a tweet has been replied to. Unfortunately, we are not able to access the reply_count attribute on a Tweet object without the Premium or Enterprise tier of Twitter's API. As an alternative, we make calls to the Twitter API's metrics field, which allows us to access engagement metrics for Tweet objects. For each politician, we use curl to request a tweet's retweet, likes and reply counts and save the output into a csv. At the end of our entire data collection process, each politician has a txt, json and csv file.

4 Methods

For this section, we discuss the three different methods we use to analyze and draw conclusions to our results. The three methods include the ratio method, the rolling/cumulative maximum method, and the permutation test method.

4.1 Ratio Metric

We analyze the community engagement by using a ratio metric. This method incorporates the number of likes, retweets, and comments a specified tweet holds. We were able to use a mathematical equation to measure the amount of community engagement with the specified given numbers from each tweet. A high ratio will generally mean the Tweet has received a positive reaction whereas a low ratio would indicate a negative reaction. We intend to track the reaction of each tweet a politician tweets over time to see the politician's approval over time.

To analyze reception to a particular tweet, we chose to use the concept of ratios or "getting ratioed" on Twitter. This is the number of replies to the number of likes and retweets a tweet receives. Ratios allow for a quantitative way to measure how controversial a tweet is, with higher ratios signaling a more disputed tweet. [5] We formally define our measure of ratio below.

$$\frac{2 * \# \, of \; replies}{\# \, of \; likes + \# \, of \; retweets} \tag{1}$$

We weigh replies more heavily than likes and retweets due to the increased amount of effort it takes to write out a reply to a tweet as opposed to liking or retweeting that same tweet.

The ratios for tweets per politician are averaged for each politician in order to determine their average ratio. Each average ratio does not include days where a politician does not tweet because the ratio would result in a zero of infinity value since the likes, comments, and retweets would be zero. These tweets are removed from the other ratios in order to prevent skewing of a politician's average ratio result.

Horizontal bar graphs were used to best represent the findings for the averaged ratios per politicians. By having the bar graph be horizontal, it is easier to view the politicians' names on the y-axis while the x-axis holds the ratio values. Data visualizations are significant in order to see trends and patterns in our technical findings.

4.1.1 Ratio Metric: Data Visualizations

In figure 1, we graphed our ten politicians we grouped as scientific. As we can see from the graph, Lisa Murkowski and Mitt Romney have the highest ratios compared to our other eight politicians grouped under scientific politicians. It is interesting to note that these two politicians represent the Republican party, while our other eight represent the Democratic Party. This is an interesting finding because a higher ratio is defined as a negative reaction from a politician's engagement with their tweets. In contrast, a lower ratio is defined as a positive reaction to a politician's engagement with their tweets.

For our group of ten politicians grouped under misinformation, we also created a horizontal bar graph to visualize the patterns and trends. In figure 2, one interesting finding is that the only politi-

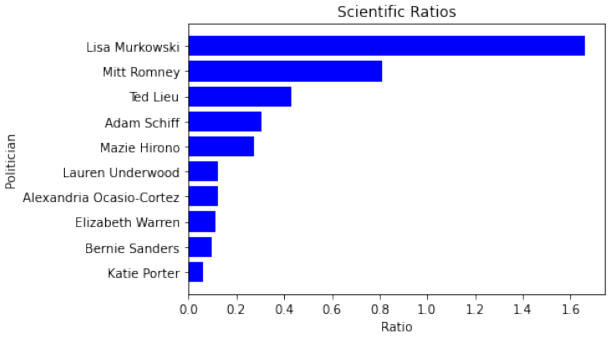

Figure 1: *Shows a horizontal bar graph for the ten politicians grouped under scientific. Represents the averaged ratios for each politician.*

cian who is democratic is Tulsi Gabbard who represents the Democratic party. His ratio is shown on the graph as the politician with the lowest ratio, meaning that his average tweet engagement is overall positive. The other nine politicians represent the Republican party. The margin of difference for each politician is not overly extensive in comparison to the Scientific Ratio graph. As seen in figure 2, Lindsey Graham, Matt Gaetz, and Joni Ernst are the three politicians with the highest ratios, meaning their tweet engagement is relatively negative.

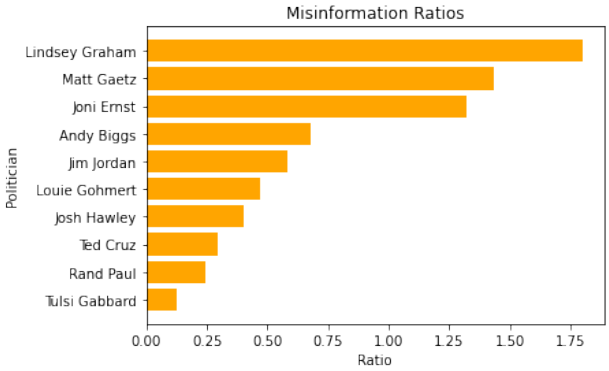

Figure 2: *Shows a horizontal bar graph for the ten politicians grouped under misinformation. Represents the averaged ratios for each politician.*

4.2 Rolling Max/ Cumulative Max Metrics

The rolling maximum and cumulative maximum methods are a way to measure a politician's growth over time. We are able to view a politician's maximum number of likes or retweets for a given day starting from the first day they begin tweeting. The start date and number of tweets for each politician will vary depending on the first day they tweeted and also the frequency in which they tweet. The rolling maximum allows us to see the popularity of a politician in a given time frame.

For each politician, we are able to analyze their likes and retweets over each month. The rolling maximum value is used to determine the number of likes or retweets for a politician over a specified window. For example, if we view Bernie Sander's rolling maximum over twelve months with a window size of 20, we are able to see his growth over the twelve months with the days grouped into sizes of 20. Each day he tweets, the maximum number of tweets for that day is extracted and we continue this process for each day. For this example, if the maximum number of likes for a tweet for the 20 day window was 1500, then the rolling maximum value for the 20 days would be 1500.

The purpose for the cumulative maximum method is to view the cumulative likes and retweets over a politician's time period on Twitter. The difference for this method is that there is no window to be specified. We take the cumulative maximum of tweets for a day and strictly use this number. This helps measure the cumulative growth over the months of activity. For example, if we view Bernie Sander's cumulative maximum for tweets, we will see the maximum number of likes or retweets per day changing each day.

4.2.1 Rolling/Cumulative Max Metric: Data Visualizations

For some graphs, the graphs analyze the tweets collected overtime while others analyze the number of tweets collected in total. The graphs showing tweets over time have a window size of 4 months for their tweets. Depending on the graph, the x-axis is tweets over time or total number of tweets while the y-axis shows the number of likes or retweets. Each graph allows us to visualize trends and patterns between each group. We are also able to make interesting findings between political parties as well as individuals from our scientific vs misinformation groups.

For the first data visualization for this metric, we show the average number of likes per month for two politicians from our scientific group. We chose the politician with the highest and lowest average ratio values in order to view any key differences from the scientific group. It is important to note that the window size for all the graphs is four months. This results in all of the graphs having the last three months as zero for the number of likes.

In figure 3, we see that Representative Katie Porter has many more average number of likes per month in comparison to Senator Lisa Murkowski. This shows a major range difference between these two politicians under our scientific group. Katie Porter's number of average likes per month exceeds Lisa Murkowski by a huge margin, meaning that Porter's tweets are less controversial.

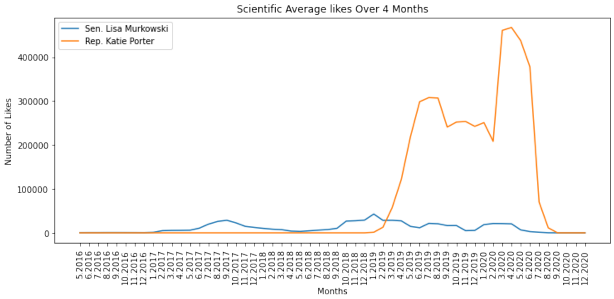

Figure 3: *This graph shows the average number of likes per month for the politicians Lisa Murkowski and Katie Porter. Both of these politicians are within our scientific group.*

In figure 4, we analyze two representatives under our misinformation group which include Lindsey Graham and Tulsi Gabbard. For this graph, it is clear that Gabbard has less average likes per month in comparison to Graham.

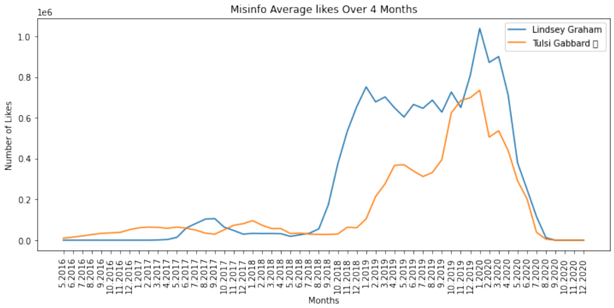

Figure 4: *This graph shows the average number of likes per month for the politicians Lindsey Graham and Tulsi Gabbard. Both of these politicians are within our misinformation group.*

For figure 5, we compare two very popular politicians from our two groups. Alexandria Ocasio-Cortez represents our scientific group and Ted Cruz represents our misinformation group. We see that Ocasio-Cortez has a much larger number of maximum likes over the four month window in comparison to Ted Cruz. We see that in this graph, Ocasio-Cortez grouped in our scientific group has more likes despite their popularity since both Ocasio-Cortez and Cruz are both equally as popular.

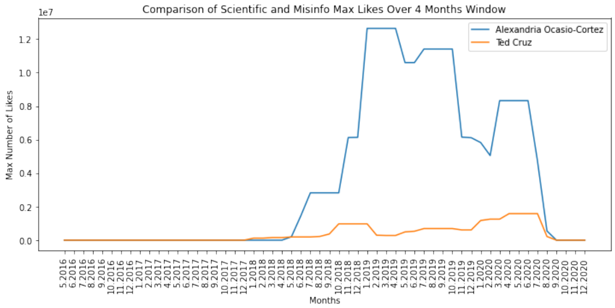

Figure 5: *This graph compares the rolling maximum number of likes for Alexandria Ocasio-Cortez and Ted Cruz.*

4.3 Permutation Test

In order to actually see if the popularity of the groups are changing over time, we run permutation tests. A permutation test takes in two samples and determines the chance that these samples come from the same population. By running this test on our likes for two groups or politicians, we can how similar in popularity our groups are or how different. The distribution that we run our test on is the normalized likes. Each tweet's likes in the current year that we are looking at is subtracted and divided by the mean likes of the previous year. This way we are able to measure the growth rather than raw numbers.

Our null hypothesis and alternative hypothesis are as follows:

Null Hypothesis: The distribution for our misinformation group is the same as the distribution for our scientific group.

Alternative Hypothesis: The distribution for our misinformation group will be different from our scientific group.

For this process, we normalize the growth of likes for each year by calculating the percentage growth from the previous year. We run three main permutation tests to determine comparison of growth. To compare the scientific and misinformation groups, we run a permutation test over each year comparing the distribution of normalized likes for each group. For example we compare Scientific 2015 vs Misinformation 2015 or Scientific 2016 vs Misinformation 2016. This will allow us to see how the growth of the two groups compare with each other. We then run two more permutation tests for the two groups themselves. This test is on consecutive years and is meant to show us if growth for each of the groups is increasing or stagnating. For example Scientific 2015 vs Scientific 2016 and Misinformation 2015 vs Misinformation 2016.

Throughout our analysis, we see that the scientific group shows stagnated growth for years 2013 and 2014. The misinformation group shows stagnated years for 2013 to 2014, 2014 to 2015, and 2017 to 2018. Comparing both of these groups, we that there is similar growth between during 2017.

4.3.1 Permutation Test: Data Visualizations

5 Results

As a result of our investigation, we found that politicians who spread misinformation often have a higher ratio value and less overall likes per tweet. This higher ratio value means that these politicians are more likely to spread controversial information on Twitter. This also shows that people who are viewing their tweets on Twitter are engaging in the politicians' tweets by having much less likes than comments or retweets on the tweet.

In contrast, we see that politicians who spread scientific information on Twitter have lower ratios and significantly more likes on their tweets. This is interesting to note because it shows a clear distinction and result between our two groups.

When comparing the two groups, we see that our scientific group has been steadily increasing in growth over the years while our misinformation group has only been growing significantly in the past recent years.

The overall result of our research shows that a politician has the most growth through spreading non-controversial, scientific information because this yields a steady growth over time in comparison to spreading controversial information.

6 Conclusion

Twitter is one of the largest social media platforms and as more politicians move to Twitter as a means of sharing their political thoughts and opinions, we see that their popularity and reputations are strongly amplified on this major social media platform. The digital world can massively transform the growth of a politician depending on the types of tweets they share.

Our ratio and rolling/cumulative maximum metrics show us that a politician's controversial tweets can heavily impact their audience engagement. Scientific, non-controversial tweets mainly spread by likes while misinformation or controversial tweets spread by having more retweets or comments addressing the tweet.

The permutation test shows us that the growth for politicians who share scientific information has been more steady since they started tweeting, whereas politicians sharing misinformation has only recently started to see a rise in growth.

These distinct patterns show how a politician can grow over time and the amount of influence they have on their Twitter followers and audience online.

The next envisioned steps of our analysis include collecting a larger sample size of politicians in order to compare each politician to a larger sample size. In addition to this, we would include some former and current presidents such as Donald Trump and Joe Biden. We may also expand on more social media platforms because this would allow us to expand our data and see what other types of content politicians are posting. Additional platforms could include Facebook and Reddit.

References

[1] Justin Littman. (2018). TweetSets. Zenodo. https://doi.org/10.5281/zenodo.1289426

[2] Wrubel, Laura; Kerchner, Daniel, 2020, "116th U.S. Congress Tweet Ids", https://doi.org/10.7910/DVN/MBOJNS, Harvard Dataverse, V1.

[3] Seddiq, O., Relman, E. (2020, September 02). Republican Sen. Joni Ernst promoted a far-right conspiracy theory that falsely claims coronavirus cases are inflated by healthcare providers. Retrieved January 25, 2021, from https://www.businessinsider.com/gop-senator-pushes-qanon-conspiracy-theory-on-coronavirus-case-count-2020-9

[4] Zadrozny, B., Collins, B. (2021, January 07). Trump loyalists push evidence-free claims that antifa activists fueled mob. Retrieved January 25, 2021, from https://www.nbcnews.com/tech/internet/trump-loyalists-push-evidence-free-claims-antifa-activists-fueled-mob-n1253176

[5] Words we're WATCHING: What is 'The Ratio' AND 'RATIOED'. (n.d.). Retrieved January 25, 2021, from https://www.merriam-webster.com/words-at-play/words-were-watching-ratio-ratioed-ratioing

A Appendix

A.1 Proposal

For our project on Political Popularity of Misinformation, we want to research the influence politicians have on Twitter, a well known social media platform for users to voice their opinions to a wider audience. The information shared on Twitter that we are interested in will be grouped into scientific information or conspiracy information. Politicians can easily sway public opinion with a simple tweet, therefore we wanted to analyze how much they can influence other Twitter users with their tweets.

This problem relates to the domain replication project because we plan to continue our investigation on the spread of misinformation on Twitter. We will continue to analyze a user's polarity based on a tweet's hashtags presented in the tweets. The polarity score for a Twitter user will show how likely a user will tweet or retweet scientific or conspiracy information. Similar data visualizations will also be created in order to illustrate the technical findings to a visual representation where we can analyze trends and patterns.

Similar to our replication report, we will be focusing on retweets on Twitter. However, our project will mainly focus on politicians as opposed to a wide array of tweets about the COVID-19. In addition, in the absence of hashtags, we'll be using Natural Language Processing to classify tweets as scientific or misinformation. This is because we found that when trying to calculate user polarities, many users don't use hashtags when tweeting. By using the text of a user's tweet as opposed to hashtags that may not exist, we'll be able to better classify tweets as misinformation. We'll also be attempting to analyze characteristics such as the location of users that retweet misinformation to see if there's a correlation between where users reside and how much misinformation they consume and tweet. To analyze reactions to a certain tweet, we'll be looking at their 'ratio'. (1) Previous work has looked at the spread of misinformation of events like the Las Vegas shooting in 2017. (2) Articles have also been written about social media stars that spread voting misinformation. (3) However, previous works have not focused specifically on politicians that tweet out misinformation and how their influence affects the spread of their tweets. Our investigation is interesting because we want to analyze the influence of politicians on Twitter and how their opinions and thoughts online can spread. We will be manually going through politicians' Twitter accounts and determining if they are more likely to spread scientific information or misinformation. We will then obtain their Twitter data through Tweepy's API which will give us information about their Twitter account and individual tweets. We plan to determine the polarity of their Tweet by processing their Tweet text with Natural Language Processing to see which side they stand on. We then find their overall user polarity by summing up the polarity of each of their tweets within a given time frame. This process will be repeated through all unique users who have retweeted or replied to the politician in question allowing us to see the influence of the politician's tweets. We also plan to analyze the community engagement for any Tweet by comparing the number of likes and retweets to the number of replies a user holds. A high ratio of likes and retweets compared to the number of replies will generally mean the Tweet has received a positive reaction whereas a low ratio would indicate a negative reaction. We intend to track the reaction of each tweet a politician tweets over time to see the politicians approval over time. The project output will be a detailed report of our investigation. We will include our analyses and data visualizations to explain the spread of scientific and conspiracy tweets created by different politicians. The details will include how we used polarity and NLP to develop our findings for labeling specific tweets for the two categories. This report will be a significant portion of our findings because it will clearly explain our analyses and how others can replicate the project as well. The data visualizations will also be a crucial part of our project in order to visually provide meaning to our technical findings. This will make it easier for viewers to identify trends and patterns for our datasets. In addition to this, data will be generated by using Twitter's API. This data will be collected and analyzed throughout our research. Looking at the data that we can obtain from tweets, we can obtain the number of likes for a specific tweet, the number of retweets for a tweet, the replies for a tweet, and the text the tweet contains. The number of likes, retweets, and replies allows for us to determine the community engagement and the reaction to see how popular a tweet is through its ratio. The text on any Tweet will be used to determine if the tweet itself spreads scientific information or misinformation which will be done by checking for keywords in the text. We can also obtain Twitter information for any specific user. The most important information we can get from a user is the number of followers a user has. This allows for us to see how the community grows around a user as they spread scientific information or misinformation. By going through any one user's tweets, we can determine the polarity of that user as a number. We can plot the distribution of user polarities from hashtags to see where most people stand in regards to posting misinformation or scientific information. This can show us how polarized a politician's community is toward any one given side. As a result for our overall project, we expect to find that conspiracy information spreads more on Twitter in comparison to scientific information. Politicians with larger followings will have a greater influence on their followers and the types of tweets they produce on the platform. Tweets containing misinformation will have a higher ratio compared to ones containing scientific information. Retweets will also be a major component for conducting our research for politicians on Twitter.