Analyzing Covid -19 Through a Sentiment Analysis of Twitter Data

Cameron Catana, Professors Jon Chun and Katherine Elkins
IPHS Senior Seminar, Fall 2021

Introduction

Covid-19 has consistently proved to be difficult to analyze and model. An article from fivethirtyeight.com (Koerth, Bronner, and Mithani, 2020) detailed some of these struggles through the example of generating a predictive model of total known deaths from Covid, due to infection rate, fatality rate, and the population. To display the difficulties in such a seemingly simple model, I have attached an image from the article (Figure 1) which shows the numerous factors that go into just fatality rate. With this knowledge in hand, I looked for an abstract method of modeling

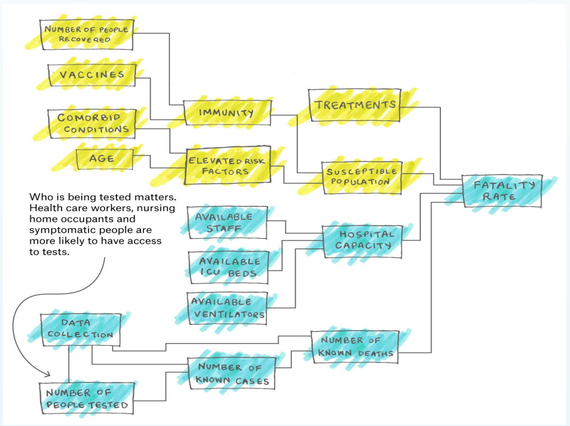


Figure 1

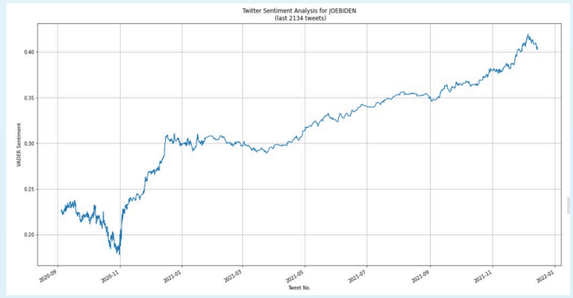
and analyzing Covid 19: twitter data. Twitter data is interesting to look at for this project because it allows us to view and interpret the words of public officials and analyze how they feel about certain important events, or at least how they want people to think they feel. From this project, I aimed to achieve a better understanding of how people really felt about Covid versus how elected officials wanted people to feel about Covid. The initial goal was to determine if there was a disparity between the words of the people and the words of those in power. For any elected official for which this was the case, such a disparity would mean that the officials are not aligned with what the people really want, which would make for some incredibly interesting analysis. Unfortunately, that goal could not quite come to fruition, as the "words of the people" proved nearly impossible to quantify for reasons I will discuss in the Methodology. I was able to generate results to essentially quantify the words of those in power, though, which posed a new, interesting question: how did different figures respond to, and treat, covid-19 through social media?

Methodology

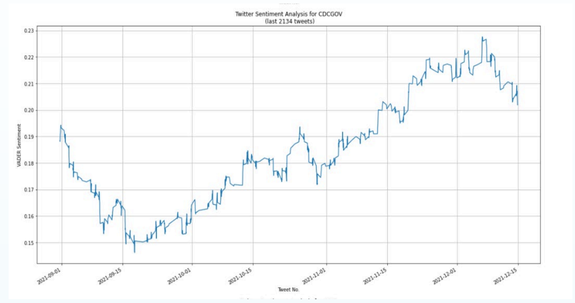
To achieve this task, I first went about gathering twitter data. This was accomplished through a jupyter notebook written in Python, with the help of Professor Jon Chun, coming with a library called Tweepy. Tweepy allows the access of the Twitter API, which allows a programmer access to Twitter in other ways which normal users do not have access to, such as collecting data through twitter via tweet extraction. I used Tweepy to extract tweets and use them as data point. The next part of the notebook was to retrieve tweets by both twitter username and by keyword, which were specified in the notebook. The tweets were then cleaned using a Python library called "tweet-preprocessor", before eventually I did sentiment analysis on the twitter data, using the Valence Aware Dictionary and sEntiment Reasoner (VADER). VADER works well for this project because it is specifically attuned to sentiment found on social media. The users selected were Joe Biden, the CDC, the WHO, Greg Abbott, who is the governor of Texas, and Gavin Newsom, who is the governor of California. I chose Biden as he was President during some very crucial turning points on the Covid timeline, such as the rollout of vaccinations, and I wanted to see how well his tweet sentiment matched with key events like that one. The CDC and WHO were chosen because they have no inherent political interest and can act as somewhat of control variables. Since they have no ulterior motives to express false or misleading sentiment, the sentiment reflected by those accounts should be an unbiased representation of Covid sentiment, with numbers possibly trending downwards due to these agencies' goal of keeping people safe and healthy, and possibly being overly cautious as a result. Abbott and Newsom, as governors of two very important and very opposed states, are included to try and detect any kind of polarity. The hypothesis is that Newsom and Abbott will have very different reactions to key Covid events, and therefore have significantly differing sentiment levels. The keywords which I selected were "masks", "vaccines", "travel", and "Covid." I chose these words because I believe they are most accurately correlated with Covid sentiment, due to all of them being specific to Covid and unlikely to encounter sentiment from unrelated tweets as a confound. Unfortunately, I encountered a technical error in collecting keyword tweets, as since the population was all tweets, there were simply too many tweets to see any relevant results without downloading an infinite number of tweets. Due to this, I looked at the user accounts and analyzed the hypotheses of sentiment behavior that I predicted earlier.

Results

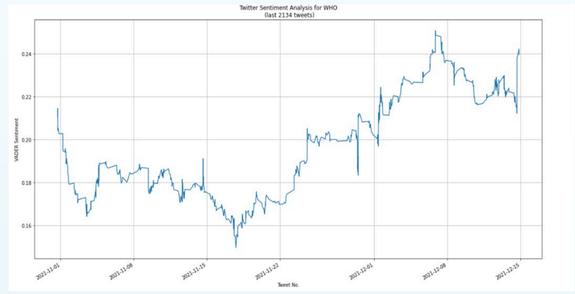
My results consist of the five different user accounts and their sentiment graphs.



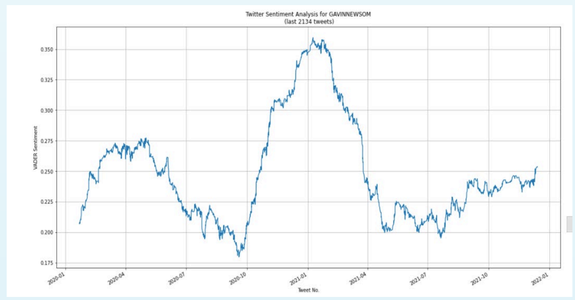
Joe Biden



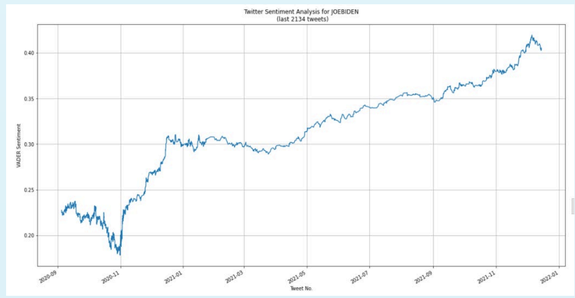
The CDC



The WHO



Greg Abbott



Gavin Newsom

Discussion and Conclusion

The most important thing to note about these results is that these graphs are non-standardized; they have different values on their x and y axes. Therefore, it is important to keep that in note when evaluating the quality of these results. First, from the Biden graph, we can see that his sentiment has generally been trending upwards since a low of below .20 right before the election. This makes sense, as Covid conditions in America have certainly improved over the past year. Next, I want to take a joint look at results from the WHO and the CDC. These accounts both have much higher tweeting volume than the other accounts, so I was only able to retrieve data from September 2021 for the CDC, and November 2021 for the WHO. We can still analyze this data while taking that fact into consideration. As I predicted, sentiment levels for these two accounts were generally lower than the other three that I used, with sentiment levels for the governors and Biden ranging from around .2-.4 in the period between September 2021 and the present day, whereas the sentiment levels for the CDC and the WHO reach a max of roughly .24 and go as low as .15. There are two possible conclusions to be drawn from this. The first is that my hypothesis was correct, and elected officials try to paint a better picture than what exists, or these organizations display lower sentiment since they are not trying to convince people to vote for them, and therefore can paint a more accurate picture of reality. The second is that the governors' and Biden's account tweets about many things unrelated to Covid, which could also explain the higher sentiment. To examine this further, let's take a look at some of the tweets from Biden and the governors during this time period. From looking at Biden's account, there are many tweets during this time period which display positive sentiment and are not related to Covid, such as "We are going to reinvest in our country and our people to continue building a better America with union jobs.", which was tweeted by Biden on December 10th. However, there are also many Covid related tweets displaying positive sentiment, such as, "Last year, more than half of the schools in America were closed. This year, 99% of schools are back open, and 20 million children ages 5 and older have already received their vaccine. Let's keep it going.", which was tweeted by Biden on December 8th. I can conclude from this that it is likely a combination of both factors; the higher sentiment is achieved by both the desire of an elected official to motivate and inspire the people, and also by unrelated non-Covid positive tweets. The last set of graphs to anaylze is the governors, as they are intended to be compared against each other to assess polarity. Interestingly enough, we can see quite similar sentiment levels displayed by both governors. We can see that they both experience a decrease between the start of Covid in March 2020 to the end of 2020, an increase in early 2021 when vaccines began to be rolled out, and a culmination of another decrease in summer-fall 2021, when Covid began to mount a comeback. From these results, it is possible that polarity between governos may actually not be as stark as I initially anticipated. More data collection is necessary to draw a more certain conclusion, but these results indicate that elected officials are trying to figure out good solutions during difficult times, and simply use different means to solve these problems, rather than the alternative, which is that partisan elected officials actually see the world differently and that is what is causing such a great divide throughout America.