Covid19 Student Engagement Award: Midterm Report

What activities and outcomes have been achieved through your project to date *

To date, we have done a lot of exploratory data analysis- data extraction, manipulation, and visualization. The data we obtained has been crawled from ids provided in the TweetsCov19 dataset, https://data.gesis.org/tweetscov19/. The data are contained in the files 'tweets.txt', 'users.txt', 'places.txt' and 'media.txt'. All tweets are contained in the file 'tweets.txt'. Each line of this file represents the information crawled for one tweet id, and has (mainly) the following JSON keys: [id, author_id, created_at, text, public_metrics, entities, geo].
From this dataset, we have:

- Created geotweet data frame by selecting only data from US tweets
- Filtered out election-related tweets
- Produced time-slider choropleth using weekly state tweet-counts
- Produced time-slider grid map using weekly state tweet counts
- Investigate several normalization techniques, including by population, #unique users and #tweets.

We then performed sentiments analysis of tweets. To allow for a meaningful comparison of sentiment and covid metrics, the dataset was divided into 2 general categories: family_df (tweets containing family-related terms), and depression_df (tweets containing depression-related terms). Word clouds and plots of positive/negative sentiment ratio were produced for each category. Topic modelling using the GenSim module in Python revealed clusters of tweets that were similar in nature for each category.
To improve the quality of sentiment detection, the 2007 Language and Inquiry Word Count was used to provide anger/sad/joy/etc counts instead of just positive/negative. After the normalized number of cases per week in each state showed not to correlate with the number of tweets in that region, a decision was made to focus on mask hesitancy, vaccine hesitancy and their relationships- performing exploratory analysis with pro-mask and anti-mask labelled tweets. We have developed a basic algorithm to distinguish mask-related attitudes by hashtags; current work revolves on improving by investigating regex matches with common phrases.

Have these activities differed at all from what you had originally planned? If yes, why?

Yes, the activities have differed from the original plan. We were hoping to implement a Machine Learning model, such as an LSTM, to predict covid cases and deaths by state. Such models are only justified if there is a correlation between the number of tweets and case/death count normalized by population. However, almost all states show no correlation at all, with r^2 values of Covid Cases/Deaths per Capita vs Number of Tweets reaching as low as 0.001 for some states.
This forced us to continue doing exploratory data analysis. We decided to investigate mask sentiments next; to evaluate the number of pro-and anti-mask tweeters by state. This posed another problem- Python Sentiment analysis using Vader is very bad at detecting mask-specific sentiment. For example, a tweet: "Wear a damn mask fools!" would be classified as negative by Vader, even though it's pro-mask, and thus we want it labelled positive. So, we worked on our algorithm for mask-sentiment detection, but it performs poorly right now. We are planning to switch from using hashtags for labelling to using regex patterns.

We are also not planning to launch this into a web framework. In our initial proposal, we stated a report or website; I think our results are more suited for a report on what twitter data is useful as a detector of various covid metrics, and what is not.

What challenges have arisen throughout the course of your project thus far? Were they expected or unexpected? How have you responded to these challenges? *

Throughout the course of this project, we have faced several challenges that forced us to continuously refine our exploratory data analysis and vision for the project.
The first challenge was noticing that a lot of the dataset was garbage- even though the entire set was extracted from scraping covid-related tweets, there were lots of tweets that referenced Covid19 but did not address the pandemic at all and had nothing to do with attitudes towards masks or vaccines. To deal with this, we used python's topic modelling libraries to identify clusters of tweets; manually identified topics for these tweets, and then filtered them out when creating our baseline "geotweets_df" data frame.
Our first visualization of the data was focused on understanding relations between case/death count and tweet count. Since this showed no correlation for almost all states, we scratched out the idea of an LSTM prediction model. We decided to focus on using twitter as sensor for identifying the population affected by quarantine, as well as locating pro- and anti-mask tweeters.
This posed another challenge; the Vader sentiment module, which is the standard library for sentiment analysis in Python, was performing very poorly because it was reporting general tweet mood rather than mask hesitancy. So, we implemented our own very basic algorithm by classifying 3000 hashtags as positive, negative, or neutral; and plan on iteratively improving it using regex match patterns.

What existing skills have you used most throughout the course of this project, eg. knowledge in a field; problem solving skills; working with others to achieve results; communication skills through presentations and reports development; inter-personal skills? *

Throughout the internship/project, I have used several skills, allowing me to further enhance them. The most important by far has been knowledge of Python, particularly data science and social network libraries like Pandas, Numpy, Folium and NetworkX. I have also used problem-solving in every stage of the project; when doing exploratory data science, a lot of the time you search for patterns that you expect already- not only does this often fail (as in our covid cases/deaths vs #tweets), it also doesn't make for a very interesting paper. Looking at problems from a broad view and allocating time to just understanding your data so you can come up with questions people in your field would care about is a skill I underestimated at the beginning of my internship but am now trying to improve actively. Working with others has been very important to the productivity of our project. The project involves 2 interns and many supervisors to provide direction. It is very important that me and my fellow intern don't do the same things to avoid redundancy, communicate progress and struggles with our supervisors, and learn the relevant data science terminology to understand their instructions & advice. Presentation skills were also important. At one point, the supervising professor invited some friends for advice and to join the

project. Me and John led the initiative to create and practice a presentation to help them with onboarding/ramp up.

What new skills have you gained or do you expect to gain by the end of your project, eg. knowledge in a field; problem solving skills; working with others to achieve results; communication skills through presentations and reports development; inter-personal skills?

Throughout the internship, I learned a lot about data science. With time, I could appreciate the art behind getting from a large dataset to conclusions about the dataset. I learned that proper data science requires that you understand your data well and is a very nuanced and iterative process. For example, data filtration and data classification are iterative and require constant testing and improvements. Outside the project, I learned a lot about social network analysis in the digital age through the book “Bit by Bit” (Salganik) and about data structures and algorithms for social network computation through the book “Networks, Crowds and Markets” (Easley, Kleinberg). I also expect to learn a lot about report writing and maybe how to write a good paper, though a publication seems out of scope for now…

How much of your funding have you spent to date? What have you spent it on? *

Kept all funding as a stipend for effort spent on the project..

Has your budget differed at all from what you had originally planned? If yes, please explain: *

No.