Sentiment Analysis of COVID-19 Twitter Trend

Nicholas Ferry[1] and Xiaomei Zhang, Ph.D.[2]

[1]Department of Computational Science, University of South Carolina Beaufort

April 2020

1 Abstract

Twitter is an excellent source to get a population's opinion on a particular topic. Through 'Tweets' users can express their thoughts in a clear and concise character-limited manner. The project aims to analyze the Tweets created in response to COVID-19 by both users and several well-known media companies, respectively. The chosen analysis that was performed on the collected tweets is sentiment analysis that takes into account the syntax used to construct the Tweet and creates a polarity ranging from [-1, 1] where 1 is the maximum for a positive sentence, and -1 is the maximum for a negative sentence. This analysis was then formed into a graphical representation of the average polarity by date and distribution of polarity by month.

2 Introduction

COVID-19 is a virus that has caused a global pandemic that has infected the world in 2020. By analyzing the influence of this event on a particular population's opinions, we can begin to understand the impact of the event on the mindset of that population. Twitter is a well-known social media platform that allows users to express their opinions in concise 140 character-limited posts known as Tweets. Twitter also allows for a process known as Retweets where users have the option of reposting the same Tweet made by a different user under their account credentials. Twitter is not only used by the general population. Several well-known media companies also utilize the Twitter platform to express their company's opinion in the same manner as a general user can. The Twitter platform proves to be an excellent source to obtain an understanding of a population's opinion on a specific subject, such as COVID-19. This study aims to examine the change of opinion regarding COVID-19 across the United States. The media companies that were chosen are well-known United States media companies. The companies chosen for the study are ABC News®, CBS News®, CNN News®, Fox News®, MSNBC News®, NBC News®, New York Times®, and VICE News®. These companies were chosen for their familiarity amongst the United States general population. Notable dates were then chosen from each company based on influential reporting events made by each company, respectively. Significant dates were also chosen for the general population's Tweets based on reports by ABC News® and the New York Times® that highlighted a timeline of significant events of impact on society caused by the virus.

Python is a high-level programming platform that, in conjunction with APIs, that can perform an extraction of Tweets and perform a form of analysis on these extracted Tweets. Python also boasts the ability to export the extracted Tweets and the analysis performed to a tangible file format and represent this analysis graphically. To extract Tweets, a development environment must first be set up within the Twitter development portal. The SearchTweets API for Python then allows a user to connect to the created development environment and perform advanced searches based on rules selected by the user of the API. Once extracted, the Pandas API enables a user to export the extracted data to a more legible file format such as the CSV file format.

Sentiment analysis is a form of natural language processing and text analysis that categorizes opinions expressed within a writer's text to understand the writer's attitude conveyed within that text. A numerical representation of the syntactical structure of a writer's text can shed

light on the positivity, negativity, or neutrality of the writer's attitude being conveyed within the text. By applying this form of analysis to the Tweets referencing the COVID-19 pandemic, we can understand the attitude of a population's opinion on the Twitter platform. The TextBlob API provides the ability to implement this analysis within Python.

3 Methodology

3.1 Tweet Collection Process

The extraction process begins with the Twitter development environment. Two potential development environments can be utilized based on the allowable time frames: a thirty-day environment or a full archive search development environment. There are then several package levels for each environment that allows for different restrictions on what the environment can achieve. This study utilized both the thirty day and full archive standard Sandbox development environments. The Sandbox product package level restrictions are based on the number of Tweets that can be requested per month, depending on the time frame environment. The thirty-day Sandbox level development environment allows for up to 250 requests and up to 25,000 Tweets, per month. The full archive Sandbox level development environment allows for up to 50 requests and up to 5,000 Tweets per month.

The SearchTweets API is a powerful library that connects Python to Twitter's development environments. The API extraction process first requires a YAML credential file enabling the user to connect to the desired development environment, either the thirty-day or the full archive environment. This YAML files directory is supplied to the load credentials function that supplies the YAML files contents to a variable that will house the credentials contained within the file. The API extraction process then requires a rule describing the desired search requests to be implemented. Specifically, the rule requires the user to specify the search terms and any restrictions on the Tweets containing the desired search term. The potential restrictions on the search term that can be used are limited to the Twitter developer environment product package level and the potential time frames of either thirty days or the full archive that can be searched through. The only restriction that was implemented in this study, within the rule itself, was restricted to the mentioning Twitter account of a Tweet, meaning the user that the Tweet originated from. This restriction was only implemented when collecting Tweets from the studies chosen media companies by specifying each media company's Twitter handle as the restriction. The next setting to decide on within the rule is the desired number of results to obtain from each Tweet extraction call and the desired to and from dates of the desired Tweets. The desired to and from dates were limited based on the development environments allowable time frame. Once the desired settings of the rule have been set, this rule is supplied to the collect results function along with the variable housing the YAML file credentials. This function is then called and the Tweet extraction can begin.

Upon extraction, the Tweets are run through several different self-created functions to separate the relevant information of each Tweet. The first step is to separate the Tweets based on the language used in their syntactical construction. Our goal is to acquire the Tweets written in English only to isolate Tweets created within the United States. Within the same function, after the English written Tweets have been separated, the Tweets text is cleaned by utilizing a regular expression function that removes any special characters and links within the Tweets, and then sentiment analysis is performed utilizing the TextBlob API within a separate function. The TextBlob API is a powerful library that contains functions that can provide natural language processing to produce a sentiment analysis of the desired text. Once the collected Tweets have been separated, cleaned, and analyzed, they are exported to a CSV file by employing the 'to csv' function contained within the Pandas API library. The exported data column headers, and corresponding information to those headers, are: "User Names", "Statuses, "Time Created At", "Sentiment Analysis", and "Polarity". This same process is utilized for both user Tweets and the selected media company Tweets. It should be noted that the user Tweets were collected with both Retweets made and no Retweets made.

3.2 Choosing Relevant Dates

To gather the best Tweet data in regards to COVID-19, a list of significant dates of impact

due to COVID-19 required constructing. A separate list of significant dates was required for user Tweets and each media company selected for this study. Tables one and two display the list of significant dates selected with the first column being the dates selected for user Tweets and the remaining columns being the selected dates for each chosen media company. To select the dates for user Tweets, two sites that contained reports on a timeline of events were used. To select the dates for each media company, impactful reporting events made by each company were researched and used as the dates contained within the list. Each list of significant dates created the basis for the range of to and from dates used when collecting Tweets for both the users and media companies.

User Dates	ABC News Dates	CBS News Dates	CNN News Dates	Fox News Dates
02-28-2020	01-30-2020	02-06-2020	01-31-2020	01-30-2020
03-03-2020	02-18-2020	02-10-2020	02-08-2020	03-10-2020
03-15-2020	02-22-2020	02-29-2020	02-18-2020	03-21-2020
03-17-2020	02-27-2020	03-19-2020	03-04-2020	03-25-2020
03-19-2020	03-12-2020	03-28-2020	03-27-2020	04-01-2020
03-21-2020	03-26-2020	04-04-2020	04-04-2020	04-05-2020
03-23-2020	04-09-2020	04-09-2020	04-10-2020	04-09-2020
03-24-2020				
03-26-2020				
03-27-2020				
03-28-2020				
03-30-2020				
04-02-2020				
04-06-2020				
04-08-2020				
04-10-2020				
04-14-2020				

Table 1: Significant Dates Selected

MSNBC News Dates	NBC News Dates	New York Times Dates	VICE News Dates
02-02-2020	01-30-2020	01-29-2020	02-13-2020
02-08-2020	02-26-2020	02-02-2020	02-27-2020
02-19-2020	02-28-2020	02-25-2020	03-05-2020
03-05-2020	03-05-2020	03-12-2020	03-13-2020
03-15-2020	03-30-2020	04-03-2020	04-06-2020
04-09-2020	04-10-2020	04-10-2020	04-09-2020

Table 2: Significant Dates Selected

3.3 Visually Representing the Sentiment Analysis

With our Tweet data now extracted, cleaned, analyzed, and exported, the results could not be analyzed and represented visually through graphs. The primary results from the data that were of interest for constructing the graphs were the average sentiment polarity per day and the distribution plot of polarity relative to the Tweet count of the polarity per month. To construct the graphs there were several self-created python functions needed to transform and aggregate the data properly into what was needed for plotting the correct graphical representation. Once the data was formulated properly, the MatPlotLib API was used to plot the data on graphs as the final results.

3.3.1 Average Sentiment Polarity Per Day

The objective here was to formulate graphs for each CSV file that represent the average polarity sentiment per day with the x-axis being the date and the x-axis being the average sentiment count for the corresponding date on the x-axis. The first function created was used to create a list that contained the index where the date changed within each CSV file's "Time Created At" column. The next function took the list of indices and computed the average sentiment polarity of the polarities contained within every other index within the list. The function then put these computed averages within a separate list, creating a list of the average sentiment polarity per day for each CSV file. This list of averages in conjunction with the list of dates was sent over to a final function that utilized the MatPlotLib API to form bar graphs of the desired data with each title being the CSV file they represented.

3.3.2 Sentiment Polarity Distribution Per Month

The next objective was to create the polarity distribution for each month per CSV file. With this graphical representation, the desired outcome was a graph with the x-axis representing the sentiment polarity and the y-axis representing the count of Tweets that have the corresponding polarity number. To accomplish this, self-created lists of indices were made containing the indices of each CSV file's "Time Created At" column where the month changed. Then a self-created function that utilized this newly formed list of month changes, formed a list polarities for each month to later be graphed. To graph this data, the histogram function contained within the MatPlotLib API library was used to form a distribution plot of the sentiment polarity Tweet count per month. Additionally, two graphs were made to visually represent the distribution over the entire span of collected dates for both user Tweets and all media companies combined, respectively. These two graphs can be seen in figures one and two.

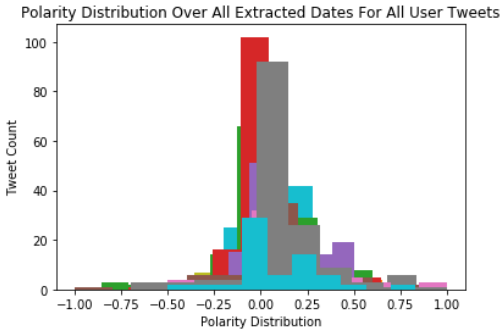


Figure 1: Total Distribution of User Tweets

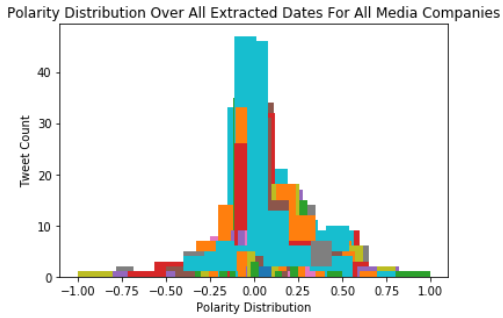


Figure 2: Total Distribution of All Chosen Media Companies

4 Results and Evaluation

The results proved to yield a significant representation of the popular opinion on Twitter regarding the COVID-19 pandemic. Through the formation of the CSV files, an understanding of this opinion by both users and media companies can be formed. By graphically representing these results, a visual understanding of the sentiment polarity change over time can be better understood.

4.1 The Collected Data Within the CSV Files

The extracted and then exported data contained within each CSV file provided an excellent insight into the opinions formed by both individual users and the chosen media companies for this study. The overall sentiment of the user Tweets proved to be mostly neutral which is surprising as you would expect the negative connotation of COVID-19 would correlate to a majority negative sentiment from users. The results of the Tweet collection process for user Tweets were limited to the package level of the Twitter development environment in several ways. The first issue that affected the results was the limit on the number of Tweets that could be collected. Expanding off of this issue, the collection process oftentimes resulted in similar Tweets obtained from a previous collection process. These two issues caused the total user Tweet results to be 2636 for Tweets without Retweets and 2482 for Tweets with Retweets. For the collected media Tweets, the results were only limited by the number of Tweets related to the subject of COVID-19 made by each media company. This is also a very interesting factor as the number of Tweets regarding COVID-19 made by Fox News only amounted to seventy-two whereas the remaining seven companies Tweets count were in a range of 347-392 with the maximum count belonging to the New York Times. The New York Times having the maximum amount of collected Tweets being the highest in comparison to the other companies is not surprising as the pandemic, numerically, affects New York the most at the time of this writing. The factor that is surprising in these count results, is the mear seventy-two collected from Fox News. Fox News is an extremely popular media station and one would expect a number similar to those of other media companies whose popularity is comparable to Fox News. Regardless of popularity, and considering the circumstances of the pandemic, a significant report count should be expected from any media company. It should be noted that the dates selected contained in tables one and two play a role in this count but should not be detrimental enough to create a 200 plus difference in Tweet count between Fox News and the other chosen media companies.

4.2 Average Sentiment Per Day

4.2.1 User Tweets

Figures three and four show the average sentiment polarity per day for user Tweets with Retweets and user Tweets without Retweets, respectively. Both figures provide surprising results as the average polarity tended to lean towards a positive average rather than a negative average polarity. The expectations would be a strong negative average due to the negative implications of COVID-19. At the very least, the expected result was for the average to be closer to zero due to the majority of sentiment being

neutral. While the average sentiment polarity does lean on the positive side, the polarity number fluctuates throughout the timeline. This fluctuation can be caused by specific events that happened that day where a less positive polarity number can be from a notable event and a high positive polarity number can be caused by a lack of notable events for that specific day. The results prove that while a pandemic like COVID-19 is detrimental to society, people tend to lead with positive attitudes towards the situation regardless of the negative impacts. The difference between the average polarity with Retweets vs. with Retweets does not seem to cause a huge impact on the resulting average sentiment. This is another surprising result as Retweets count for a multitude of the same Tweet which should affect the average more than it appears to.

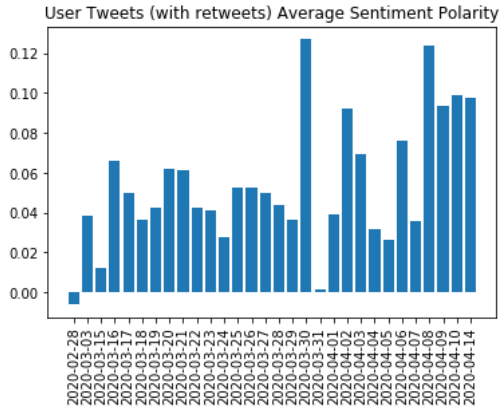


Figure 3: User Tweet Average Sentiment Polarity By Date With Retweet

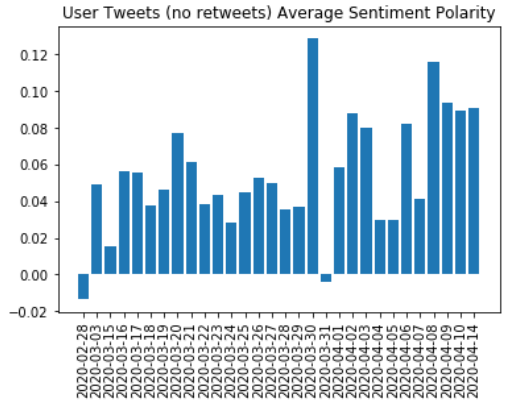
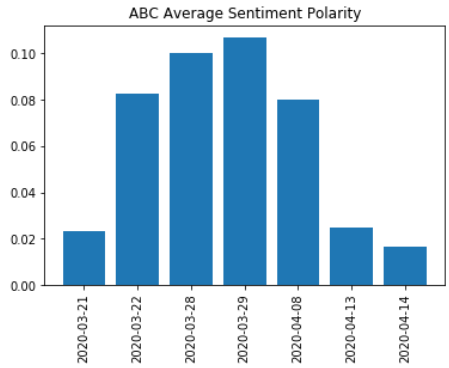


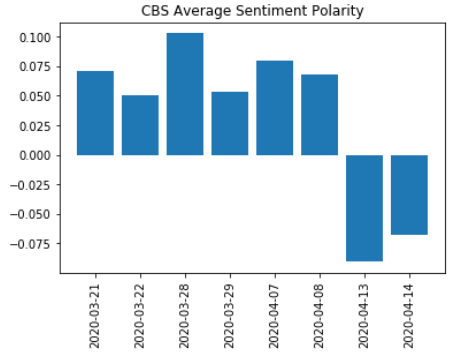
Figure 4: User Tweet Average Sentiment Polarity By Date Without Retweet

4.2.2 Chosen Media Company's Tweets

The resulting average polarity graphs for each media company overall are not surprising. The average polarity seemingly correlates to the sentiment of the report made on the corresponding dates about the COVID-19 pandemic. Most of the media companies, except for CBS News, Fox News, and VICE News, had an overall positive average sentiment polarity during March. However, in April, almost all of the media companies took a tremendous dive in average polarity and the average value dropped tremendously in comparison to March. The media company that had the most fluctuation was VICE News which is notoriously known for factual focused reporting thus in our opinion, this makes VICE News the most influential media company data to examine. To save space, we have displayed only the media company's graphs with the most prominent results. Figures five and six contain the chosen company's results to display.

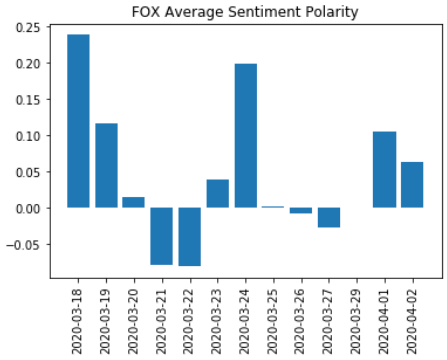


(a) ABC News

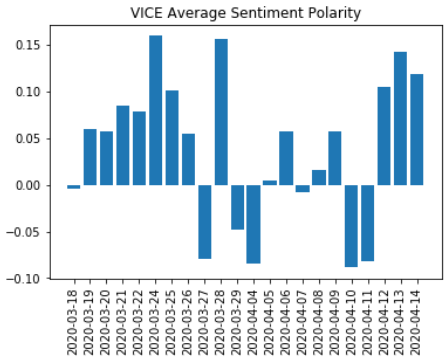


(b) CBS News

Figure 5: Some of the Chosen Media Company's Average Sentiment Polarity Tweets Per Day



(a) FOX News



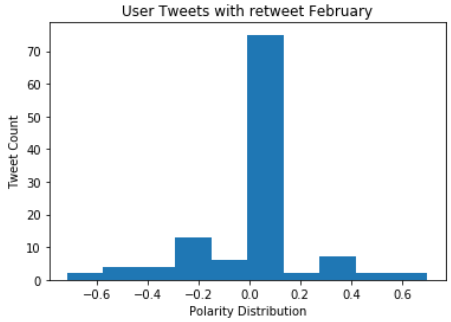
(b) VICE News

Figure 6: Some of the Chosen Media Company's Average Sentiment Polarity Tweets Per Day

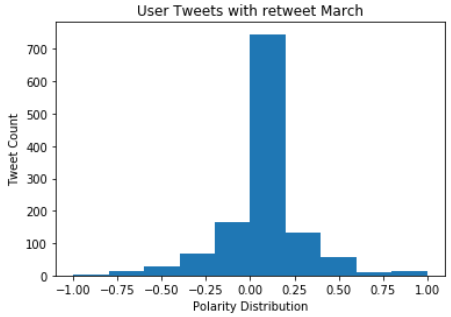
4.3 Polarity Distribution by Month

4.3.1 User Tweets Polarity Distribution

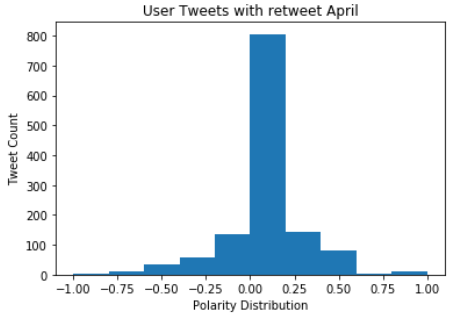
The polarity distribution provides a better insight into the count of each polarity number. This insight provides a better understanding of the popular opinion throughout February, March, and April. February's results are limited in comparison to March and April due to the number of Tweets collected in February being below 120. The results prove the original visual analysis to be true where most of the user Tweets are centered around a neutral sentiment polarity. However, although most tweets are centered around neutral, they are mostly positive proving again that the common opinion during the time from the collected Tweets is mainly positive. This result holds for all three of the months resulting from the Tweet collection. Another noticeable factor from the results is that the negative Tweets were more prominent in March and were almost non-existent in April. This is due to the increase in the significance of the impact of the pandemic increasing in March in the United States with the primary event of significance being state-implemented quarantines. The only primary difference between the graphs displaying user Tweets with Retweets and user Tweets with Retweets is the overall count for each sentiment polarity level. As expected, the user Tweets with Retweets have a higher count at each sentiment polarity level. Figures seven and eight display the results of the average sentiment polarity for user Tweets with Retweets and user Tweets without Retweets.



(a) February

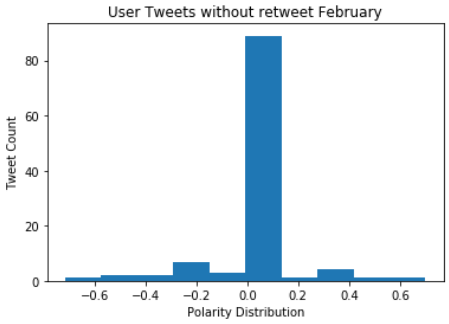


(b) March

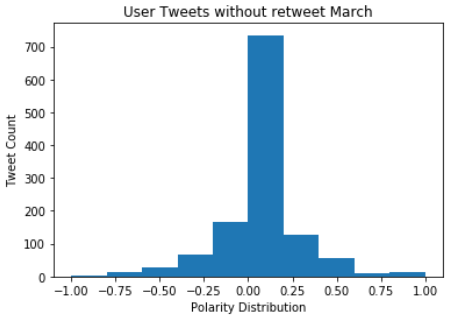


(c) April

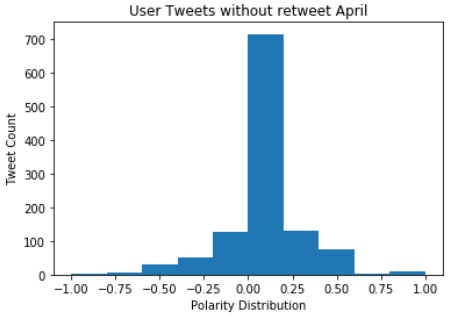
Figure 7: User Tweets Sentiment Polarity Distribution With Retweets



(a) February



(b) March

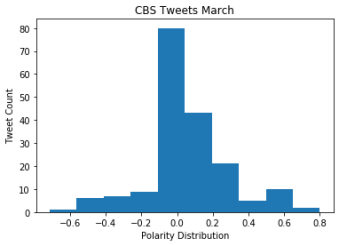


(c) April

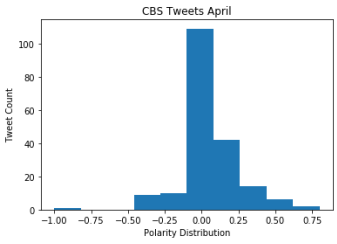
Figure 8: User Tweets Sentiment Polarity Distribution Without Retweets

4.3.2 Chosen Media Companies Polarity Distribution

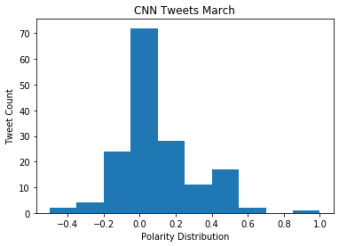
The sentiment polarity distribution graphs for the chosen media company's Tweets include March and April as these were the months that resulted from each of the Tweet extractions performed on the chosen media companies. The results from the sentiment polarity distribution differ from the average sentiment polarity in that most companies had a close to neutral tone throughout each month. However, it appears that CBS, CNN, and NBC had a large count of negative sentiment polarity Tweets in March. This negative count can be again attributed to the increases in the significance of the impact of COVID-19 on the United States during March. One of the most surprising results of the polarity distribution graphs is the majority of the Tweet count by the New York Times being positive. This is surprising as New York is the most heavily impacted state in the United States. Figures nine and ten display most significant sentiment polarity distribution graphs.



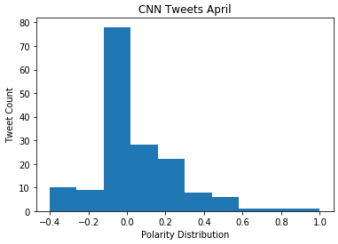
(a) CBS News March



(b) CBS News April

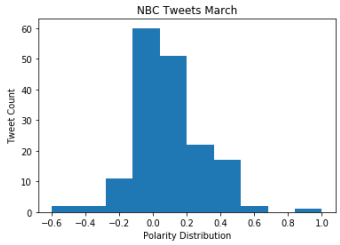


(c) CNN News March

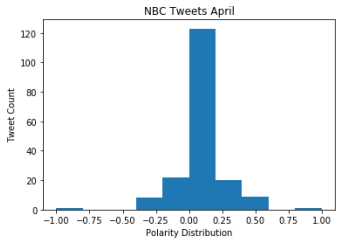


(d) CNN News April

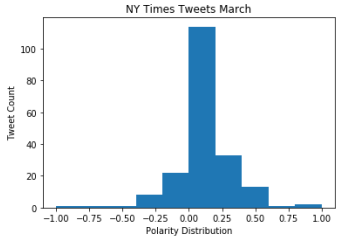
Figure 9: Some of the Chosen Media Company's Sentiment Polarity Distribution by Month



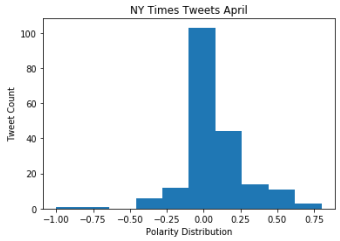
(a) NBC News March



(b) NBC News April



(c) New York Times March



(d) New York Times April

Figure 10: Some of the Chosen Media Company's Sentiment Polarity Distribution by Month

5 Future Works

In the future, we would like to perform the Tweet extraction based on geographic location to analyze the change in sentiment linked to geographical locations. This additional analysis parameter would provide an insight into how different geographic locations are impacted by the COVID-19 pandemic. These different insights can then be compared to postulate how different cultural backgrounds affect a population's opinion. Another additional form of analysis that we would like to perform would involve analyzing the emojis used in Tweets. Emojis help to understand the tone trying to be conveyed within the syntax which is often unrecognizable strictly from a writer's syntactical structure. The tone of the Tweet would improve the overall sentiment analysis.

6 Conclusion

The overall process brought many welcome challenges but resulted in subsequent answers to the objective question: what is the population's reaction to the COVID-19 pandemic? This study in its limited capacity answered these questions through graphical representations of the collected data. This study included many prominent APIs and a powerful form of natural language processing. The CSV file extraction allows for a concise way to view the collected Tweets and the analysis performed on them. The graphical representation of the results provides an excellent insight into the significance of the collected Tweets.

7 References

1. http://www.twitter.com
2. https://developer.twitter.com/en/
3. https://www.python.org/
4. https://pypi.org/project/searchtweets/1.0/
5. https://textblob.readthedocs.io/en/dev/
6. https://matplotlib.org/
7. https://pandas.pydata.org/
8. https://numpy.org/
9. https://www.nytimes.com/article/coronavirus-timeline.html
10. https://abcnews.go.com/Health/timeline-coronavirus-started/story?id=69435165