STEM Fellowship Big Data Challenge 2020

Public Health in the Digital Age: Analyzing the Relationship Between the Influence of Twitter Social Media on Social Distancing Measures and COVID-19 Cases in the United States

Grace Huang[1], Jessie Wu[1], Nicholas Chung-Hun[2], and Varun Mokashi[2]

[1]University of Western Ontario
[2]University of Waterloo

June 1, 2020

Abstract

The public health strategy to "flatten the curve" has recently become a widely trending phrase across social media platforms as a means to share content related to COVID-19. Given today's digitally-focused society, we aimed to investigate the relationship between the Twitter tweet sentiments on critical COVID-19 social distancing policies in the United States (US) and the presence of COVID-19 cases. Using data sets from John Hopkins and Harvard University, we established the relationship between Twitter sentiment, as collected through natural language processing sentiment analysis techniques, and the change in COVID-19 cases in each US state over time. This uncovered potential to build a predictive growth model based on the Granger-Causality test examining how sentiments expressed through tweets would set a precedent for fluctuations in new COVID-19 cases. Additionally, we sought to understand the degree to which social media externally influences users to drive certain social distancing behaviour by building a linear regression model analyzing the relationship between likes and retweets of tweets expressing different sentiments and growth of cases. As a result, our conclusions suggest a three day time-lag relationship between social distancing tweet sentiments and fluctuations in COVID-19 cases. The conducted study was limited by virtue of varying state laws, discrepancies in state populations, and the exclusion of non-English tweets, which elicit the need for multiple focused studies that more accurately explore these phenomena. In the future, iterations of the predictive growth model should account for more comprehensive tweet data sets in tandem to the geographical influence for the spread of the virus on a day-to-day timeline. Subsequent studies should also investigate leveraging social media as a tool to maintain or improve public health.

Keywords COVID-19, natural language processing, sentiment analysis, social distancing, Twitter

1 Introduction

Since the World Health Organization declared COVID-19 a pandemic on March 11, 2019, many countries had already begun implementing non-pharmaceutical interventions by restricting travel to minimize the spread of the virus [1]. The coronavirus has a R-naught value, which is the estimated new infections that will occur from one case, of roughly 2 to 2.5, making it more contagious than the seasonal flu [2]. Thus, given the absence of pharmaceutical treatments, the key step of imposing social distancing measures was implemented by many countries to prevent the spread of the virus and "flatten the curve."

Many scientific studies have confirmed the positive impacts of social distancing on decreasing both the spread and mortality rate of the coronavirus, with the latter partially due to preventing an overwhelming number of hospital intensive care units [3]. Historically, during the outbreak of severe acute respiratory syndrome coronavirus (SARS-CoV) in 2003, non-pharmaceutical interventions such as quarantine, border controls, and contact tracing successfully limited the spread of transmissible disease [4]. Researcher Joel Koo and colleagues "adapted an influenza epidemic simulation model to estimate the likelihood of human-to-human transmission" of the COVID-19 disease in a simulated Singaporean population [4]. Considering three infectivity scenarios, the combined intervention, encompassing self-isolation, school closure, and workplace distancing reduced infection rates to the greatest extent [4].

For the first time, a global health pandemic has surfaced in a digital age, where social media has become the primary outlet for spreading news, movements, and personal sentiments. The Twitter social media platform's design elements allow an immediacy that encourages its users to share "stream-of-consciousness" content that fully expresses sentiment toward social distancing measures. This allows for data to be extracted as a means to predict the degree to which individuals will take social distancing precaution. In tandem with allowing users to acquire a follower base and interact with other Twitter users through acts of "retweeting" to multiply the spread of information and sentiments, users can also use hashtags to start social movements through a phenomenon coined "hashtag activism" [5]. By investigating the overall tweet sentiment of users who used social distancing related hashtags, we can begin to understand social media's contribution to herd mentality and algorithms that prioritize content which leads to negative public health effects.

In the US, COVID-19 cases have been rising sharply and has quickly become the country with the most cases [6]. With newly imposed social distancing government policies, different US states have polarized views. Across the country, backlash from anti-quarantine protests and responses in social media posts surrounding these new practices are rising [1]. The dominating role social media platforms play during this pandemic inspired the research question for this paper: how does the sentiment around social distancing expressed on social media differ from state to state, and how is this reflected in the extent to which US citizens are abiding by stay-at-home orders?

This investigation aims to study any relationship between social distancing tweet sentiments and COVID-19 cases over time, separated by state. Establishing this correlation would act as a predictor of the growth of cases in the 50 states over time, as well as provide greater insight into the degree to which social platforms influence individuals' beliefs and behaviour.

2 Materials & Methods

2.1 Data Collection

First, the US COVID-19 cases dataset used for this investigation was obtained from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [7]. The dataset provided COVID-19 cases over time between March 3rd, 2020 and May 1st, 2020 broken down by the city in each state. The data was filtered to combine all the cities in the same state into a single entity for data of cumulative cases over time separated by state [8]. Next, the Twitter tweets dataset was obtained from Harvard University. The dataset provided IDs for tweets relating to COVID-19 [9]. To collect all tweets with the targeted social distancing topic for the purpose of this investigation, the tweets were further filtered by searching the list for geotagged tweets containing the following hashtags:

*"socialdistance, socialdistancing, stayhome, lockdown, shutdown, quarantine, stayathome, stayin, stayinside, 6feetapart, sixfeetapart, selfisolation, isolation, prosocialdistancing, f***corona, endthelockdown, endtheshutdown, f***socialdistancing, f***covid, openup, letusfree, freedom, lifttheban, reopen, liberate, coronaisfake, antisocialdistancing"*

The tweets were then hydrated back to their original form using Twitter's API to receive the text of the tweet, the location of the tweet, and the date of the tweet. Due to the scope of the research, the tweets were further filtered by language and country code to only include those written in English in the US.

Afterwards, sentiment analysis was performed using Google Cloud Natural Language Processing API to obtain sentiment scores ranging from -1 (negative sentiment) to 1 (positive sentiment). Given the 1.4% API error rate, a portion of tweets were failed to be analyzed and were given a sentiment score of 0 (neutral) by default by the API. This faulty data was omitted from the data set, along with the rest of the tweets with a neutral score. Due to the nature of this investigation, it was most appropriate to analyze values indicative of decisive sentiments. This sentiment data was then averaged to produce a mean tweet sentiment data set by state [8]. Due to Twitter's API policy that limits the dissemination of full tweet information, the text from each tweet has been removed in the datasets that are provided. However, the sentiment data extracted has been kept in the datasets.

2.2 Statistical Analysis

Also, for the purpose of investigating how the spread and reach of Twitter content might have a direct influence on the change in COVID-19 cases, we recalculated the sentiment scores by factoring in each tweet's likes and retweets. Tweet likes were assigned an impact value of 0.5X that of the original tweets as it represents a passive agreement with the sentiment of the liked tweet that passes into the Twitter algorithm's of giving content with higher levels of engagement more visibility. The same is true with retweets, which was assigned an impact value equal to that of the original tweet as it presents an active intent to spread the tweet and simultaneously fuels the Twitter algorithm. Based on these adjusted tweet sentiment values, the mean sentiment values for each state were recalculated to more accurately display likely more negative or positive overall sentiments [8]. This was analyzed against the growth rate of COVID-19 cases over time from March 3rd to May 1st [8].

Using this extracted data, the COVID-19 cases and tweet sentiment data were first displayed in isolation to show its changes over time, then a linear regression model was built to examine the relationship between mean tweet scores and total COVID cases in different states. For a comprehensive display of all variables, data visualization using Tableau Dataprep was used to create a bubble map. This map visualized the number of COVID-19 cases in each state over time with each state's fixed mean tweet sentiment value was created, where number of cases is presented by circle size and sentiment value is represented through circle colour. Finally, a linear regression model was built around the growth rate of COVID-19 cases in each state and degree of tweet spreading on the platform to investigate the influence of social media on the number of cases.

As a final test, the Granger-Causality test was used, which is a hypothesis to determine whether data from one time series can be used to forecast another. In our case, it was looking at whether negative tweet sentiments can help us predict coronavirus cases over time.

Overall, all statistical analyses and graphs were developed using Python in Google Colaboratory. A heatmap visualization over time was created using Tableau.

3 Results

3.1 Pattern Analysis of COVID-19 Cases and Tweet Sentiment Scores Individually

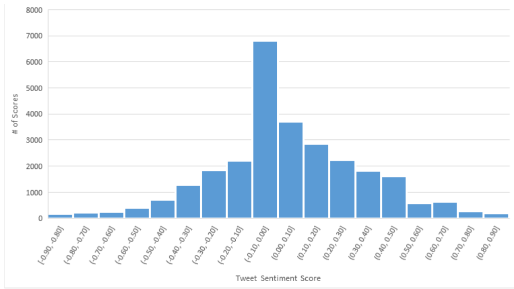

Figure 1: Distribution of Tweet Sentiment Scores

After gathering the results of the Twitter sentiment analysis, we first wanted to see the distribution of the scores. As seen in Figure 1, the scores were quite normally distributed, with a large number of scores near or at 0, indicating neutral sentiment. There were also slightly more positive scores than negative scores, indicating that the majority of sentiments expressed on Twitter regarding staying at home was positive in the US.

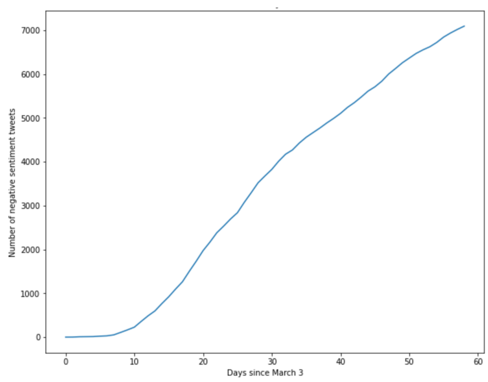

Figure 2: Cumulative number of negative social distancing tweets over time (March 3 - May 1, 2020)

As shown in Figure 2, we displayed how negative tweets increased over time so we created two different plots showing cumulative negative tweets over time and percentage of negative tweets over time.

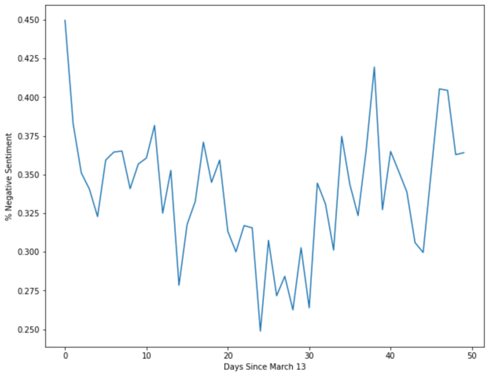

Figure 3: Percentage of number of negative social distancing tweets per day (March 13 - May 1, 2020)

In Figure 3, we displayed the percentage of negative tweets per day to determine what proportion of the social distancing related tweets in the US as whole were negative. With data beginning from March 13, when US President Donald Trump declared an emergency order, we can conclude a decrease from an initially high proportion of negative tweets, then slight overall increase in proportion of negative tweets.

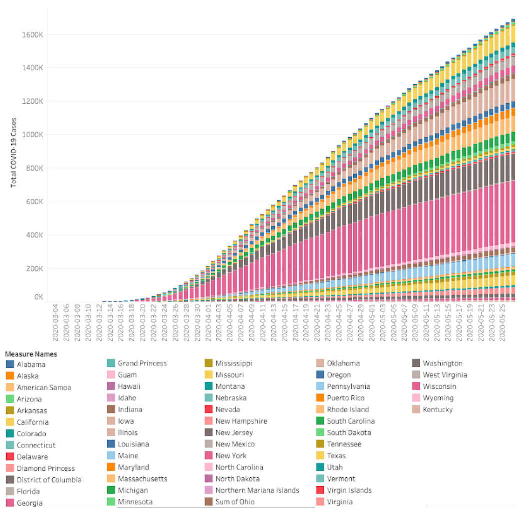

Figure 4: Cumulative number of COVID-19 cases, per state

Figure 4 shows the cumulative number of COVID-19 cases in the USA, with each bar of the graph divided into each state, indicating the proportion of the total cases each state makes up.

3.2 Correlational Analysis of COVID-19 Cases and Tweet Sentiment Data Sets

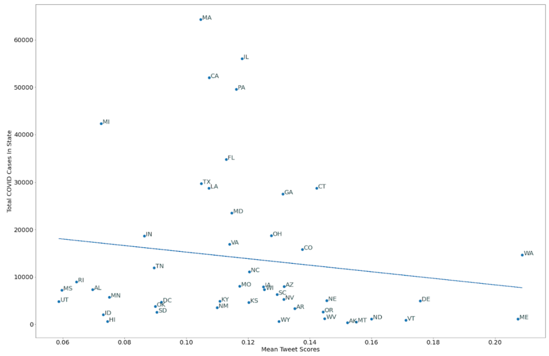

Figure 5: Cumulative number of COVID-19 cases, per state

In Figure 5, the mean tweet sentiment scores across the two months of March and April are plotted against the total cumulative COVID-19 cases, beginning April 30, 2020, in each state. After conducting an outlier test, the identified outliers of New York, New Jersey, and New Hampshire were removed so the data could be better modeled. The regression line can be modeled with $y = -69182x + 22147$, and it has an r value of -0.15.

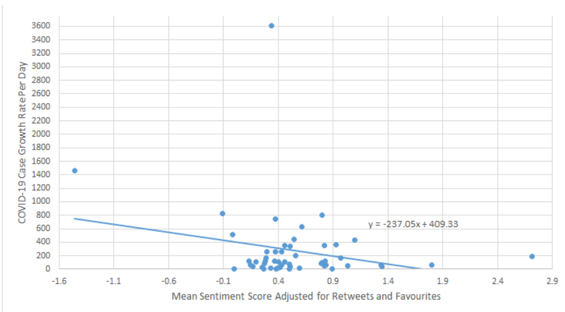

Figure 6: Relationship between adjusted mean sentiment score and COVID-19 case growth per day in the US

By plotting the mean adjusted sentiment score against the rate of change of cases, a linear regression analysis presents a weak negative relationship.

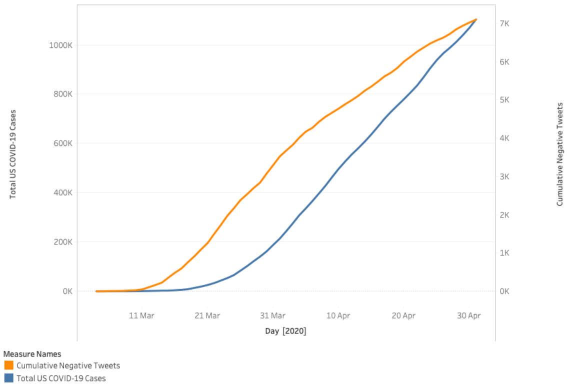

Figure 7: Cumulative number of COVID-19 cases and cumulative number of negative tweets over time from March 3, 2020 to May 1, 2020.

In Figure 7, cumulative number of tweets with negative sentiments regarding social distancing (right y-axis, orange) was plotted against the cumulative number of COVID-19 cases in the USA (left y-axis, blue). Visually, the shapes of the two plots appear similar, with a slight time lag between the increase in negative sentiments starting at around March 16th and an increase in COVID-19 cases starting at around March 22nd.

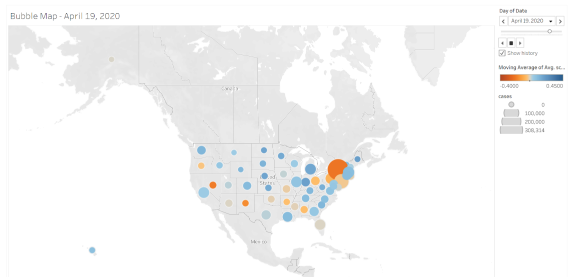

Figure 8: Bubble map which depicts the relationship between the moving average of Twitter sentiment with regards to the number of cumulative COVID-19 cases per state, where each circle colour represents sentiment and the larger the circle size, the greater the number of cumulative COVID-19 cases. This map is animated to show the changes in both variables, or each circle, over time. The interactive bubble map can be found here.

Through an interactive view, the changes with respect to tweet sentiments and the simultaneous change in COVID-19 cases over the period March 3, 2020 to May 1, 2020 is shown. Specifically, the changing sentiment values are a moving average, where for each state, each day's sentiment is calculated by taking the average of the sum of the current day's average sentiments and previous its two days. This shows a more current representation of user sentiments and shows how recent sentiment trends online pose a delayed effect on encouraging or discouraging social distancing behaviour, which is reflected in the cases count.

Overall, the bubble map highlights the volatility of tweet sentiments over time - many states fluctuate between having positive and negative average sentiments and the map display isolates its effect on changes in cases as seen in the varying degrees to which the circle sizes increase.

4 Discussion

Overall, the results demonstrate that there is a weak correlation between the negative sentiments on Twitter and COVID-19 cases, however there is some potential in using the Twitter tweet sentiment data to predict the COVID-19 data. In Figure 5, the weak negative relationship suggests correlation between tweet sentiment and total number of COVID-19 cases. This correlation roughly indicates that states with lower mean sentiment scores would practice social distancing to a lower extent and would raise possibilities to build predictive models for each state forecasting the number of COVID-19 cases based on users' online sentiments.

Moreover, the majority of the states who have high incidence of COVID-19 cases (above 15 000 as of May 1, 2020), had mean sentiment scores that were closer to a negative sentiment. This is suggestive of a greater proportion of negative tweet sentiments regarding social distancing, and a higher chance of those individuals disobeying distancing orders, which may be correlated to higher COVID-19 cases. While this relationship is not highly statistically significant given an r value of -0.15, it may be useful to revisit this topic as data is released for the months of May and onward to determine how the relationship between tweet sentiments and COVID-19 cases change across a longer period of time.

In Figure 6, the negative relationship suggests that although weak, social media interaction plays a causal role in the thinking and behaviour of users, especially given the widespread discussion of the pandemic online. This would in turn affect the COVID-19 growth rates per day. It is important to note that by adjusting the sentiment scores for retweets and favorites, this considers the role Twitter's algorithm plays in proliferating content, thereby increasing engagement with content expressing certain sentiments.

The shape of the graph seen in Figure

7 is highly similar to the graphs of cumulative negative sentiment tweets and cumulative COVID-19 cases over time for each state (not shown). The state graphs show a similar pattern where the cases of COVID appear to lag a couple days behind the cumulative negative sentiment line. Thus, it could be concluded that increases in negative sentiment tweets may be predictive of COVID-19 cases. For Figure 7, we ran a Granger-Causality test, which is a hypothesis to determine whether one time series can be used to forecast another. Following the test, it was concluded that a time-lag of 3-4 (days) resulted in the smallest p-value, indicating that we could reject the null hypothesis that the tweet data does not explain the variation in COVID-19 data. Granger causation describes a state of precedence between two time series variables - in our scenario, we can conclude that negative tweet sentiments "Granger-cause" COVID-19 cases (i.e. the negative tweet sentiments contain data that can help us predict coronavirus cases over time, with a time lag of approximately 4 days).

The animated bubble map in Figure 8 provides insight into the geographical spread of COVID-19 while also displaying trends and fluctuations in the moving average for tweet sentiment. From a purely geographical perspective, it is evident that the majority of cases are clustered around New York and the surrounding states within the northeast. The colours represent a moving average of tweet sentiments as the pandemic progresses. There appears to be several fluctuations within the states as per their moving average, however there are certain time stamps that provide a unique opportunity for insight. For example, March 6 shows a unanimously negative sentiment towards social distancing across the United States. Mid-March and April is when the size of the bubbles begin to expand rapidly (that is, COVID-19 cases increase dramatically), primarily in the northeast. As these cases rise meteorically, the tweet sentiments trend towards the positive spectrum, albeit with fluctuations towards the negative side on days such as April 19 and April 29, emphasizing the variance in tweet sentiments.

The data within the Figure 8 map offers the greatest opportunity for extension, particularly in predictive modelling with respect to using the tweet sentiments as a feature to predict COVID-19 cases. The time lag of the moving average concluded from the Granger-Causality test, which suggested that there is data within the tweet sentiments that can be leveraged to predict COVID-19 cases, offers the opportunity to explore various predictive regression models at the state and national level while considering geographical proximity. Specifically, activity that fails to respect social distancing laws have the potential to impact the spread of the disease in nearby states, so in further explorations, it would be significant to model the impact of the tweet sentiments in tandem with the geographical influence on the spread of the virus.

One limitation of this study is the presence of confounding variables, namely differing timelines of each state's lockdown laws and the differences in each state's population size. Moving forward, it would be more accurate to represent COVID-19 cases over time as a relative value of cases per 100,000 individuals, as it would prevent the amplification or compression of case count data. Alternatively, it would be viable to split the analysis into low and high density states for investigation with low level scope to improve accuracy. Another limitation derives from the tweet data being restricted to English only. This made our data set exclusive to areas with more speakers of Spanish or other languages, which could create bias in our data. To prevent this, greater efforts to ensure data is more comprehensive would further randomize our data set to ensure accurate results. Finally, in subsequent studies, analysis of the sentiment score variance of each state can be conducted using box-and-whisker plots. By determining the variance, we can examine the degree of sentiment polarization within each state and how changes in polarization affect the dynamics of COVID-19 cases.

5 Conclusions

Through the analysis of the data that we obtained about tweets sentiments and COVID-19 cases, it is shown that, although weak, there is a relationship between tweet sentiments towards COVID-19 cases. The Granger-Causality test revealed that the tweet sentiments can be used to forecast COVID-19 cases in a geographic area to some certainty. Moreover, it is shown that tweet engagement such as retweets and favorites do play a large role in spreading sentiment. The ability that individuals have to express their opinions on any subject to the world is an important consideration when analyzing human behaviour especially with the growing integration of technology into individuals' lifestyles. Given the findings, relationships concluded in this report can be adapted to build a predictive model of COVID-19 cases focusing on specific individual geographical region or grouping geographical regions with similar characteristics.

This investigation's findings hold significance for a number of reasons. In the domain of public health, it would be beneficial to conduct further studies using more comprehensive data sets of tweets to track retweets and likes on a daily basis, indicating to what extent social media content is spreading across the platform. This value representing the amplification of tweets with both positive and negative sentiment values will be analyzed against the daily change in COVID-19 cases to ultimately track the aftermath effects of social media platforms on individuals' behaviour on a timescale. Additionally, it would be interesting to track the organization of anti-lockdown protests and rallies on social media to determine its location, predict its subsequent impact on the spread of disease, and take appropriate action to dispel such events that threaten public health. On a broader scale beyond the COVID-19 pandemic, more studies on the influence of social media on public health could provoke important conversations with social media companies. This industry-government partnership would allow social media's opaque, fast-changing algorithms to be used as a tool to implement public health surveillance and improve the accuracy of health information.

The information presented in this paper should be revisited in the future as more data available. While there is a significant correlation found at this time, the limited Twitter data that was accessible and identified limitations may not have been appropriate to draw a conclusion for social media as a definite causal factor of the spread of the COVID-19 disease in the US.

Acknowledgements

We would like to thank our mentor Tom Ouellette for his guidance and support throughout the writing process. We are grateful for the opportunity that STEM Fellowship has granted us through the 2020 Big Data Challenge to grow our skills in data analytics and contribute to the research community.

References

[1] D Cucinotta and M Vanelli. Who declares covid-19 a pandemic. *Europe PMC*, 91(1):157–160, Mar 2020.

[2] Morgan McFall-Johnsen. An average coronavirus patient infects at least 2 others. to end the pandemic, that crucial metric needs to drop below 1 - here's how we get there. *Business Insider*, Apr 2020.

[3] Michael Greenstone and Vishan Nigam. Does social distancing matter? *Becker Friedman Institute*, Mar 2020.

[4] Joseph A Lewnard and Nathan C Lo. Scientific and ethical basis for social-distancing interventions against covid-19. *The Lancet Infectious Diseases*, 20(6):631–633, 2020.

[5] Elena Cresci. 12 ways twitter changed our lives. *The Guardian*, Mar 2016.

[6] Adam Gabbatt. Us anti-lockdown rallies could cause surge in covid-19 cases, experts warn. *The Guardian*, Apr 2020.

[7] CSSEGISandData. Cssegisanddata/covid-19. *GitHub*.

[8] Jessie Wu. jessie-wu22/unbdc-2020. *GitHub*.

[9] Kerchner, Daniel, Wrubel, and Laura. Coronavirus tweet ids. *Harvard Dataverse*, May 2020.

[10] Meg Anderson. U.s. sees exponential growth in coronavirus death toll. *NPR*, Mar 2020.