CLMS Project 2020
Analyzing Correlations between Twitter Topics and COVID-19 Cases

Nitya Sampath
University of Washington
Department of Linguistics
nsampath@uw.edu

Abstract

This project analyzes the correlations between social media topics of Twitter and COVID-19 case numbers. It uses a Twitter dataset containing tweets in English concerning the pandemic posted between January and May. In addition, data about COVID-19 cases, deaths and recoveries provided by Johns Hopkins University was used to examine the correlation pattern between the tweets and the progression of the pandemic. From these datasets, the system extracts tweets relevant to specific topics used as cues for case numbers. These tweets are further processed to analyze how their frequency and sentiment correlates with changes in case numbers of COVID-19.

1 Introduction

This project aims to look at the correlations between discussion topics on Twitter and the progression of case numbers in the COVID-19 pandemic. Specifically, it investigates three topics of conversation as potential cues for changes in case numbers.

The three topics I focus on are ”masks”, ”social distancing” and ”quarantine”. For each topic, I look at how positive or negative sentiment in Twitter data related to each of these topics correlated with case numbers in the pandemic and analyze what this might mean in relation to communication on Twitter.

To do this, I build a system that takes in Twitter data and outputs data about the number of tweets per day and the sentiment of these tweets over a period during the pandemic from late January to late May 2020.

In this paper, I will discuss my work on a system to analyze correlations in Twitter topics and COVID case data. First, I will discuss the datasets I used and the approach I took to build the system. Then, I will present the results outputted by the system, and finally, I discuss the conclusions that can be drawn from this work.

2 Datasets

2.1 Twitter Data

This system used a dataset of Twitter data compiled by Panacea Lab (Banda et al., 2020). This dataset consisted of Tweet ID’s for tweets related to COVID-19. The full dataset contains more than 600 million tweets and retweets in three languages: English, Spanish and French. The number of unique tweets (no retweets) in the dataset is more than 100 million. The dataset begins on January 1 and is updated daily.

For the purposes of this project, all retweets were removed, and only English tweets were considered. In addition, this project only used the data between January 25 and May 19.

2.2 Benchmark Datasets

Johns Hopkins University dataset

The dataset on COVID-19 cases used by the system was supplied by Johns Hopkins University (Dong et al., 2020). The JHU data included information about daily case numbers, deaths and recovered patients for several countries and regions since January 21. The data is also updated daily.

This project used the cases and deaths data between January 21 and May 19, the dates corresponding with the period covered by the Twitter data. In addition, the project did not make use of the information on recovered cases provided in the JHU dataset.

Apple Mobility Data

Apple’s mobility data (Apple Inc, 2020) was also used as another benchmarking dataset. This consisted of information on daily requests for directions from Apple Maps beginning from

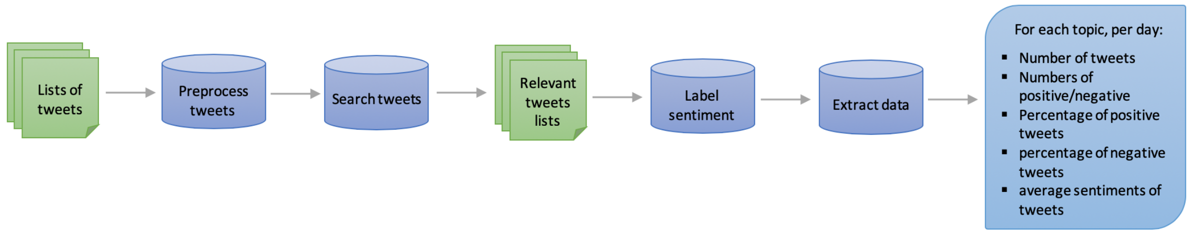

Figure 1: Diagram of system steps

January 13 and updated daily. Data is given for different regions in several different countries by date.

For the purposes of this project, the mobility data over the Twitter data time frame, January 25 to May 19, was extracted for use in the system.

Google Trends data
The last dataset used was Google Trends data, retrieved by manually searching Google Trends for each of the three topics of focus for this project. This data consists of daily samples of searches made on Google. The search data is normalized to the time and location of the search request and scaled between 0 and 100.

As with the mobility data, the Google Trends data from January 25 to May 19 was extracted for this system.

3 Approach

My approach to building this system began with preprocessing of the data. Then, the data was searched for relevant tweets and information about those relevant tweets was extracted as the output of the system. Figure 1 shows the steps of the system.

3.1 Preprocessing

The Twitter dataset included several fields of information in each tweet that swelled the size of the dataset to over 100 GB. In addition, many of those fields were unnecessary for ths purposes of this project.

In order to make the dataset more manageable, I had to preprocess the tweets to decrease the size of each tweet. Only the full text of each tweet, the tweet ID, and the date each tweet was created were saved. The rest of the fields were removed from the dataset. From these stripped tweets, the tweet text was preprocessed using regex to remove usernames and URLs, though hashtags were retained in the data. Each tweet was lowercased and tokenized using NLTK’s wordtokenizer.

3.2 Search

The next step was to search preprocessed tweets to extract all the tweets relevant to the three cues: wearing masks, social distancing and quarantine. To do this, I compiled lists of relevant keywords and hashtags to search for in the tweet tokens for each topic. These words and hashtags were compiled manually from the Twitter data, by searching through the tweets for terms that referred to each of the three topics.

Table 1 shows the keywords and hashtags used to search for tweets relevant to each topic. For the search, a tweet which contained at least one hashtag or at least two keywords for a particular topic was flagged as “relevant” for that topic. Each of these tweets were added to a “relevant” list for its topic.

Topic	**Keywords**	**Hashtags**
masks	’wear’, ’wear’, ’wearing’, ’mask’, ’masks’, ’cover’, ’face’, ’facial’, ’N95’, ’respirator’	’wearamask’, ’facemask’, ’mask’
social distancing	’distancing’, ’social’, ’distance’	’socialdistancing’
quarantine	’quarantine’, ’quarantined’, ’quarantining’, ’self-isolate’, ’self-isolating’, ’isolating’, ’isolate’	’quarantine’, ’workfromhome’, ’stayhomesavelives’, ’selfisolation’, ’stayathome’, ’lockdown’

Table 1: Keywords and hashtags for search

3.3 Sentiment Analysis

From the relevant lists, the sentiment of each tweet was calculated using Textblob (Loria, 2013). The sentiment scores consisted of two numbers: the polarity score and the subjectivity score. The polarity of each tweet was scored within a range from -1 to 1, with -1 being very negative and 1 being very positive. The subjectivity refers to whether the tweet stated a fact or gave an opinion. It was scored over a range from 0 to 1, where 0 meant very objective and 1 meant very subjective. For the purposes of this project, I chose to focus on the polarity score (negativity or positivity) of each tweet, rather than the subjectivity.

3.4 Data Extraction

After scoring the tweets, information about the relevant lists for each topic had to be extracted. The objective was to extract data about the volume of tweets posted each day and their sentiments by the date on which they were posted. The data extracted at this stage consisted of:

- total number of tweets per day,
- the count of positive tweets per day
- the count of negative tweets per day
- the sentiments of each tweet posted per day

All tweets having a sentiment score strictly greater than 0 were counted as ”positive” and all tweets having a score strictly less than 0 were counted as ”negative.

The counts were then smoothed to account for certain anomalies in the data, such as a particular date having an unusually low number of relevant tweets. This was done by setting the counts of total/positive/negative tweets for each day to the average count of tweets for that day as well as the two previous days and the two subsequent days.

From this information, the percentages of positive tweets per day were calculated, as well as the percentage negative and the average sentiment of the tweets posted per day. The percentages were calculated by taking the ratio of the positive (or negative) tweets over the total number of tweets. The average sentiment scores were calculated by taking the sum of the sentiment scores for each date and dividing it by the total number of tweets for that date.

Next, I combined this data with data on case numbers and deaths by date extracted from the JHU dataset. I calculated the number of new cases and deaths per day by subtracting the previous day's numbers with the current day's numbers for each date. These case numbers and deaths were also smoothed in the same manner as the tweet counts. From this data, I calculated the percent change in new cases by taking the ratio of the current day's new case numbers over the previous day's new case numbers.

Finally, the correlations between the new case numbers and the percentages were calculated as well. The correlation coefficients over the entire time period from January to May, as well as two subsets of the time period, from January to March and from March to May, were calculated.

4 Evaluation Methods

4.1 Manual Evaluation

To evaluate the quality of the relevant tweets returned by the system, I took a random sample of 100 tweets from the relevant list. I manually evaluated each of these sample tweets on three measures of the accuracy of the search. This process was repeated for each topic.

First, I counted the number of tweets in the samples that were actually relevant to the topic. All tweets that discussed the topic in some form were counted as ”actually relevant”.

Second, I counted the number of tweets that were labeled with the correct sentiment, regardless of their topic. So, a tweet that expressed a positive sentiment and received a positive polarity score would be counted, even if it was not actually relevant to topic it was flagged as relevant to by the search.

Third, I counted the number of tweets for which the polarity score aligns with whether the tweet is in favor of the topic. So, for example, a tweet in the ”masks” relevant list that expressed an opinion in favor of wearing masks and received a positive polarity score was counted. In addition, a tweet in the ”masks” relevant list that expressed an opinion opposed to wearing masks and received a negative polarity score would also be counted. However, a tweet that did not express an opinion in favor of ”masks” but still scored positively would not be counted, even if it was actually relevant.

4.2 Correlations

To evaluate the correlations, I looked at how the data I extracted was correlated with the cases data over the entire time frame of the tweets, January to May. I also looked at the correlations over two distinct periods within that time period, the first from January to March and the second from March to May.

To further evaluate how well the system performed, I analyzed the correlations between the extracted data and the mobility data and Google Trends data over the same three time periods.

The mobility data was used based on the assumption that increased support for example, for quarantining, might reflect a real world practice of this action. So, we could expect that the number of tweets about, or in favor, of quarantining or social distancing would correlate somewhat negatively with mobility trends.

The Google Trends data was used as a benchmark because it could be assumed that an increased interest in a topic would be reflected similarly in Google searches and tweets. So the number of tweets about a topic should show similar trends to the number of Google searches about the same topic.

5 Results

5.1 Manual Evaluation

The results from the manual evaluation are given in Table 2. The values in the table are the counts (out of 100 tweets in each sample) of tweets for each of the evaluation measures.

Topics	**Relevant**	**Sentiment**	**In Favor**
Masks	85	77	58
Social distancing	84	67	59
Quarantine	83	65	53

Table 2: Manual evaluation results

Here, we can see that overall, the majority of the tweets returned by the system were actually relevant to the topic they were flagged for. There were, however, several tweets that ended up not actually being relevant, despite meeting the heuristic the system used to identify likely relevant tweets. Some of these tweets included several which contained a hashtag relevant to the topic, but whose main text did actually not discuss the topic in any form, which were not considered to be ”relevant” to the topic.

The sentiment analysis had somewhat lower counts in the manual evaluation, but still managed to correctly identify the sentiment of the majority of the tweets. Still, there were several tweets that were scored incorrectly. Many of these were tweets that contained sarcasm or humor of some kind.

The third measure looked at whether the tweets' polarity scored actually reflected whether or not the tweet was in favor of the topic it was flagged as relevant for. This measure had consistently lower counts for all three topics, with the highest being 59 out of 100 for social distancing. These counts are explained both by the accuracy of the sentiment scores themselves, as well as by several factors relating to how people used language to express their opinions in the tweets.

In some cases, certain tweets' sentiment did not align with their stance on the topic. For example, a tweet denouncing the actions of people not practicing social distancing, might have been scored as negative. In this case, while the polarity was correctly scored, the score did not actually reflect the fact that the tweet was ultimately in favor of social distancing.

On the other hand, there were several cases where a tweet like this was in fact incorrectly scored for polarity, but it ended up reflecting the actual opinion of the tweet toward the topic.

In other cases, tweets may have expressed a sentiment, either positive or negative, about something other than the relevant topic. For example, a tweet that expressed a positive sentiment about ways to pass time during quarantine might not necessarily express any opinion on quarantining itself. In this case, it could not be counted because its sentiment was positive but it did not express a view in favor of the topic.

Notably, a few of the tweets contained news headlines that did not actually express a sentiment or argue an opinion. Almost all of these were given a polarity score of 0, which happened to be correct for both the polarity and the tweet's stance on the topic. In this case, the tweet was counted for both measures. For social distancing in particular, tweets such as this were especially frequent.

5.2 JHU Data

Here, I present the correlation results for the JHU COVID-19 case data. For each topic, a chart showing the correlations between the cases data and the percentage of positive tweets extracted from the data is given. The charts display the new cases per day and the percentage of positive tweets for each day over the full time period from January to May to show how the correlation patterns changed over time. In addition, a table displaying the correlation coefficients of the case data and the extracted positive percentages period over the entire time periods, as well as subsets of the time period, is shown. The correlations for the number of new cases as well as the ratios of the change in new cases from the previous day are shown in the table.

Masks

Figure 2 shows the correlation patterns for the topic ”masks.” The chart shows that over the entire time period from January to May, the patterns in correlation changed considerably. From January to March, the percentage of positive tweets and the daily new cases are not very correlated, with both fluctuating a lot during this time period. However, from March to May, the correlation increases and becomes strongly positive as both the percentage of positive tweets and the daily new cases increase as time goes on. Ultimately, we see a general pattern that the number of new cases per day rises as the percentage of positive tweets rises.

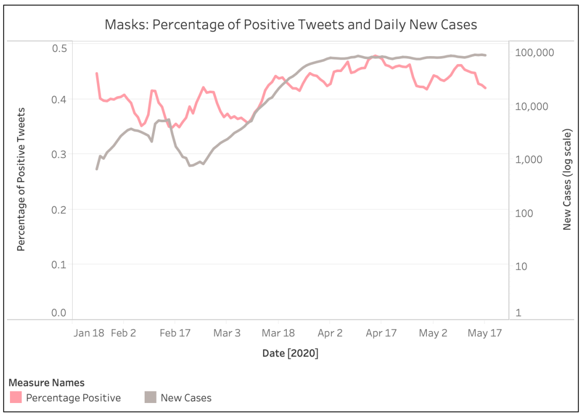

Figure 2: Chart of correlations for topic ”masks”

Table 3 shows the correlation coefficients for ”masks” over the time period. The correlation for new cases and percentage of positive tweets between January and March is rather low, while between March and May is is much higher. Overall, the correlation between the new cases and the percentage positive is quite high. This supports the patterns seen in Figure 2.

	New Cases	Ratio
Jan - May	0.81627	-0.00016
Jan - March	0.26411	0.55325
March - May	0.80975	-0.59236

Table 3: Correlations for topic ”masks”

For the ratio of new cases, we see overall a negative correlation between the ratio and the percentage of positive tweets. However, the correlation is weak over the entire time period, because it changes drastically over time. The correlation from January to March is positive, while the correlation from March to May is negative. Both correlations are similarly strong, which contributes to the weak overall correlation.

Social Distancing

Figure 3 shows the correlation patterns for the topic ”social distancing”. The chart shows that the percentage positive for social distancing was quite weakly correlated with the daily new cases over the time period. On the whole, the correlation appears to be negative, especially between January and March, when the percentage of positive tweets fluctuates considerably between its highest point and lowest point observed over the entire period. The number of new cases rises only slightly and then drops to its lowest point during this period as the percentage of positive tweets reaches a high point. Between March and May, the correlation seems to decrease as both the positive percentage and the new cases seem to become more even. Still, we have a negative correlation as the number of new cases rises over this period, while the percentage of positive tweets drops.

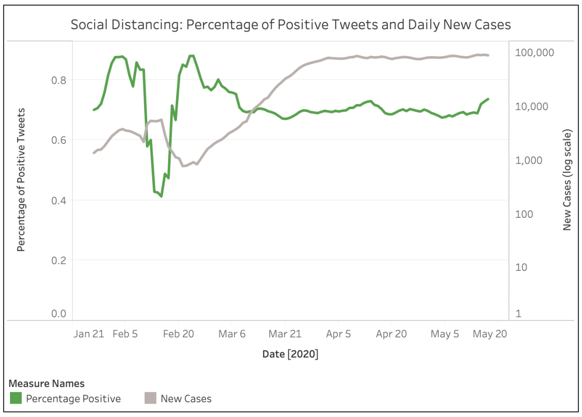

Figure 3: Chart of correlations for topic ”social distancing”

In Table 4, we have the correlation coefficients for ”social distancing”. The values largely support the patterns in the chart, with the overall correlation being quite low, rising from January to March and dropping again from March to May.

	New Cases	Ratio
Jan - May	-0.22754	0.05230
Jan - March	-0.532041	0.06166
March - May	-0.28624	0.08249

Table 4: Correlations for topic ”social distancing”

The correlations for the ratios of new cases and the percentage of positive tweets show that overall the correlation is positive, meaning that the ratio of new cases increases as the percentage of positive tweets increases. However, this correlation is very weak. It is weaker from January to March than from March to May, but still quite weak over the entire time period.

Quarantine

Figure 4 shows the correlation patterns for the topic ”quarantine”. The chart shows an overall strong, positive correlation between the percentage of positive tweets for ”quarantine” and the daily count of new cases. We see that the correlation is weaker at first, but then becomes stronger after March as both values increase together.

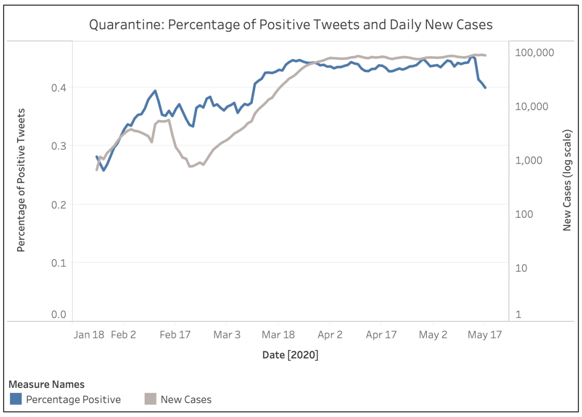

Figure 4: Chart of correlations for topic ”quarantine”

Table 5 shows the correlation coefficients for ”quarantine”. Here, as with the topic ”social distancing”, the coefficient values largely support the patterns in the chart. However, the correlations change over time. From January to March, it is lower, but still positive, while from March to May there is a highly positive correlation. The coefficient overall from January to May is highly positive, meaning that in general, the daily number of new cases increases with the percentage of positive tweets.

	New Cases	Ratio
Jan - May	0.78603	-0.12947
Jan - March	0.27423	-0.13497
March - May	0.71131	-0.34866

Table 5: Correlations for topic ”quarantine”

5.3 Mobility Data

This section presents the correlation results from the benchmarking tests with Apple’s mobility data. For each topic, the correlations between the benchmarking dataset and the percentage of positive tweets is shown in a chart and a table of the correlation values over the entire time frame and sections of the time frame is also given. In addition, I have shown the correlations between the mobility data and the daily number of cases.

Masks

Figure 5 shows the correlation patterns for the mobility data for the topic ”masks”. The chart shows that the correlation between mobility and the percentage of positive tweets starts off very low in late January. The mobility fluctuates a lot over this period. While the percentage of positive tweets also rises and falls during this time, it does not show the same level of fluctuation. There appears to be a somewhat negative correlation from January to March, as the low points of the fluctuations in mobility seems to correspond with high points in the positive percentage at various

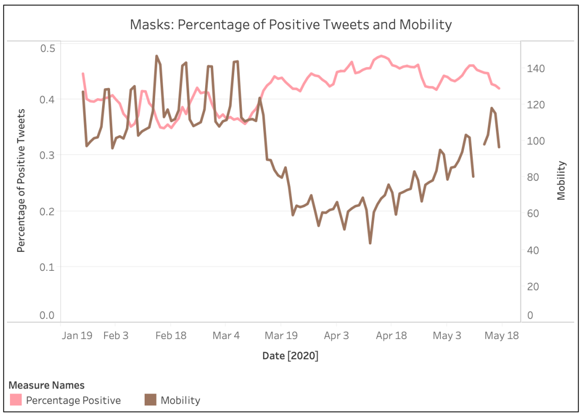

Figure 5: Chart of mobility correlations for topic ”masks”

times. Still, the correlation is generally quite weak during this period. However, around mid-March, the correlation becomes more strongly negative. The mobility sharply decreases as the percentage of positive tweets increases over the period from March to May.

Table 6 shows the correlation coefficients for the mobility data for ”masks”. The correlation between January and March is negative and relatively low, but becomes a stronger negative correlation between March and May. Overall the correlation is very negative, so as the percentage of positive tweets increases, the mobility decreases.

	Mobility Data
Jan - May	-0.74237
Jan - March	-0.21782
March - May	-0.70535

Table 6: Mobility correlations for topic ”masks”

Social Distancing

Figure 6 shows the correlation patterns for the mobility data for the topic ”social distancing”. Overall, the correlations seem to be relatively low, particularly during the first period From January to March. During this period, the correlation seems to be somewhat negative as mobility is very slightly increasing on average, while the percentage of positive tweets rises, but then drops considerably. Over the period form March to May, the correlation is also not very strong.

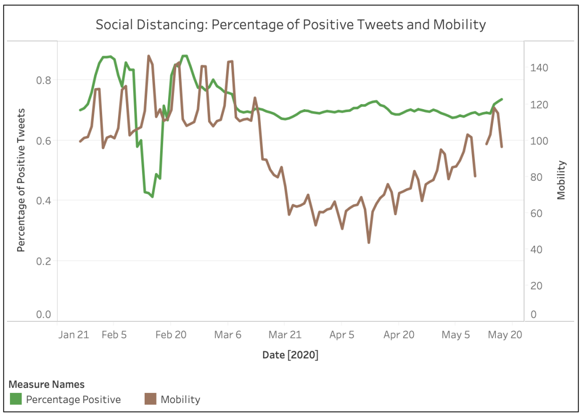

Figure 6: Chart of mobility correlations for topic ”social distancing”

Table 7 shows the correlation coefficients for the mobility data for ”social distancing”. The correlations show that initially, from January to May, there is a weak, negative correlation between mobility and the percentage of positive tweets. However, between March and May, this becomes a somewhat positive correlation between the two. Overall, however, there is a very weak, positive correlation between mobility and percentage of positive tweeets.

	Mobility Data
Jan - May	0.13742
Jan - March	-0.22583
March - May	0.40639

Table 7: Mobility correlations for topic ”social distancing”

Quarantine

Figure 7 shows the correlation patterns for the mobility data for the topic ”quarantine”. The chart shows a somewhat positive correlation between mobility and positive tweets over the period from January to March. However, afterwards, it shows a strong negative correlation between March and May, as the mobility drops sharply and the percentage of positive tweets slightly increases.

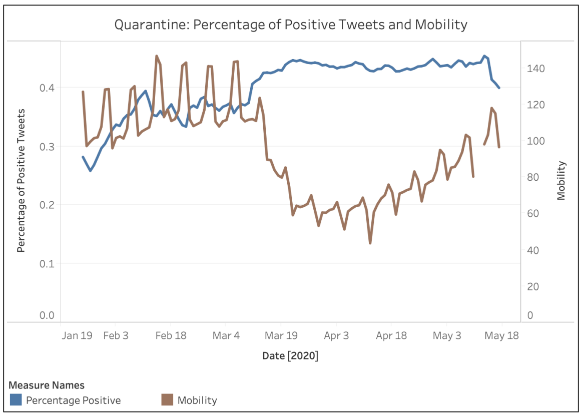

Figure 7: Chart of mobility correlations for topic ”quarantine”

Table 8 shows the correlation coefficients for the mobility data for the topic ”quarantine”. The table shows that for January through March, the

	Mobility Data
Jan - May	-0.67961
Jan - March	0.18076
March - May	-0.70910

Table 8: Mobility correlations for topic ”quarantine”

correlation is, in fact, low and positive. Then, from March to May, it becomes a strong, negative correlation, supporting the patterns seen in the chart. Overall, there is a relatively strong, negative correlation, showing that generally, the mobility decreases as the percentage of positive tweets increases.

New Cases
In addition to the three topics, in Figure 8 we have a chart showing the relationship between the mobility data and the daily new cases. Here we see a strong negative correlation between the number of daily new cases and the mobility data. The correlation coefficient between new cases and mobility over the period from January to May is -0.71892, which supports the strongly negative correlation seen in the chart. This suggests that the number of new cases rises as mobility falls and falls as mobility increases.

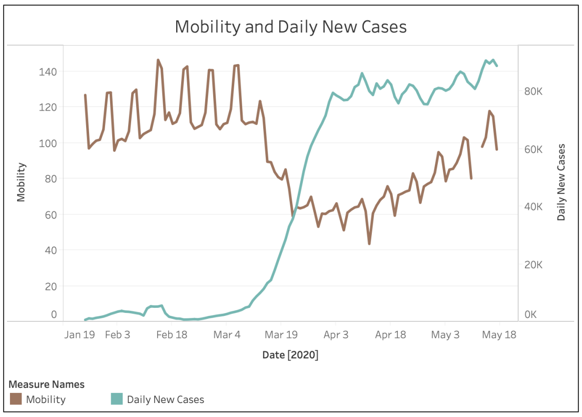

Figure 8: Correlations between mobility and new cases

5.4 Google Trends

This section presents the correlation results from Google Trends data. For each topic, the correlations between the Google Trends score and the total counts of tweets is shown.

Masks
Figure 9 shows the correlations with the Google Trends data for the topic ”masks”. The chart shows an overall positive correlation between the Google Trends score and the number of tweets over the period from January to May. The correlation is stronger early on, between January and March. Afterward, the correlation weakens as the Google searches spike but then decrease slightly over time. However, during the same time period, the number of tweets increases and spikes in late May.

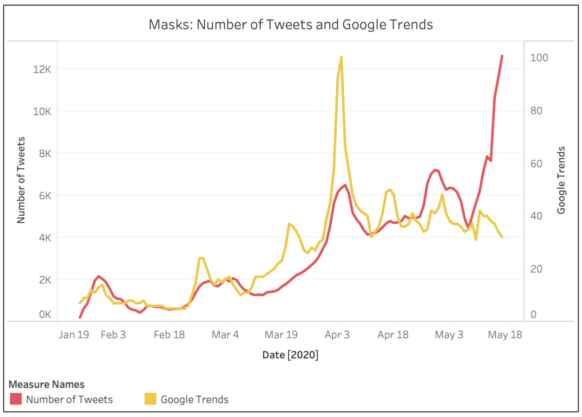

Figure 9: Chart of Google Trends correlations for topic ”masks”

Table 9 shows the correlation coefficients for the Google Trends data for the topic ”masks”. The table shows a very strongly positive correlation between January and March that later weakens between March and May, confirming the patterns shown in the chart. Overall, the correlation between the Google Trends score and the number of tweets is strongly positive, showing that the number of tweets on the topic ”masks” extracted by the system roughly mirrored the number of Google searches around the same time.

	Google Trend Score
Jan - May	0.73981
Jan - March	0.74761
March - May	0.56099

Table 9: Google Trends correlation results for topic ”masks”

Social Distancing
Figure 10 shows the correlations with the Google Trends data for the topic ”social distancing”. The chart shows overall a high correlation between the number of tweets and the Google trends score. The period from January to March appears to show a very high correlation between the number of tweets and the Google Trends score, while it would seem that later period shows numbers that are less strongly correlated.

Table 10 shows the correlation coefficients for the Google Trends data for the ”social distancing”. The coefficients actually show that the period from January to March had a lower correlation than the period from March to May. This might be because

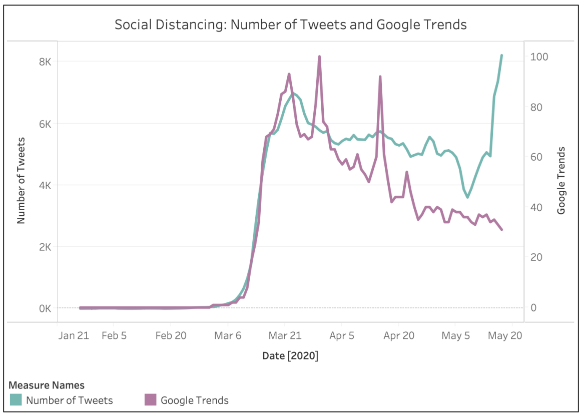

Figure 10: Chart of Google Trends correlations for topic ”social distancing”

	Google Trend Score
Jan - May	0.91101
Jan - March	0.487347
March - May	0.78524

Table 10: Google Trends correlation results for topic ”social distancing”

the Google Trends scores for the dates from late January through the first days of March were actually all 0. The number of Google searches for the topic ”social distancing” during this time period was especially low, which reflects in the scores being uniformly 0 over this period. However, the number of tweets, while also very low during the same period, also shows some change over time, with the numbers fluctuating between very low numbers for much of the time period. This issue in the data would have resulted in low the correlation coefficient, and appeared on the chart as a very high correlation.

The later period from March to May also had a high correlation coefficient, one which was not skewed by anomalies in the data. It supports the pattern seen in the chart that the number of tweets extracted by the system roughly follows the Google Trends score.

Overall, the Google Trends score is very high, which likely reflects the strange patterns in the Google Trends data during the first half of the time period. Still, in general, the number of tweets extracted by the system does closely follow the Google Trends data.

Quarantine

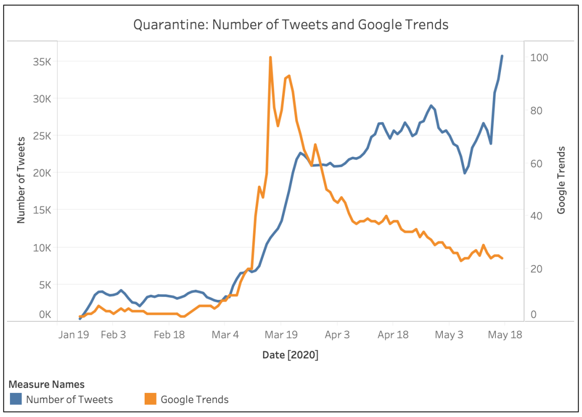

Figure 11: Chart of Google Trends correlations for topic ”quarantine”

Figure 11 shows the correlations with the Google Trends data for ”quarantine”. The chart shows that the number of tweets for ”quarantine” was much less correlated with the Google Trends score than for the other two topics. Overall, we see a positive correlation between the two, with the correlation being highest early on, until around early March. However, after that point, we see that the Google searches initially spike before dropping to low numbers, while the number of tweets generally increases over the same period. During April and May, the number of tweets and the Google Trends score are only weakly correlated.

Table 11 shows the correlation coefficients for the Google Trends data for ”quarantine”. Here, we see that the coefficient between January and March is, while relatively low, still much higher than the coefficient between March and May. Overall, the correlation is positive, and somewhat strong, but still much weaker than the correlations for the other two topics.

	Google Trend Score
Jan - May	0.56086
Jan - March	0.45032
March - May	0.08265

Table 11: Google Trends correlation results for topic ”quarantine”

6 Discussion

6.1 Analysis

The correlation patterns we see in the Twitter data might suggest several things about the relationship between trending discussion topics on Twitter and the progression of the COVID-19 pandemic.

For ”masks”, we see that the percentage of positive tweets is very strongly positively correlated with the daily number of new cases. This would suggest that as the sentiment in tweets about ”masks” becomes more positive, the number of cases increases faster. This conflicts with expected patterns, if we assume that a more positive Twitter sentiment correlates with more support for wearing masks. We would instead expect the number of new cases to decrease with increased positive sentiment for ”masks”. However, this pattern is more likely showing that support for wearing masks increases as the pandemic worsens. This would mean that the increase in positive sentiment for wearing masks would be in response to the increase in the number of cases, rather than the other way around.

For ”social distancing”, we see a much weaker correlation between percentage of positive tweets and daily new cases than for ”masks”. This suggests that Twitter support for ”social distancing” has less of an relationship to the number of cases than support for ”masks” does. On the other hand, we do see a somewhat negative correlation between positive sentiment for ”social distancing” and the number of new cases. This means that as positive sentiment for ”social distancing” increases, the number of cases decreases. This relationship, though weak, might suggest that increasing support for practicing social distancing corresponds to a decrease in new case numbers.

For ”quarantine”, we see correlation patterns similar to the ones seen for ”masks”. Overall, we have a high positive correlation between the percentage of positive tweets and the number of new cases. At first the correlation is lower, but it increases as time goes on. The positive correlation means that it is unlikely that increased positive sentiment in tweets for ”quarantining” would correspond to a change in cases. However, just as for masks, this relationship could actually indicate that Twitter support for quarantining might be a response to the worsening situation of the pandemic.

With the mobility data, we see that for ”masks” and ”quarantine”, increased positive sentiment on Twitter is, in fact, strongly correlated with a decrease in mobility. This suggests that the Twitter sentiment might be reflective of actual practice of these public health measures. On the other hand, we do not see as strong of a correlation between mobility and ”social distancing” suggesting that the Twitter discussion of social distancing might not be an accurate reflection of actual behavior. We also see a correlation pattern that suggests that new cases rise as mobility decreases, which seems to conflict with the expectation that new cases would fall as mobility decreases due to quarantining. However, this could also be explained similarly to the potentially counterintuitive correlations we see for masks. It is likely that the decrease in mobility is a response to the increasing new cases instead of a lower number of cases being a response to decreased mobility.

6.2 Issues and Weaknesses

This system does have several weaknesses. The search terms for extracting relevant tweets have currently been compiled manually to a list. This could potentially leave out tweets that are actually relevant, but would not match with the current keywords.

In addition it makes certain assumptions about the data that might not be accurate. In particular, with the sentiment analysis, the system has to assume that a positive sentiment is in favor of the topic. However, from a manual evaluation of the data, this is not necessarily true. Positive tweets might be expressing sentiments about a different topic, even if they are correctly flagged as relevant to the topics focused on in this project.

7 Conclusion

This paper presents the results of a system analyzing correlations between topics in Twitter data and the progression of cases of COVID-19.

There are many ways this work can be expanded in the future. Currently, the search is done using manually compiled keywords. As there are several issues with the accuracy of such manual keywords in searching for relevant texts, the system could be expanded to instead use keywords selected more systematically.

Future work could also focus on methods to more precisely extract the number of tweets in favor of each of the three topics, rather than simply analysed as expressing a positive sentiment, as well as an accurate method of automating the selection of keywords to search on.

8 Relation to CLMS

I felt that the CLMS coursework prepared me for this project and the work necessary to complete it. I felt that for the most part, the concepts I had to use for this project were ones I was able to understand because of prior experience in CLMS coursework.

However, I found managing the massive amounts of data this project dealt with very difficult. It was the most difficult hurdle I ran into with this project. I had not had to deal with such large amounts of data before, so figuring out how to efficiently process the data for every step in the system proved very difficult. Ultimately, the initial processing steps that pass through the entire Twitter dataset became the most time consuming part of the project to implement and run. It would be beneficial to have had coursework that prepared us to deal with large amounts of data.

References

Apple Inc. 2020. Apple covid-19 mobility trends reports.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalina, and Gerardo Chowell. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration. This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates. Release: We have standardized the name of the resource to match our preprint manuscript and to not have to update it every week.

Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20(5):533 – 534.

Steven Loria. 2013. Textblob.