CS432: Final Project Report
COVID-19 Tweets Sentiment Analysis

Alison Lee, Ji Hoon Park

I. Problem

Since the first appearance of COVID-19 at the end of 2019, the virus has become a pandemic, spreading throughout the world as well has becoming a major topic in discourse online. With Twitter being one of the top social media platforms, we thought it would be an abundant source of data for analyzing thoughts on the coronavirus. Using tweet sentiment scores, we analyzed its correlation to major topics and case/death rates, as well as trends over time and geographical location. The tasks we focused on were:

A. Charting sentiment data on a timeline and comparing it to a Covid-19 Timeline.
B. Placing each sentiment data point on a geospatial scatter map to see if there's any correlation between sentiment and geographical location.
C. Finding the most popular topics for each sentiment category (very happy, happy, neutral, unhappy, very unhappy). This can be used to analyze topics causes of happiness and concern related to the virus.
D. Comparing geographic sentiment data to case and death rates for states and countries.

We used datasets from the IEEE Dataport [1], The Atlantic's COVID Tracking Project [2], Our World in Data [3].

II. Software Design and Implementation

The majority of the project was done using Python, utilizing PyMongo, Twarc, Plotly, pandas, and certifi libraries as well as the JSON and CSV data formats.

Tweet IDs were hydrated using the Twarc library, and the hydrated data was attached to sentiment scores provided in the dataset. After cleansing the data, there were over 500,000 geo-tagged tweets up for analysis. We decided on MongoDB Atlas for our choice of NoSQL database due to its support for JSON data formats and our familiarity with JSON-based querying/aggregation.

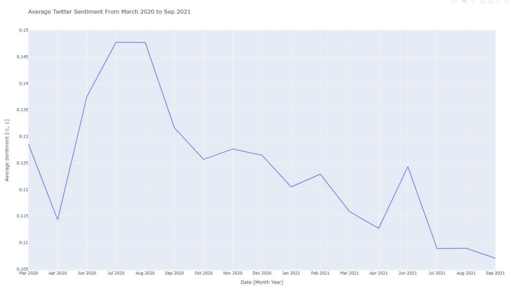
Ji Hoon: Implemented tasks A and B.
Alison: Implemented tasks C and D.

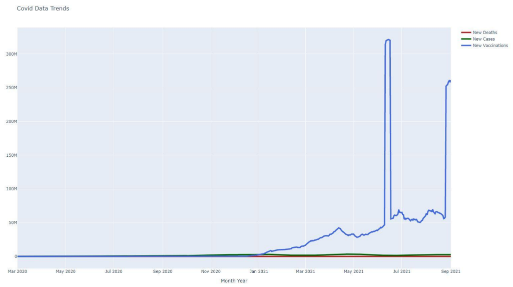
III. Project Outcome

In the following graphics, the sentiment scores have a range of [-1,1] with -1 being the most negative and 1 being most positive.

A. Average Sentiment Data Over Time & COVID-19 Trends Over Time

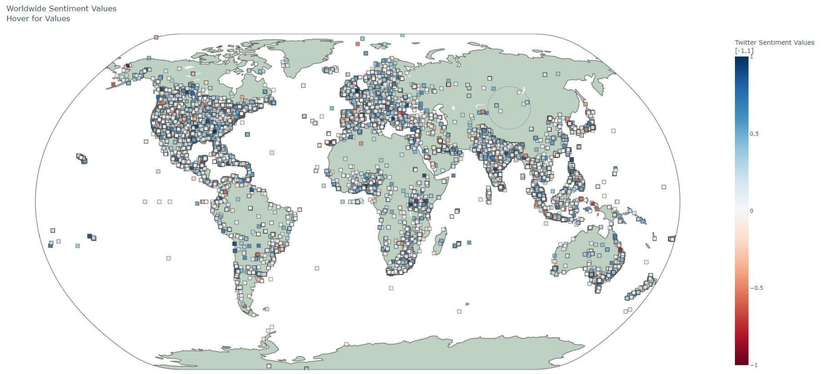
We found that sentiment values peaked following large spikes in new COVID-19 cases and related deaths. There were also instances where sentiment dipped following lulls or dips in new COVID-19 cases or related deaths. An interactive version of these visualizations can be found here.





B. Sentiment Plotted on Geospatial Scatter Map

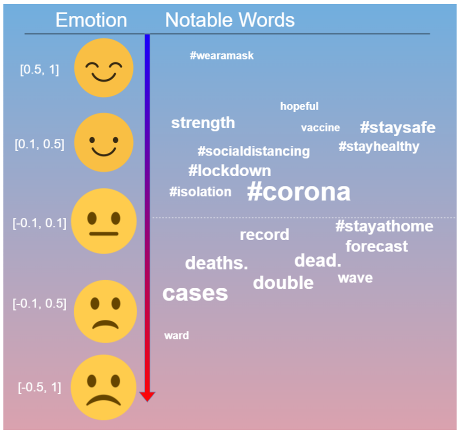
This data reflected the generally neutral-positive sentiment found across the world, however a trend could be seen when taking a deeper look at regions such as North America, Europe, and South Asia. Sentiment values were usually higher in more datapoint populated areas and lower in less datapoint populated areas. Overall, sentiment had a tendency to be more negative on the outside of datapoint population centers. An interactive version of these visualizations can be found here.

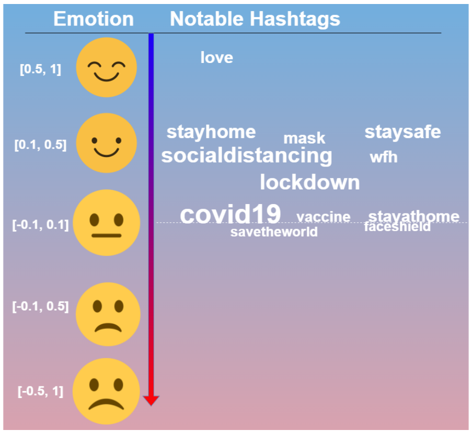


C. Most Popular Words and Hashtags by Sentiment

Due to the size of the dataset, complicated group by operations could only be done on a random sample. This may explain the more varied sentiments on words in comparison to hashtags which were, instead, collected over the entire database which skewed positive. In the random sample of tweet text, the most common words had negative sentiment as shown by the text size in the graphics.

(click the image hyperlinks for larger images)

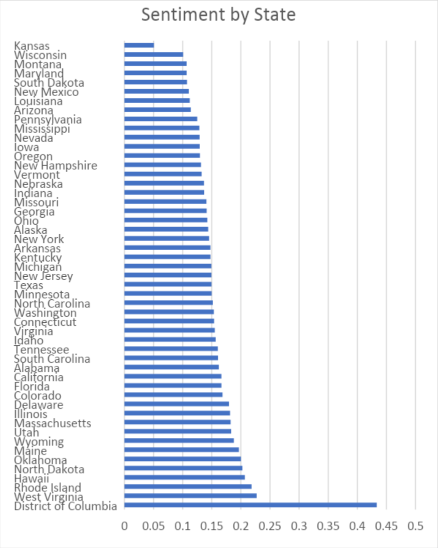


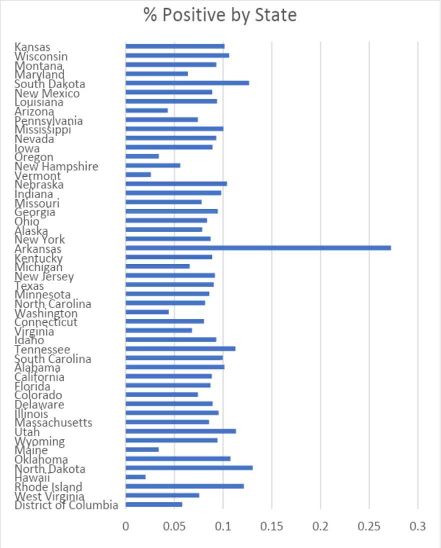


D. Geographical Sentiment Data Compared to COVID-19 Case and Casualty Rate

(click the image hyperlinks for larger images)

Most of the Twitter data came from the United States as the dataset held English tweets. Comparing the average sentiment by state to the percentage of cases and deaths, we concluded that there was no correlation. This lack of correlation also translated to data for countries and their case/death rates. A possible explanation for this is that it is due to the sentiments being positive on average throughout the entire database.





IV. References

1. Rabindra Lamsal. (2020). Coronavirus (COVID-19) Geo-tagged Tweets Dataset. IEEE Dataport. https://dx.doi.org/10.21227/fpsb-jz61
2. (2020) The COVID Tracking Project. The Atlantic. https://covidtracking.com/data
3. Hannah Ritchie, D., & Max Roser (2020). Coronavirus Pandemic (COVID-19). Our World in Data. https://ourworldindata.org/coronavirus-source-data