COVID-19: Comparing Twitter Volume and Sentiment to COVID cases by Country

Alex Gu, Andrew Kim, Benjamin Shaman, Chris Solé

Dartmouth College

This manuscript was compiled on November 21, 2023

Our group examines the spread of COVID-19 through the number of Tweets posted in each country and the prevailing sentiment within the Tweets. We initially hypothesized that as a country's number of active cases increases, the number of COVID-19-related Tweets will increase proportionally. Additionally, we wanted to examine the relationship between the mean sentiment of the Tweets and COVID-19 cases. To test this, we examined two datasets: the WHO's new cases data and COVID-19 Twitter data. We performed many visualizations of the time series data to help explore the relationships. We performed Granger's causality test to determine the existence and strength of causality between variables. Lastly, we subsetted our data to specific topics to analyze public engagement with government policies. We found that the volume of COVID-19 Tweets was more likely to have a causal relationship with new cases than COVID-19 Tweet sentiment.

COVID-19 | WHO Data | Sentiment | Granger Causality | Twitter

Significance Statement

This paper engages with the public sentiment surrounding COVID-19 through the analysis of Twitter data. Understanding public discourse on global crisis events such as this pandemic provides valuable insights into the collective consciousness, which in turn influences the dissemination of information as well as public response. As such, we come to the conclusion that there is a causal relationship between social media variables and the viral spread of COVID-19. Using this information, we can enact more adaptive policies and initiatives that are more likely to be effective based on public opinion. Appending previous methods of forecasting with this ability could improve early warning systems and public health intervention strategies and policies.

1. Introduction

Since the outbreak of COVID-19, there have been 800 million cases across nearly every country in the world according to the World Health Organization (1). Due to the restrictive measures imposed to contain the pandemic, people have become significantly more active on social media, especially on micro-blogging platforms such as Twitter or Instagram, with a significant amount of online discourse centered around COVID-19. Therefore, this project aims to quantify this trend by modeling the spread of COVID-19 in countries analyzing both the total volume, and prevailing sentiment of Tweets. We conducted a Granger Causality F-score test on the full dataset, and researched imposed government policies to perform targeted sentiment analysis on keyword subsets in the COVID-19 twitter data. This methodology was implemented on a dataset containing 1.9 million COVID-19 related Tweets from various countries, as well as the WHO's international COVID-19 data. The paper will comprise of 6 main sections:

Related Work: Discuss similar research that examined the trend of Tweets to new COVID cases
Data: Brief overview of the data sources
Methods: Explanation of the data cleaning and specific methods used in analysis
Results: Explanation and analysis of findings
Discussion: Situate the broader context, relevance of results , and acknowledge possible inaccuracies, shortcomings, and steps taken to mitigate them
Conclusion: Drawing an overarching conclusion, pointing to potential future investigations

2. Related Work

Previous research has further reiterated the efficacy of analyzing social media posts to comprehend the situational nature of crisis events, and we believe this ability will extend to the pandemic. Studies have been conducted into similar topics of whether the viral COVID-19 spread could be modeled by various social media analytics ((2), (3), (4)). Each of these studies took into account slightly varied factors, including sentiment, location, and volume, yet used these independent variables to predictively model the spread of COVID-19. All three of these cases concluded that there was, "the presence of a relationship between latent social media variables and COVID-19 daily cases." (2). Our study will build on previous research into this area which have already identified possible causation between COVID cases and changes in social media usage. Focusing first on the top 20 COVID Tweet countries and then

Author affiliations: [1] Dartmouth College

shifting to the United States specifically, we will test if there is possible causality of Tweets, sentiment score, and new cases in certain countries around the world.

3. Data

WHO Data. We pulled our data on COVID-19 cases per country from the World Health Organization which records the new daily confirmed COVID-19 cases, deaths, and countries affected. From the WHO Data, we pulled one Excel sheet that provides information about the date reported, country code, country region, new cases, total cumulative cases, new deaths, and total cumulative deaths (1).

In our study, we focused solely on the new cases in every unique date reported. When cleaning this dataset, we decided to remove the country code and deaths columns to get an extensive view on each country's daily trend of COVID-19 cases during the height of the pandemic from January 2020 to June 2022. The final columns were 'date reported', 'country code', 'country region', 'new_cases', and 'total cumulative cases'. Also, the 'date reported' in the WHO dataset was reported in the year-month-day format. We, instead, decided to group the dates in the year-month format to display monthly trends of COVID cases in each country and create an easier merge in our methods section. The result of the cleaning totaled 10,902 unique reports of 237 country and territorial observations.

One of the limitations in the WHO Dataset is countries report their COVID cases at different times often going several days without reporting any updates in cases. Another limitation is that the number of unique countries exceeds the 195 official countries in the world. Some territories of countries separately record their cases leading to slight underestimations of reported case data. Another possible limitation is the underestimation of COVID-19 cases because countries may lack the social welfare and government funds to make accurate recordings.

COVID-19 Twitter Data. We used a 0.01% sample from the entire pool of COVID-19-related Twitter posts, comprising of 1.9 million Tweets and 72 unique characteristic columns. The Tweets covered the time period of January 2020 through June 2022 containing posts every day in various countries around the world. The original sample comprised of columns describing specific traits that distinguished one Tweet from another. For example, the Tweet text, twitter id, location, and time of publication were only a few columns within the dataset providing an opportunity to subset.

We opted to focus solely on English Tweets originating from countries that appeared in the location column. A challenge in the COVID-19 Tweet dataset was discrepancies in the location and Tweet text columns. There were several examples of figurative locations and non-word phrases throughout the Tweets causing us to filter out Tweets that were not in English, and contained locations that were not actual countries. After running multiple data cleaning strategies to find Tweets that were posted in actual countries and separating English and non-English Tweet texts, we were able to create a filtered dataset of 656,000 Tweets comprised of English text and country locations.

4. Methods

Data Cleaning

WHO Dataset. The WHO data set that we retrieved was already mostly formatted for our purposes. In "00_WHO_data_cleaning", We converted the date column into datetime format and then into year-month format using the "dt.to_period("M")" method. We then grouped the number of cases by year-month and country.

COVID-19 Twitter Dataset. By contrast, the Pickle file for the Twitter dataset was much larger and required extensive cleaning for our purposes. Our entire data cleaning process for this set can be found in "01_Twitter_data_cleaning". We began by subsetting to English language Tweets and choosing the most relevant columns: ("tweetid", "location", "text", "date", "norm_country"). We decided to only use English to simplify our sentiment analysis. The "norm_country" column was especially useful because it contained the cleaned names of Tweet countries of origin, reflecting the result of the creators of the dataset having already done some cleaning. However, we noticed a large proportion of missing values in this column and thought it worthwhile to do further cleaning of the location column to try and add more usable rows. The location column had very few missing values, but it also contained values that could not be matched to countries, even by hand, such as 'Outer space,' 'ur mom's house,' and various emojis.

In cell 5, we subsetted to Tweets where "norm_country" was missing. Our first method for matching uses the function in cell 2, "extract_country_names", and is executed in cells 6 and 7. It tokenizes the location variable, searches for geopolitical entities within the text, and then matches them to a dictionary of countries, ISO3166_countries. Our function used the NLP package for named entity recognition (NER) of the text. We used list comprehension for each of these steps, and created a new column in the original dataframe, "match_1", to reflect the new matches. This first method resulted in an additional 25,744 usable rows of data. We then once again subsetted to unmatched Tweets, and repeated this process with a second method. For the second matching method, we used a package called geonamescache with the method search_cities to find cities within the location variable. The output of that method is a dictionary containing many facts about the

city, including the country code for the country in which it is located. This was also done using list comprehension. We then applied the function abv_to_full (from the same notebook as the ISO3166_countries dictionary) to convert country codes into full names. This second method added an additional 14,896 usable rows of data. Finally, we combined all of the original values from "norm_country" with the new ones from each of the matching methods to create the column "location_clean".

Lastly, we subsetted to rows where "location_clean" exists, and the columns "tweetid", "text", "date", and "location_clean". We exported this dataframe as a CSV, which we each stored locally, as it was more than 100mb and could not be added to GitHub. One important factor to note is that this code takes a very long time to run, so while it was possible with our sample of 1.9 million rows, it would be unreasonably slow if we scaled by another factor of ten.

Sentiment Analysis. For our sentiment analysis, we were interested in the mean compound sentiment expressed by each country over time. Our goal in "02_Perform_sentiment_analysis", was to add the "compound" column to the "location_cleaned" dataset for the sentiment analysis scores of each Tweet. Before performing our sentiment analysis we did some substantial preprocessing to improve our accuracy. First, we converted all of the text to lowercase. Then, we used the remove_users and remove_links functions to remove all user mentions and links to external websites. These functions use regular expressions and the re.sub method. Finally, in the content_preprocess function, we tokenized the Tweets and removed English stopwords, all numbers and symbols, and words with fewer than three characters. Using list comprehension, these functions give us the "processed_tweets" column. To extract the compound sentiment scores, we used Python's SentimentIntensityAnalyzer and our function, "ranking_tweets", which helped us apply the analysis to a time series. Lastly, we concatenated this result with our original dataframe of all the Tweets, and saved it as a CSV. This file is also over 100mb and has to be stored locally.

Full Dataset Comparisons and Analysis. To start exploring the full dataset, it was helpful to know the sample sizes of Tweets that we were dealing with for each country. Therefore, at the top of "03_Full_data_comparisons", after importing the Twitter and WHO datasets, we created Fig. 1 to show the proportion of Tweets among the top locations.

Next, for our analysis of Tweet volume and sentiment versus new COVID-19 cases, we had to first convert our data into time series. We decided to use months as our unit. For the Twitter set we used the method "dt.to_period("M")" to convert date to year and month format. We then grouped by "location_clean" and year-month and aggregated the count of "tweetid". This gave the number of Tweets from each country each month (4366 rows). Then, for both datasets, we created the column "ym_str" to be used later for merging, as we cannot merge dataframes on period objects.

We then used the function graph_c_v_tweets to visualize the data and create the base for the graphs seen in Figs. 2 through 5. The function takes five arguments: df1, df2, country, stat1, and stat2. df1 and df2 are the year-month formatted versions of the Twitter and WHO data respectively, and country takes a country name as a string. stat1 and stat2 are the two variables that the graph will compare. For our Tweet volume analysis that is "tweetid" and "New_cases", but the function is also generalizable to compare any two variables. For our Tweet sentiment analysis, we compare "tweetid" and "compound". The function uses a left merge to combine the two dataframes on the "ym_str" column. It then normalizes each of the inputted statistics so that the graph makes sense visually. Finally, it uses the matplotlib.pyplot package to create a graph object, two lines, and some basic labels, which can be visually customized outside of the function.

We used the function graph_c_v_tweets, described above, to graph Tweet volume against new COVID-19 cases as shown in Figs. 2, 4a, and 5a. We graphed mean compound sentiment against new COVID-19 cases as shown in Figs. 3, 4b, and 5b (see Figures below).

Granger Causality. For our project's purposes, we could not perform a simple correlation between new COVID-19 cases and Tweet volume or sentiment because they are not independent variables. The Granger Causality test is a statistical hypothesis test used to determine whether one time series can predict another, or whether there is likely to be a causal relationship between the two (5). Finding that there is Granger Causality between two time series does not imply true causation, but it suggests predictive causality based on the observed historical patterns in the data. The test involves comparing the predictive accuracy of a model with lagged values of both variables against a model with lagged values of only one variable. If including lagged values of one variable significantly improves the predictive accuracy of the model, it suggests that the first variable Granger causes the second variable.

We used Granger Causality to quantify the relationship between Tweet volume, Tweet sentiment, and new cases in different countries. To perform the Granger Causality test, we used the Grangercausalitytests module from statsmodels.tsa.stattools. The Granger Causality test also requires that the time series being compared are stationary, so we used the Augmented Dickey-Fuller test to check for this and wrote functions to take differentials where necessary. In contrast to our other analyses, in which we divided the data by month, causation analysis requires shorter time periods. For example, we cannot expect case numbers from 1-3 months ago to affect present-day social media activity. We attempted to perform Granger Causality using weekly data, but this too did not yield enough significant results. Therefore, we performed the Granger Causality test using a 7-day moving average of the daily data. As a result, it made sense to set the maximum number of lags equal to 7. We also only performed the test on the top 20 countries with the most Tweets, because countries further down the list did not have enough data.

The functions we used and our executions of the tests are found in "04_Granger_causality". First, the analysis required reshaping the data by day instead of by month. We once again used the ".to_period" method to achieve this and then grouped by day and country for both the Twitter and WHO data. We added the "d_str" to both dataframes for merging purposes. With the data now in the right format, we used the functions "prep_stationary", "prep_df", and "make_all_stationary" to execute the analysis. The "make_all_stationary" function uses the functions "check_stationarity" and "make_stationary" to take all columns in a dataframe (assumed to be time series) and make them stationary if they are not already. The "prep_df" function takes two dataframes (twitter and who) and for each one creates a 7-day moving average for the relevant variables. It then merges them on "d_str", puts "d_str" in the index, and drops all rows with missing values. Lastly, "prep_stationary" combines these two functions and normalizes the resulting columns.

We execute the tests in cell 13 of "04_Granger_causality", looping over the list of countries with the most Tweets. The output of the "grangercausalitytests" function is a dictionary of test results, which we first transform into a matrix of p-values. An example of this is shown in Table 1. Each cell can be read as "variable x Granger causes variable y." The p-values across the diagonal are all 1, as comparing two identical predictions will result in no improvement. In this example, only "New_cases" on "tweetid" has a p-value low enough to reject the null hypothesis that there is no Granger causation.

	tweetid_x	New_cases_x	compound_x
tweetid_y	1	0.0151	0.1825
New_cases_y	0.207	1	0.6319
compound_y	0.9679	0.7856	1

Table 1. Example of the Granger causality test output

From these matrices, we extract information about whether the p-values are significant, and if so, the associated F-score, or strength of the predictive value of the x-variable on the y-variable. These results are formatted in tables 3 and 4 for the relationships between cases and Tweet volume and cases and Tweet sentiment, respectively.

Finding Trends by Category: Subsetting with Regex. After analyzing the full data set we were curious to see if some trends were more pronounced in Tweets about certain topics. Thus we performed some subsetting for a more targeted analysis. Using regular expressions (regex) to search for keywords, we subsetted to Tweets that referenced the topics: Masking, Vaccination, and Social Distancing. Table 2 lists the different key words we searched for within our regex equations to create subsetted datasets we could use to uncover further possible correlations between Tweet volume, Tweet sentiment, and new cases.

Subjects	Keywords
Masking	mask, masking, face covering, kn-95
Vaccines	vax, vaccine, pfizer, moderna, booster
Social Distancing	social distance, six feet, distance, distancing

Table 2. Key Subjects and Words focused on in the Regex Patterns

For each of these subsetted categories, we compared Tweet volume with new monthly COVID cases in the United States due to the majority of our Tweets originating from the country. We also conducted sentiment analysis on the cleaned Tweet text data and compared the compound sentiment with new cases in the United States. We graphed these relationships and analyzed the trends around specific events, such as Tweets from the CDC and new COVID-19 regulations or mandates to see how government policies impacted overall public sentiment and social media activity.

The code for subsetting to these categories is found in the "05_subsetting_regex" notebook. We load the Twitter dataset, remove instances where "processed_tweets" is missing, and use regular expressions to create three binary indicators, one for whether a Tweet falls into each of the three topics. The vaccine topic contains 75,202 Tweets, 11.48% of the overall dataset, with 43,073 Tweets, 57.3%, from the US. The masking topic contains 43,988 Tweets, 6.72% of the overall dataset, with 29,040 Tweets, 66.0%, from the US. The social distancing topic contains 14,130 Tweets, 2.16% of the overall dataset, with 7,564 Tweets, 53.5%, from the US.

We then loaded the WHO data, and prepped each of the subsets for graphing using the "prep_subset" function, which takes a dataframe, groups by location and month, and creates "ym_str" for merging. Finally, we used the "graph_c_v_tweets" function from the "03_Full_data_comparisons" notebook to create graphs of data from the United States. These are shown for

masking and vaccines in figures 4 and 5.

We have different figures for the social distancing subset because we found an interesting trend in the cross-national data. We found it more useful to create bar graphs to demonstrate those points. We used the count of "location_clean" in the data subset to create figure 6 (a), and we used the same monthly sentiment data as calculated above for figure 6 (b).

5. Results

Full Dataset Analysis. In our entire 656,553 Tweet dataset, we discovered that the majority of English-language Tweets originated in the United States. Fig. 1 graphs the top 10 countries with the highest number of Tweets during the time period. We initially thought that the majority of Tweets coming from the United States might limit our comparisons between countries. However, we determined that the top twenty countries had enough data available to perform analysis, with the cutoff being Brazil at 1,956 Tweets.

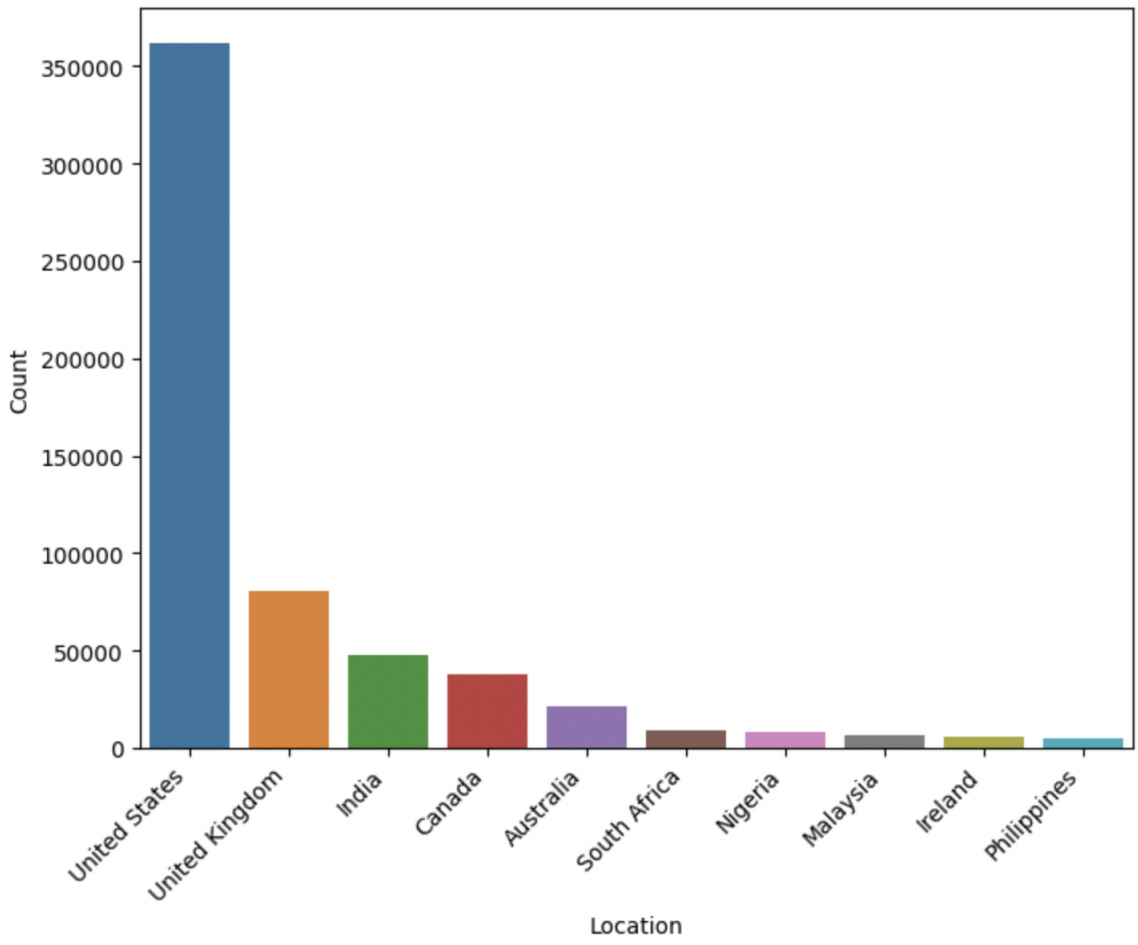
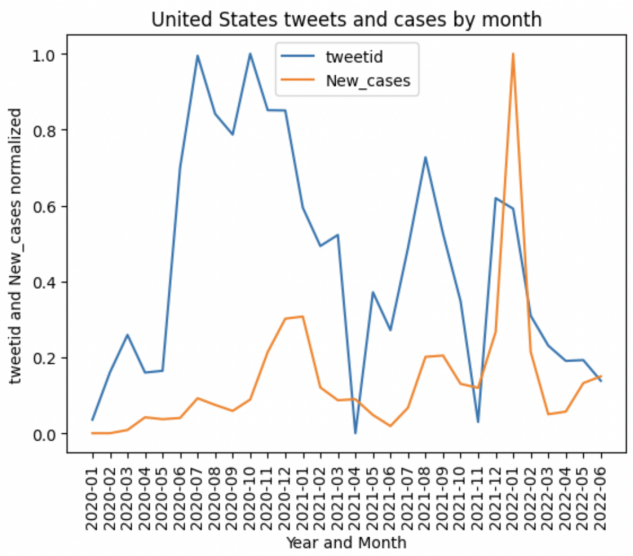


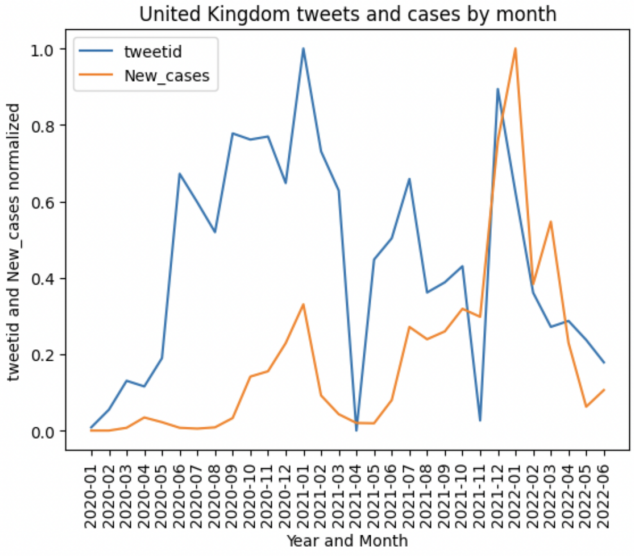
Fig. 1. Bar graph of the count of Tweets from each of the top 10 English-language countries in the full dataset

Fig. 2 are line graphs for both the United States and the United Kingdom comparing the normalized volume of Tweets and new cases by month throughout the pandemic. They indicate that in both countries Tweet volume peaked toward the beginning of the pandemic, while cases peaked toward the end. Looking more closely at the spikes in the data, we see that in both countries it looks like cases and Tweets have local maximums and minimums at around the same time. This trend was also observed in the same graphs for other countries.

Fig. 3 shows line graphs for both the United States and the United Kingdom comparing the normalized compound sentiment of Tweets and new cases by month throughout the pandemic. The overall data for these two countries seems to show little correlation between compound sentiment and the number of new cases. Sentiment will become more relevant in our analysis of specific subsets of the data.

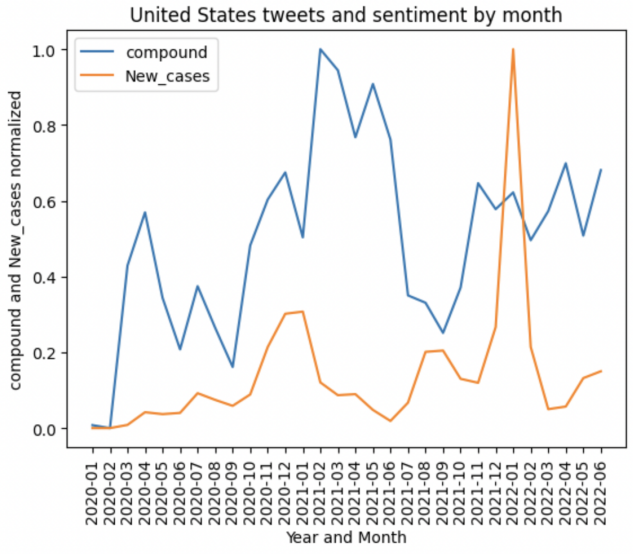


(a) Unites States Tweet volume and new COVID-19 cases by month

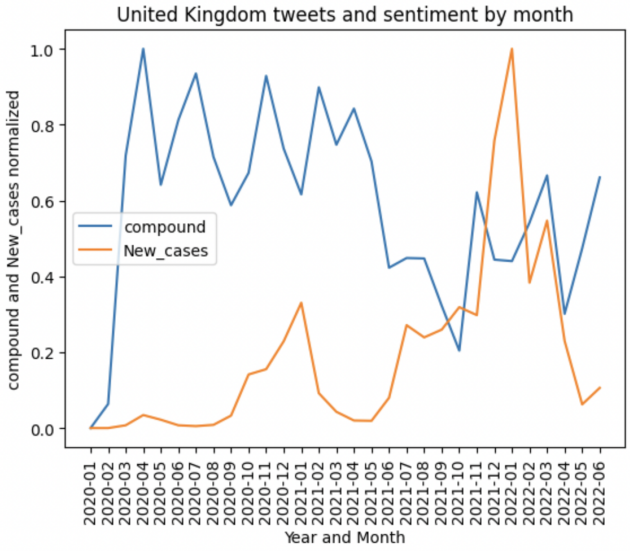


(b) Unites Kingdom Tweet volume and new COVID-19 cases by month

Fig. 2. Comparison between the US and the UK Tweets and cases



(a) Unites States Tweet compound sentiment and new COVID-19 cases by month



(b) Unites Kingdom Tweet compound sentiment and new COVID-19 cases by month

Fig. 3. Comparison between the US and the UK Tweet sentiment and cases

Granger Causality. Table 2 lists the countries in order of the overall number of Tweets originating from there. The next two columns give booleans for whether the p-value for the number of cases causing Tweet volume and vice versa is significant ($p < 0.05$) in each country. If there is a significant prediction value, the associated f-score, a statistical measure used to evaluate the significance of the Granger Causality relationship, is printed. Otherwise, *nan* is printed. For the countries where both Tweet volume and new cases are predictive of the other, the last column shows the ratio of their significance, with a value greater than one indicating that new cases have a greater effect on Tweet volume than vice versa.

The table results illustrates that cases are a better predictor of Tweet volume in India, Canada, Australia, Pakistan, Kenya, and Indonesia, while Tweet volume is a better predictor of cases in the United States, South Africa, Malaysia, the Philippines, New Zealand, Japan, and Brazil. There was no Granger causation exhibited between these time series in the United Kingdom, Nigeria, Ireland, Germany, France, Spain, and the Netherlands.

Table 3 displays the same data with the regression performed on Tweet compound sentiment versus new cases. As expected,

there are many fewer cases where Tweet compound sentiment and new cases are good predictors of each other. In a few countries, (Canada, Nigeria, the Philippines, Pakistan, and Indonesia) compound sentiment is a good predictor of new cases.

	country	p˙case˙predicts	p˙vol˙predicts	f˙case	f˙vol	case˙vol˙ratio
0	United States	True	True	8.8653	16.3483	0.5422765669824997
1	United Kingdom	False	False	nan	nan	
2	India	True	False	7.7335	nan	
3	Canada	True	True	4.2836	2.8445	1.5059237124274916
4	Australia	True	True	4.8251	3.1745	1.5199558985667034
5	South Africa	True	True	13.3045	21.5776	0.6165884991843393
6	Nigeria	False	False	nan	nan	
7	Malaysia	False	True	nan	46.2846	
8	Ireland	False	False	nan	nan	
9	Philippines	False	True	nan	16.9323	
10	Pakistan	True	True	11.4722	6.6078	1.7361602954084567
11	Germany	False	False	nan	nan	
12	Kenya	True	False	2.2462	nan	
13	France	False	False	nan	nan	
14	New Zealand	False	True	nan	41.8599	
15	Spain	False	False	nan	nan	
16	Japan	False	True	nan	12.582	
17	Netherlands	False	False	nan	nan	
18	Indonesia	True	False	4.9666	nan	
19	Brazil	False	True	nan	7.2209	

Table 3. Full dataset Granger Causality for Tweet volume vs. cases, p-value significance and f-scores

Sentiment Key Word Analysis

The next section will focus primarily on the United States as over half of our Tweets within our dataset originate from the country. We will be analyzing twitter patterns based on the regex subset of key words and subjects from Table 2. These subset tests help to confirm the results of the Granger Causality Test as the volume of Tweets had a higher chance of causation in new monthly cases compared to the overall sentiment score which was determined to have little effect.

Masking Tweets. Filtering for tweets where the masking indicator was True, we counted the number of tweets in our 'location_clean' dataset that mentioned masking by using the 'value_count' function, which calculated what proportion of mask-related tweets were posted in each country. Our analysis revealed that 43,988 Tweets contained masking-related keywords, constituting 6.72% dataset, and most of these masking tweets were from the United States, with 29,040 Tweets, or 66.0%, tweeted in the US.

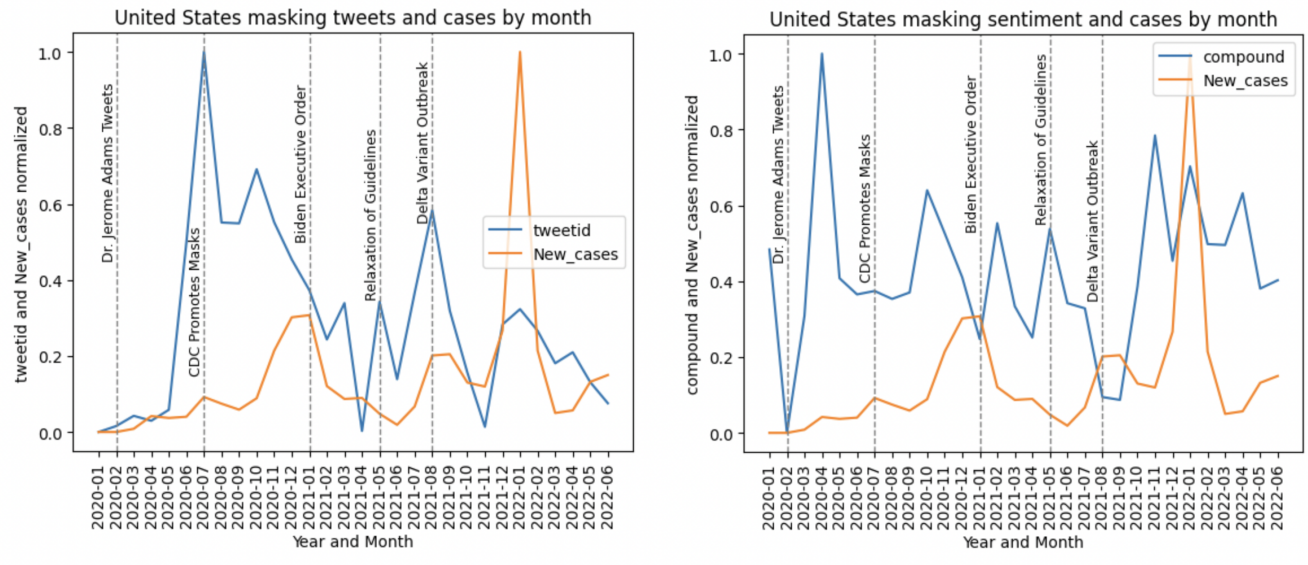
Fig. 4 (a) shows the line graphs for masking Tweets in the United States, comparing the normalized volume of masking Tweets and new cases by month throughout the pandemic. Similar to the overall data set, masking Tweet volume peaked toward the beginning of the pandemic, while the number of new cases peaked toward the end of the months in our data set. This pattern follows the trend found in the full data set.

Fig. 4 (b) shows line graphs for masking Tweets in the United States, comparing the normalized compound sentiment of masking Tweets and new cases by month throughout the pandemic. The masking subset for the United States does not show positive association between compound sentiment and number of new cases. However, sentiment appears to accurately reflect government policies and CDC guidelines. For example, Dr Jerome Adams who was the US Surgeon General made Tweets claiming that masking would not be effective at stopping the spread of the virus in February 2020. These Tweets were later deleted, but our graph shows a sharp dip in compound sentiment in February 2020. Compound sentiment rebounded and gradually increased as CDC studies found masking effective for reducing the spread of the coronavirus. After a rise in new cases in late 2020, the Biden administration implemented executive orders mandating masking. In the next few months, the compound sentiment rose, then new cases declined, and then the compound sentiment reversed, declining with the number of new cases falling. The compound sentiment of masking Tweets continued declining through delta variant outbreak in August 2021, reaching its local minimum in November 2021. Masking Tweet sentiment sharply rebounded in January 2022 when the

	country	p˙case˙predicts	p˙sent˙predicts	f˙case	f˙sent	case˙sent˙ratio
0	United States	False	False	nan	nan	
1	United Kingdom	False	False	nan	nan	
2	India	False	False	nan	nan	
3	Canada	False	True	nan	24.8177	
4	Australia	False	False	nan	nan	
5	South Africa	False	False	nan	nan	
6	Nigeria	False	True	nan	40.9871	
7	Malaysia	False	False	nan	nan	
8	Ireland	False	False	nan	nan	
9	Philippines	False	True	nan	8.0256	
10	Pakistan	False	True	nan	12.6753	
11	Germany	False	False	nan	nan	
12	Kenya	False	False	nan	nan	
13	France	False	False	nan	nan	
14	New Zealand	False	False	nan	nan	
15	Spain	False	False	nan	nan	
16	Japan	False	False	nan	nan	
17	Netherlands	False	False	nan	nan	
18	Indonesia	True	True	2.7243	9.8774	0.27581144835685506
19	Brazil	False	True	nan	2.358	

Table 4. Full dataset Granger Causality for Tweet compound sentiment vs. cases, p-value significance and f-scores

number of new cases spiked to its absolute maximum. We speculate that the fall in sentiment towards masking in United States Tweets reflected a decrease in following CDC masking guidelines—which are negatively associated with contagion when properly followed (6)—thus leading to the spike in cases during the delta variant outbreak (May 2021 through June 2022 in Fig 4 (b)). We believe we can accurately interpret the sentiment analysis of masking Tweets on number of new cases to reliably reflect Americans' opinions and perhaps compliance with masking.



(a) Masking Tweet volume and new COVID-19 cases by month

(b) Masking Tweet sentiment and new COVID-19 cases by month

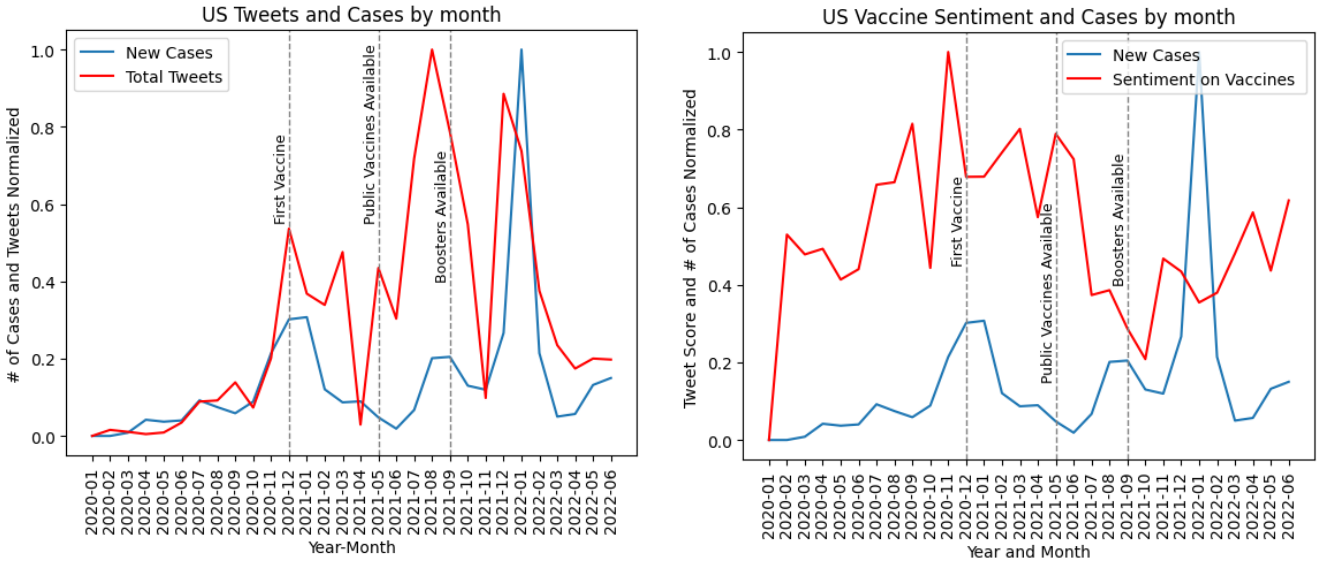
Fig. 4. Comparison between the US "Masking" Tweet volume and sentiments vs cases by month

Vaccine Tweets. For Tweets where the 'vaccination' indicator reported True, we conducted a count using `value_counts` to determine the proportion of vaccine-related Tweets within the cleaned COVID-19 dataset. Our analysis revealed that 75,202 Tweets, 11.48% of the filtered COVID-19 dataset, mentioned vaccine-related keywords, with 57.3% of these Tweets originating from the United States. Due to the substantial volume of Tweets, our group performed decided to conduct sentiment analysis on this subset of Tweets in the United States during the COVID-19 pandemic, spanning from January 2020 to June 2022. We compared the monthly trends of COVID Tweets and new cases by normalizing both columns on a scale of 0 to 1, labeling them 'US_t_norm and 'US_c_norm', respectively. After performing the previous Granger Causality tests and analyzing the subset of masking Tweets in the United States, we wanted to see if the pattern of Tweets Granger causing cases continued.

In Fig. 5 (a), the red line notes that the number of total vaccine Tweets peaked when the government announced the eligibility of the vaccine and booster shots for the elderly and the general population. This demonstrates that United States Twitter users were desperate during spikes of the virus causing an increase in the number of Tweets discussing the vaccine. All of the Tweet volume spikes in December 2020, May 2021, and August 2021 also exhibited sudden rises in COVID-19 cases. The blue line of Fig. 5 (a) maps the rise and fall of COVID-19 cases in the United States. The similarity trend in both lines confirm the Granger Causality test that there is possible causation between the overall Tweet volume about vaccines and the number of new monthly cases in the United States.

While the monthly Tweet volume and new cases trend demonstrate signs of similarities in Fig. 5 (a), the same cannot be said when comparing the monthly sentiment score of vaccination Tweets to the monthly count of new cases. The red line of Fig. 5 (b) maps the monthly COVID-19 vaccine Twitter sentiment by the number of new United States COVID cases per month. The monthly sentiment trend is fascinating as the public Twitter opinion of vaccines reached its highest positive ranking in December of 2020 when the government first issued widespread doses of the vaccine to the elderly population. At the time, United States citizens might have been in belief that vaccines could contain the spread of COVID cases hence the optimistic uptrend in the beginning of the pandemic. However once the vaccines became available to the general population, the monthly mean compound sentiment score of vaccine Tweets declined. This may be the reason why in second half of the pandemic, people started to resent getting more than one shot to prevent getting COVID questioning if the original vaccine was effective. While the blue line remains the same as the previous figure, the overall sentiment scores compared to new cases is nearly the opposite suggesting little causation between both trends.

Overall, the sentiment ranking of vaccine Tweets was positive at the start of the pandemic but declined over time. The comparison of the number of Tweets and monthly reported cases indicated potential causality as the red and blue lines matched similar peaks and valleys in Fig. 5 (a). In this specific subset of COVID-19 Tweets, there are hints of a causal relationship between the total number vaccine Tweets and number of monthly COVID-19 cases in the United States but not in the average compound Tweet sentiment and number of new cases.



(a) Vaccine Tweet volume and new COVID-19 cases by month

(b) Vaccine Tweet sentiment and new COVID-19 cases by month

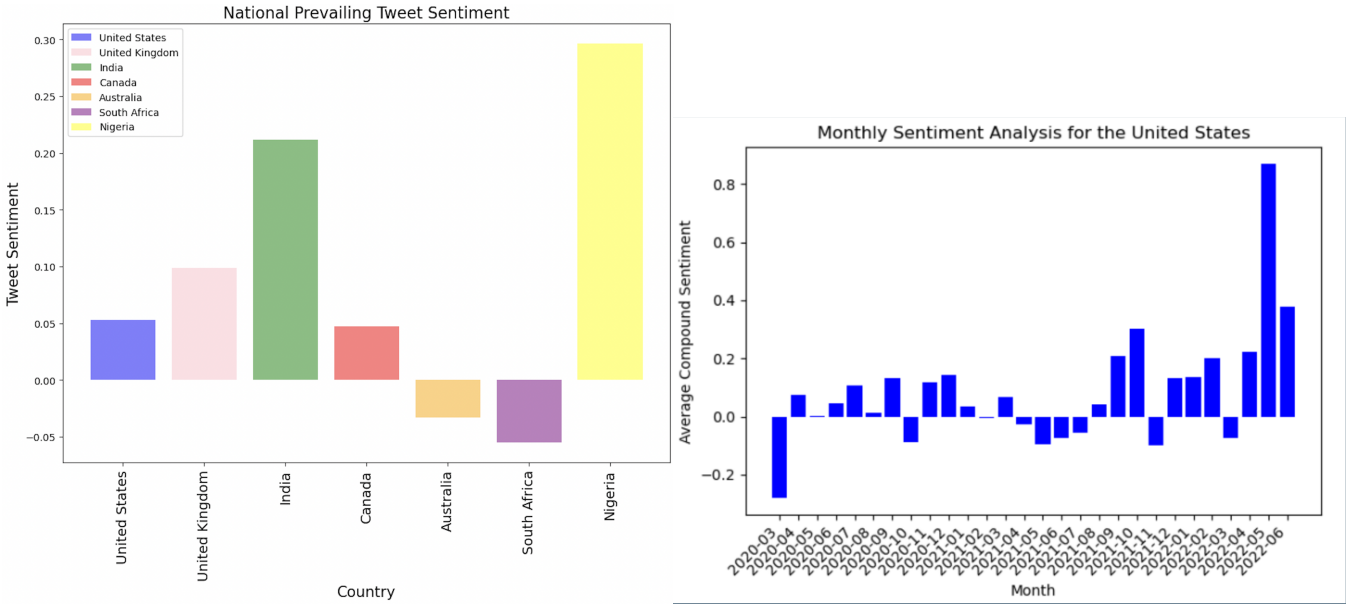
Fig. 5. Comparison between the US "Vaccine" Tweet volume and sentiments vs cases by month

Social Distancing. Similar to the other targeted sentiment analysis, for Tweets where the 'social-distance' indicator reported True, we used value_counts to conduct an initial overview of the proportion of Tweets within the dataset which discussed social

distancing. The analysis revealed that 14,130 Tweets, or 2.16%, were related to social distancing in some context, with the United States accounting for 53.5% of the Tweets, and 6 other countries accounting for another 36%. As such, our group performed sentiment analysis to obtain the compound sentiment for these top 7 countries throughout the pandemic (March 2020 - June 2022), as well as analyzing the monthly sentiment in the United States.

Fig. 6 (a) denotes the prevailing Tweet sentiment in regards to social distancing by each of the top countries represented in the dataset. This data is important to analyze as the prevailing sentiment in each nation evidences public opinion on government policies and general adherence, making it a strong predictor for cases. Furthermore, the sentiment seems to be correlated with the degree of strictness to which social distancing was enforced. Take Australia for example, with a compound sentiment of -0.03, a figure corroborated by the Australian Government Department of Health's strict policies, with the social distancing fully implemented in March 21 2020, and not lifted until February of 2022 (7).

Fig. 6 (b) represents the prevailing sentiment regarding social distancing in America, when Tweets were grouped by month. This graph shows several interesting trends, firstly, we can see the initial outcry against social distancing in March of 2020, where the compound sentiment reaches -0.24. The graph fluctuates significantly throughout the pandemic, but there is a visible uptick in sentiment positivity from September 2021 to November 2021. This coincides with the time period when the booster shot's availability became widespread, and thus may have caused more positive discourse on social distancing. Finally, from April 2022 to the end of our data we can see a huge spike in positive sentiment regarding social distancing. At this time, social distancing requirements began easing around April or May, and were lifted in August, closely following the Tweet's sentiment (8).



(a) Social Distancing Average Tweet Sentiment by Country

(b) Social Distancing Monthly Tweet Sentiment in the United States

Fig. 6. Comparison between the US "Vaccine" Tweet volume and sentiments vs cases by month

6. Discussion

After performing Granger Causality on the top-20 COVID-19 Tweet countries, and sentiment analysis of specific COVID-19 key words on the United States, we found that there are instances when the total volume of Tweets correlates to the number of new COVID cases over a given period of time. While the number of Tweets and new cases may display a common pattern, the same cannot be said for overall sentiment score on the number of new cases as the majority of the countries in our test were found to have no relationship between either variable.

We believe there were two primary reasons for the patterns in our findings. In the first case of COVID-19 Tweet volume affecting the number of new cases, people tend to talk more about topics or issues if there are a rise in the number of people affected. Twitter automatically groups large sums of Tweets about the same topic under one subject in the "Trending" tab typically containing global events that spark the interest of people within a country and around the world. When the sudden spike of COVID-19 began in the beginning of 2020, more Twitter users took to the site to discuss various events of the pandemic. From lockdowns, closing of borders, potential vaccine timelines, and implementation of distancing became widespread conversations leading to an increase in the total number of COVID-19 Tweets. This social media trend could explain why a higher majority

of countries demonstrated potential causal relationships between COVID-19 Tweet volume data and the number of new cases.

The second comparison of overall COVID-19 sentimental scores having little effect on the number of new cases can be explained through limitations. The first limitation of our study had to do with the personalized nature of tweeting. Sentiment analysis is a good testing program that can accurately rank how positive or negative a group of words can be. However, the majority of Tweets in the original dataset proved difficult to rank as not all the Tweets contained words. Numerous Tweets were not in English, contained emojis, or exhibited clear instances of sarcasm that might not necessarily align with conventional sentiment analysis. This inherent ambiguity introduces the possibility of Tweets being taken out of context based on the surrounding words. Consequently, interpreting sentiment accurately becomes more complex when considering the nuanced and varied nature of Twitter communication during the pandemic. To mitigate potential inaccuracies in the compound Twitter score, we found it necessary to filter out stopwords, URLs, retweeted accounts, and limit the preprocessed texts to words longer than three characters. While we did find some countries that indicated some signs of potential causal relationships of sentiment to new cases, others demonstrated little correlation as there may have been Tweets in different languages from countries originating in the Granger Causality test but had to be filtered out since they were not in English.

Further studies could expand beyond our scope of English speaking countries and the relationship between COVID-19 twitter data and the number of global reported COVID cases. Some future projects could be to analyze particular continents (i.e. Europe, Asia, Africa) or by language (French, German, Spanish, Chinese, Arabic) to potentially discover possible COVID sentiment patterns as one language or region experienced a polar opposite trend compared to another part of the world.

7. Conclusion

Our tests and analysis discovered that COVID-19 Tweet volume is significantly more likely to have a causal relationship with the number of COVID cases by country compared to COVID-19 Tweet sentiment score. The Granger Causality test resulted in 13 of the top-20 COVID-19 Tweet countries noticing a relationship between the number of Tweets and new cases compared to only 5 of 20 in the sentiment score and new cases test. Our keyword sentiment analysis indicated that the United States, a country that demonstrated patterns of causation in the first Granger test (see Table 2), confirmed the findings as Tweets mentioning vaccines, masking, and social distancing rose whenever new cases spiked. While it is difficult to correlate the number of people tweeting as a result in the number of COVID-19 cases, it increases the probability of a causal relationship. Overall, we conclude that the COVID-19 Tweets during the time of the COVID-19 pandemic demonstrated signs of causality in some countries around the world.

ACKNOWLEDGMENTS. Thank you Professor Chang and the TAs for a wonderful term of QSS20 and your continuous support in our research.

8. Works Cited

1. Who covid-19 dashboard. (2023).
2. AH Rabindra Lamsal, Twitter conversations predict the daily confirmed covid-19 cases. (2022).
3. MG Oguzhan Gencoglu, Causal modeling of twitter activity during covid-19. (2020).
4. D Bisanzio, Geolocated twitter social media data to describe the geographic spread of sars-cov-2. (2020).
5. S Prabhakaran, Granger causality test in python. (2022).
6. J Shaman, Mask wearing and control of sars-cov-2 transmission in the united states. (2021).
7. P Library, Covid-19 state/territory government announcements. (2021).
8. C for Disease Control, Prevention, Cdc media statement: Cdc updates covid-19 guidance for travel. (2022).