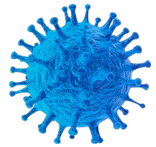
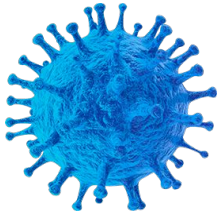
Public Opinion of the COVID-19 Pandemic

Laura Stagnaro & Ian Byrne
SIADS 591 & 592
University of Michigan School of Information

Table of Contents





All manipulated data that was used for the project can be found here under Final Project. A sample of the raw IEEE data along with the COVID time series data can be found in the Original Data folder

All scripts can be found here. Data gathering and manipulation scripts are under Data-Gathering-Scripts. The analysis and visualization notebooks can be found under final_project

A World Event Unlike Any Other

Motivation

The COVID-19 pandemic is a world event unlike any other in recent history. As of August 10, 2020 there have been 20 million cases worldwide (Pineda, 2020). According to the CDC it is "spread mainly from person to person" and can be prevented through measures which reduce contact between people, including social distancing and wearing a face mask (CDC, 2020). It is clear, based on the CDC recommendations, people's behavior impacts the spread of COVID-19. Behavior is often influenced by beliefs and opinions so understanding how people view the pandemic may give some insight into why infection rates change.

If a relationship between people's opinions and infection rates could be established this information could potentially be used to predict when surges or drops in cases may occur. It could also potentially help us better understand the trajectory of other health crises, like flu epidemics. The goal of this project is to do a preliminary exploration of the relationship between people's thoughts and feelings regarding COVID and the spread of COVID. We seek to accomplish this by looking at how tweets about COVID have changed with the number of daily new infections and global spread. This includes looking at topics discussed as well as sentiment. It does not seek to determine if trends could be predicted based on Twitter data but could provide the first step for a more in depth analysis.

Inspiration for this project came from the Medium article "Covid-19 Outbreak: Tweet Analysis on Face Masks" by Yanqing Shen.

The goal of that analysis was to look at tweets regarding face masks specifically. We took some of her analysis techniques such as looking at the most common words and sentiment and broadened our scope to look at Tweets about COVID in general as well as layered in COVID infection data to see the relationship between tweets and infection spread.

Questions:

1. What are the most common words used when tweeting about COVID and how have those trended over the course of the pandemic?
2. How has sentiment changed as trends in number of daily confirmed cases has changed?
3. Is there evidence of burnout? Has the conversation moved to apathy towards the virus?
4. How have tweets about COVID changed as it has continued to spread globally and is there a pattern which shows tweets are impacting the spread?

Johns Hopkins Global COVID-19 Data

Data Source

The Center for Systems Science and Engineering at Johns Hopkins Global COVID-19 Time Series Data consists of a CSV file with the cumulative number of confirmed cases for each country by day. The dates span from January 22, 2020 through present day. We filtered the data to only include March 19, 2020 through September 1, 2020 to keep it consistent with the data available for the Twitter dataset.

The CSV is in a wide data format so each row represents a country with a separate column for each date. While there are only 188 countries represented in the data there are actually 267 rows since several countries are broken down by state or province.

Johns Hopkins collects the data from multiple sources, including the World Health Organization, the European Center for Disease Prevention and Control and various US state departments of health. New data is added on a daily basis and missing or inaccurate data is corrected..

Johns Hopkins Global COVID-19 Time Series
Location:
https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series
Size: 263 KB
Format: CSV
Variables: Country, dates
Date Range: 3/19/2020 - 9/1/2020

COVID-19 Tweets Dataset

Data Source

The "Coronavirus (COVID-19) Tweets Dataset" consists of tweet IDs and sentiment scores for over 500 million English tweets spanning from March 19, 2020 through present day. The data is formatted in one CSV file per date and is located on the IEEE website. This project gathered 1% of all tweet IDs (4,823,702) from March 19, 2020 through September 1, 2020.

The original tweets were gathered by Rabindra Lamsal from the School of Computer and Systems Sciences at Jawaharlal Nehru University. The collected tweets contained at least one of over 90 COVID related keywords such as, "#covid", "quarantine", "lockdown" and "wear a mask". The CSVs only contained tweet IDs in order to comply with Twitter's redistribution policy which restricts the sharing of Twitter information to tweet ID or user ID. The sentiment score was calculated using TextBlob. TextBlob calculates a polarity score which has a range of -1 to 1 where -1 is very negative and 1 is very positive.

All tweet IDs had to be "hydrated" with the actual tweet content using the Twitter API . Any tweet IDs which were no longer available were not included in the final dataset. This reduced the final size to 4,217,125. There are various reasons these tweets may not have been available such as the user deleting the specific tweet or their account.

IEEE
Location: https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset
Size: 4,823,702 Tweet IDs
Format: CSV
Variables: Tweet ID, Sentiment Score
Date Range: 3/19/2020 - 9/1/2020

Twitter
Location: www.twitter.com
Size: 4,217,125
Format: JSON
Variables: Tweet ID, timestamp, text, hashtags, language
Date Range: 3/19/2020 - 9/1/2020

Gathering the Tweets
Manipulation

STEP 1: Create the 1% Sample of Tweets

one_perc_sample.py
This python script loops through a folder containing the daily CSV files of COVID-19 related tweets originally accessed on IEEEDataport. Each CSV file contained all of the tweet IDs that were identified as relating to COVID-19 in some way. As each CSV was accessed a one percent sample of each day was taken with pd.DataFrame.sample() and appended to a new dataframe that contained the previous one percent samples gathered by the program. Random state of 52 was used here. The final aggregation of samples was written to a CSV.

STEP 2: Gather the Tweets

twitterAPIScript.py
Splits all the tweet IDs into groups of 100 IDs and then uses the Tweepy library to gather 100 tweets in one API call. It then parses the returned JSON to get the tweet ID, timestamp, text, content, hashtags, language, and sentiment score. Any tweets which were no longer available were not gathered by the API. This left us with approximately 87% of all tweet IDs we initially sampled from the IEEE data. The script then outputs the Twitter information into a CSV.

A challenge we encountered at this stage was after initially hydrating all of the tweet IDs, hundreds of thousands of them were formatted incorrectly in the CSV file. For example, instead of the tweet ID being in the “id” column, the tweet content was in the “id” column. After exhaustively investigating why the error occurred we were unsuccessful in identifying the cause so we decided to re-pull all of the tweets. Fortunately, we did not encounter this error again.

combineTweetCSVs.py
Since the above script was run in batches this script combines all CSVs which contain Twitter data into one pandas DataFrame, sorts them by timestamp and outputs them into a CSV.

Tokenizing the Tweets

Manipulation

STEP 3: Tokenize the Tweets

tweetTokenizer.py

Removes any duplicates based on tweet ID and adds a date column. All of the tweets are then normalized by lowercasing the text, removing hashes, links and non-ASCII characters. Initially we wanted to include emojis but we ran into an issue where characters in other languages were leading to errors in the data. We felt the removal of emojis would not negatively impact our analysis so removing non-ASCII characters was the best fix.

After normalizing the tweets they are tokenized by using the NLTK TweetTokenizer module. We decided to use this tokenizer since it takes into account the ways which people tweet. For example, it understands that “can’t” is meant to be one word instead of “can” and “t”.

Any unnecessary punctuation and stopwords are then removed. The tokens are then lemmatized and put into bi-grams so we can see more context than just a single word. Lemmatization was used so that words with a similar meaning would be turned into the same word. For this step we did not use a part of speech tag but this is an improvement we would make in the future.

Once a list of bi-grams are created for each Tweet, the DataFrame is reduced to just the tweet ID, date, and bi-grams before using the explode method to create a new row for each bi-gram. The data is then aggregated to find the frequency of each bi-gram per day, another column is added which indicates the phase of the pandemic the tweet belongs too and it is output to a CSV.

tweetTokenizerSingleWord.py

The single word tweet tokenizer script works almost identical to the the previous script except it removes common words like “coronavirus” and “covid” and returns single words instead of bi-grams.

COVID-19 Data
Manipulation

STEP 1: Get the counts of confirmed cases for each country per day

covidCountsCountryDay.py

Transposes the initial CSV from wide format to long format so the countries are the columns and the dates are the index. It then filters the data down to the same date range as the Twitter data. In the original data there are several countries with more than one row since they are broken down by province or state. In order to get total confirmed cases per country each day these rows are aggregated by country and date. The script then adds a column which indicates which "phase" of the pandemic the data falls into.

During initial exploration of the data we noticed that the change in average number of new cases per day had four distinct changes. We thought it would be interesting to look at how the Twitter information changed depending on the pandemic phase so we included a "phase" label in both the COVID dataset and the Twitter data.

Lastly, the script outputs the date, country, number of confirmed cases and phase label to a CSV.

STEP 2: Get the total global confirmed cases per day and enrich the data

covidTimeSeries.py

This script performs the same manipulation as covidCountsCountryDay.py except instead of adding the total number of confirmed cases per country each day it calculates the total number of cases per day. It then uses the .diff() method on the total number of confirmed cases column to create a column with the total number of new cases per day. The .rolling() method is then used to calculate the seven day rolling average for number of new cases. A min-max scaled version of this column is also created so it could be visualized with the seven day rolling average sentiment score which is much smaller than average number of new cases per day.

The sentiment score and timestamp columns from the pre-tokenized Twitter data are then read in as a dataframe. The daily average sentiment score, seven day rolling average sentiment score and min-max scaled seven day average sentiment score are then calculated as additional columns

Lastly, the COVID dataframe and sentiment DataFrame are merged based on date, the pandemic phase label column is added and the data is output to CSV.

Top Mentioned Words with Total Cases

Analysis

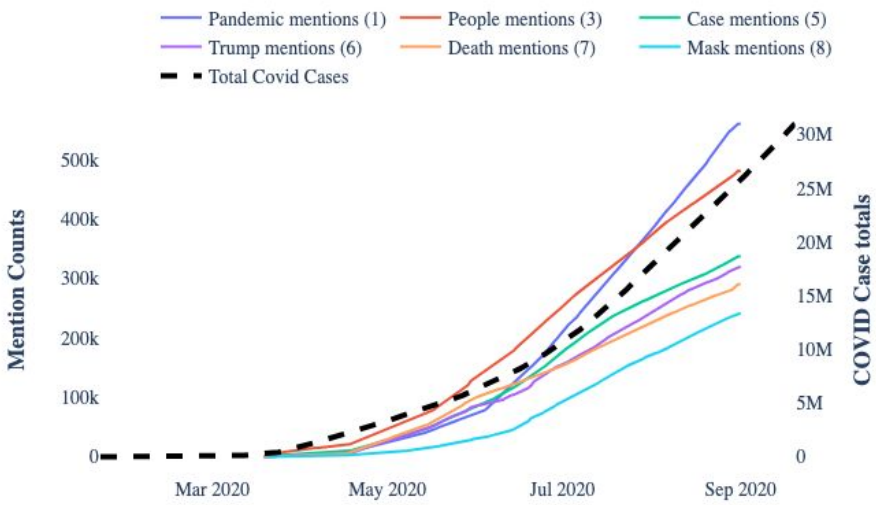
What are the most common words used when tweeting about COVID and how have those trended over the course of the pandemic?

Of the most top 10 most tweeted words over the course of the pandemic we have selected six to compare with the rise in total cases throughout the world. The six most mentioned words were “pandemic”, “people”, “case”, “Trump”, “death”, and “mask”. Several words deemed meaningless such as “ha” and “wa” were mentioned in the top 10 as well, but we removed them due to a lack of context.

It is not particularly surprising to find “pandemic” and “case” as the first and fifth most mentioned words throughout the pandemic. “People” was the third most mentioned word, likely mentioned in the context of social distancing, avoiding people, or mentioning nostalgia when seeing friends was more normal.

We do not see “Trump” or “mask” accelerate significantly until mid June, when they both see upticks in the mentions per day. During this same time period global case numbers also see an increase per day. “Trump” being the 6th most mentioned term also highlights how politicized the pandemic has become since March. While only looking at English tweets narrows down what governments would be referenced, it is still interesting that no other politician or world leader is mentioned within the top 100 terms.

NOTE: On April 18th the IEEE data added more coronavirus related terms when collecting tweet ids.



COVID and Sentiment Trends

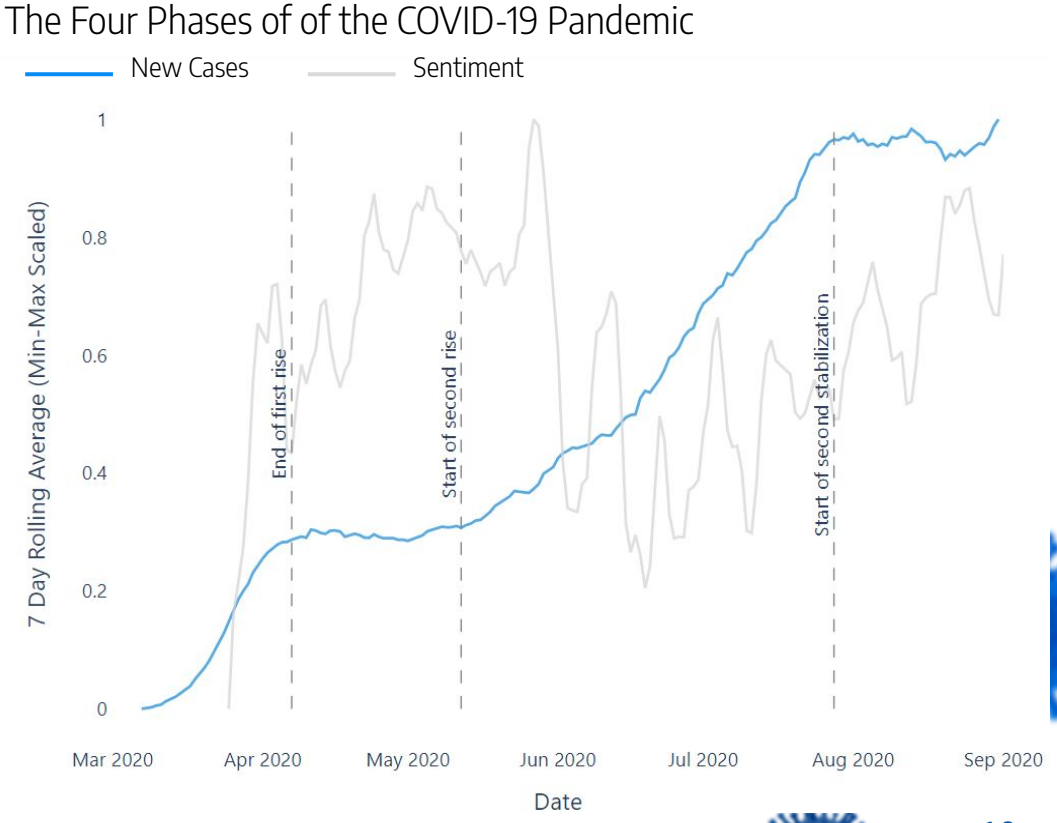
Analysis

How has sentiment changed as trends in number of daily confirmed cases has changed?

Exploration of the average number of daily new COVID cases shows there have been four distinct “phases” since the beginning of March 2020. In the first phase, the rolling seven day average of new cases shows a steep increase while sentiment appears to start off lower and increase toward the end of this phase. In the second phase we see a plateau with very little change in the average number of new cases per day while sentiment was on the high end of the scale, especially during the second half. The third phase includes another sharp increase in cases which lasts for a few months. Initially sentiment appeared higher and then dropped in June until entering the fourth phase which is another plateau but with a much higher average number of new cases.

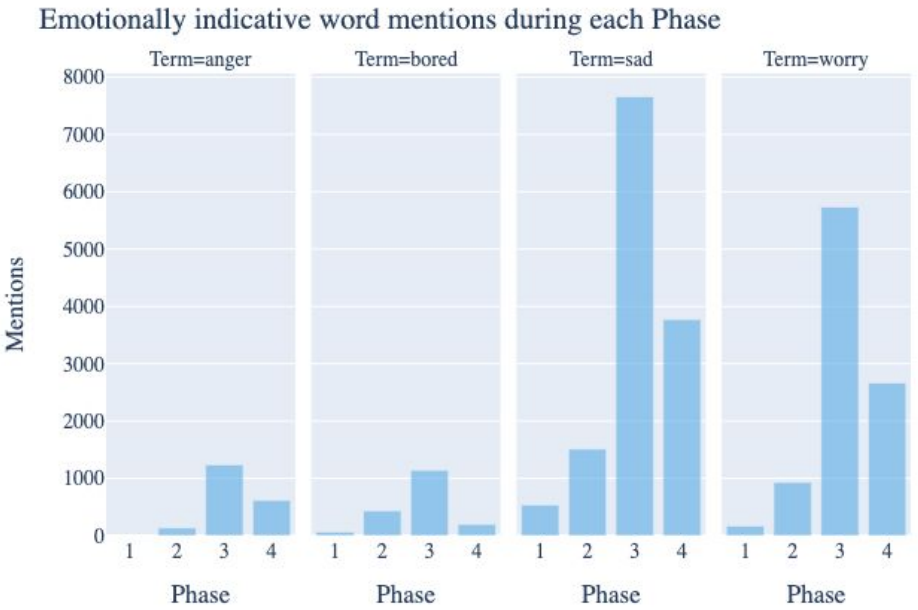
In order to plot the line chart we calculated the seven day rolling averages and then min-max scaled them since the original scales of the data are vastly different.

Although there appears to be a very large range in sentiment scores, due to this scaling, the maximum seven day average sentiment score only reached .067 while the minimum was .027. This means in general there was a very small range which hovered around 0. Part of this small range in sentiment scores may be due to the fact we sampled only 1% of the data or it could be that average sentiment has stayed consistent throughout the pandemic.



Insight into Emotional Wellbeing

Analysis



Is there evidence of burnout? Has the conversation moved to apathy towards the virus?

It is easy to identify broad topics or conversational trends by looking at common words but when it comes to identifying emotions like apathy and burnout that is significantly more difficult. Analyzing entire tweets and conducting more sentiment analysis would be the most effective way to gather that level of information but since this project focused on words and bi-grams we decided to see what we could find by looking at those.

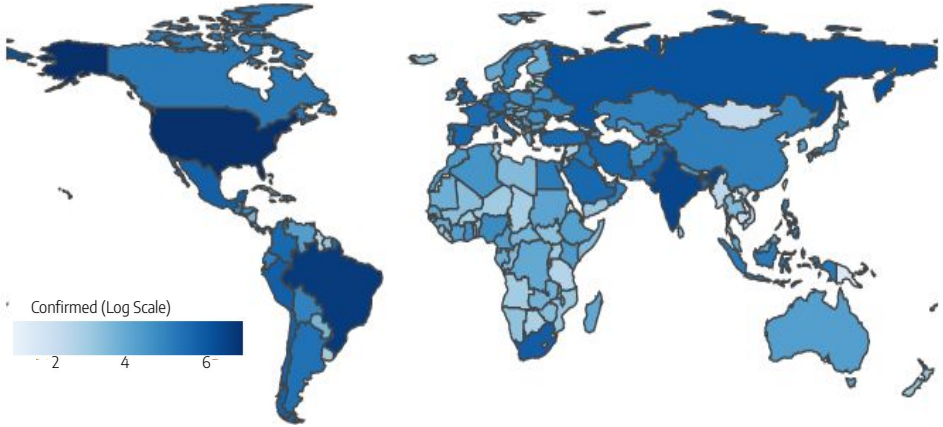
One proxy we came up with was mentions of words that would indicate the pandemic taking an emotional toll as well as looking at the sentiment scores that came with the IEEE data. As shown on the previous slide, in June sentiment dropped, but has been on a choppy upward trend ever since. However, as noted earlier, the average sentiment has stayed pretty consistent throughout the pandemic so this upward trend does not tell us a lot.

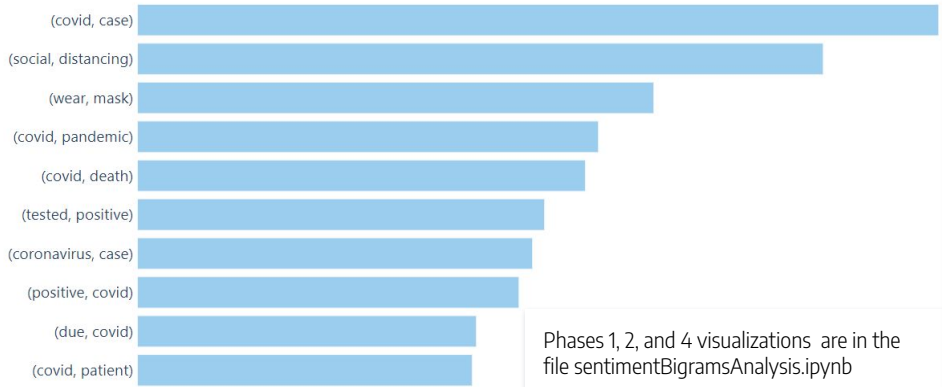
If we look at mentions of the words “sad” and “worry” we see they have significant volume in phases 3 and 4. This indicates that more people are experiencing emotions associated with burnout as the pandemic continues to progress. Further sentiment and emotive word analysis on a larger segment of data would need to be conducted to truly understand if there is evidence of burnout or apathy. If that analysis did show a similar trend of increasing apathy over the last couple months it would be interesting to see how that might impact COVID spread in the coming months.

Global Spread and Common Topics

Analysis

Covid Global Spread and Top 10 Phrases During Third Phase





Phases 1, 2, and 4 visualizations are in the file sentimentBigramsAnalysis.ipynb

How have tweets about COVID changed as it has continued to spread globally and is there a pattern which show tweets are impacting the spread?

At the end of the first phase, COVID had already spread worldwide with the US, Western Europe, and China being some of the main hotspots. During this period, tweets included phrases like “fight corona”, “stay home”, and “relief fund”. It is particularly interesting that “stay home” was one of the top phrases since this shows a focus on prevention which preceded the second phase plateau.

At the end of the second phase, the US was still one of the countries with the highest number of total cases but places like India, Brazil, and Russia saw large increases. Phrases on Twitter also shifted to “death”, “patients”, and the “white house”.”Fight corona” was still one of the top 10 phrases but dropped to number nine in the second phase and “stay home” did not make the top 10. Besides “fight corona”, none of the top 10 phrases had anything to do with prevention or combating the virus.

Once we hit the end of the third phase the hotspot countries did not change but their numbers increased dramatically. Conversations on Twitter again shifted to discussing more preventative measures like “social distancing” and “wear mask”. This is interesting since these conversations precede the fourth phase which is again a plateau in the number of new daily cases, although at a much higher rate. It would be interesting in future analysis to investigate if this pattern of increased frequency of tweets regarding ways to prevent the spread of COVID prior to plateaus in number of daily new cases is actually causal or significant.

Similar to the third phase, the actual countries with the most cases at the end of the fourth phase did not change. The number of cases also did not rise as significantly, except in India which doubled. On Twitter the top phrases were still focused on prevention with “social distancing” and “wear mask” taking the top two spots.

Statement of Work

Overall

This project is the results of a collaborative effort. We held meetings at least once a week to discuss progress, roadblocks, and the next steps. We used collaborative environments like github and Google Drive to host our scripts and data to make sure each team member had access to the entire project and could make improvements. Although we each took the lead on different sections we worked together to determine the best methods to use and helped each other through challenges.

Data Sources

Laura identified both the IEEE data and the Johns Hopkins COVID data. Ian identified Google Trends as another data source but unfortunately due to time constraints we were unable to incorporate this data source.

Data Manipulation

Ian created the scripts to manipulate the IEEE data. Initially he had wanted to incorporate a web scraper which would pull the CSV files from IEEE automatically but after a concerted effort and help from the instructional team this was unfortunately not possible.

Laura created the scripts to get the tweets using the Twitter API, combine them and tokenize them. When deciding how to complete the tokenization step Ian researched using Spark and Laura researched NLTK. The NLTK method seemed the best option since Laura was able to get what we needed rather quickly.

As for the Johns Hopkins COVID data, Ian initially created a script that combined all of the daily report CSVs but upon inspecting the output we realized that there were errors in the data. We found that Johns Hopkins does not change these reports if there are updates to the data after the fact. Fortunately, the time series data is updated for accuracy so we were able to use this data instead. Previously, Laura had manipulated the time series data for another project so she created the two COVID scripts. Lastly, with regards to the github repository, Laura created the repo but Ian took the lead on managing it including setting up the README file and the .gitignore file.

Data Analysis

Laura completed the analysis and visualizations for the COVID and Sentiment Trends as well as the Global Spread and Common Topics. Ian completed the Top Mentioned Words with Total Cases and the Insight into Emotional Wellbeing. Ian also created several interactive visualizations which allow the viewer to focus on very specific days within the pandemic and view the top 10 words mentioned that day. All visualizations were done with the Plotly library allowing the user to have a tooltip with more information on any of the charts in the presentation notebooks.

References

"How to Protect Yourself & Others (2020). https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html

Pineda, K. (2020, August 11). The world surpasses 20 million COVID-19 cases: A look at the milestones and setbacks. Retrieved from https://www.usatoday.com/story/news/nation/2020/07/23/united-states-coronavirus-cases-deaths-timeline/5485674002/

Shen, Y. (2020, March 16). Covid-19 Outbreak: Tweet Analysis on Face Masks. Retrieved from https://towardsdatascience.com/covid-19-outbreak-tweet-analysis-on-face-masks-27ef5db199dd

Slide template from www.slidescarnival.com