**Introduction**

Twitter is one of the biggest social media platforms in the modern age. Started in 2006, the platform distinguished itself from other social media by popularizing the use of microblogging, which is a shortened version of blog posts which are meant to give updates throughout the day rather than longer blog posts which typically take a little while to read. For the past decade almost, Twitter has also been a great source of data. With close to 350 million tweets per day, Twitter is a hub for people to discuss anything and everything regarding their lives. In the past Twitter data has been used as a data source for everyone from politicians to scientists. The millions of tweets per day serve as a sample of the mind of the population.

In 2020, the world was hit by the Covid-19 pandemic. The Covid-19 virus is a respiratory disease which is quite contagious and causes a range of symptoms depending on the person from minor coughs and colds to more fatal respiratory damage which has led to death in many cases especially in people with advanced age. The disease caused a nearly worldwide shutdown with many forced to stay inside to protect themselves from Coronavirus. This long period of staying indoors caused was a major event and a lot of the sentiment from that period can be seen reflected on Twitter. After nearly a year the first vaccines were made available in December of 2020. Despite assurances by the FDA, there was still a lot of skepticism from people regarding the vaccine with concerns regarding the speed with which the vaccine was developed, the interests of the parties which created it, to various conspiracy theories surrounding true intentions of the vaccine. With all this skepticism there was an on-going war between those who believed in the vaccines and those who were skeptical and refused to take it. There were many efforts, however, it was quite difficult to mandate vaccines for citizens so there was always an aspect of anti-vaccine sentiment in the United States. There was no better place to see the minds and the strongest opinions of both opposing sides than Twitter as it was a place for people to speak their mind, connect with people with similar opinions, as well as interact with people from the opposite side in an often uncivilized manner.

I had a few goals regarding this project split into personal goals, and project goals.

Personal Goals

1. Gain experience in working with larger datasets. Specifically, using optimizations techniques as to not bloat run times for analysis of big data.
2. Learn more about sentiment analysis and use it to classify text data based on a subject area.
3. Analyze two sets of data to extrapolate insights and make an informed, data-backed conclusion on whether they are correlated or not.

Project Goals

1. Identify key points of negativity and positivity on Twitter in 2021 and investigate what could have caused them
2. Identify the key points of negativity and positivity and correlate them with real Coronavirus statistics to see how those events affected the Coronavirus statistics.
3. Determine how correlated the sentiment of Twitter is with the Coronavirus statistics.

**Related Works**

**Data Collection**

The main dataset I used to source the Tweets was created by Doctor Christian E Lopez and Doctor Caleb Gallemore, both from Lafayette College in Pennsylvania. The dataset was created for a paper proposing the dataset as an archive for research to be done on the social discourse of the Coronavirus. The way the data was collected was the researchers at Lafayette college used the Twitter API’s research license which allows 10 million tweets to be accessed per month. With that research license they set up a continual collection of tweets using a set of keywords related to Coronavirus such as covid, coronavirus, and more. In addition to the research license the researchers also used other popular datasets being created in academia at the time to supplement their dataset. The dataset itself has over 2.2 billion Tweets collected from 2020 to 2022 from Twitter regarding Covid-19.

The dataset I used for getting Coronavirus data was the one created by the publication called Our World in Data. Our World in Data is an online resource which creates dataset and graphs regarding global issues to make available to the public for research and other purposes. There is a multitude of researchers who have worked to keep this data up to date and maintained for easy usage including Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian, Esteban Ortiz-Ospina, and Max Rose. The dataset contains data regarding all aspects of the Coronavirus for as many countries as they could include without breaking any rules. The columns we focus on for this paper are the new vaccinations column, which tells us the number of new vaccines administered on a given day, and the new cases column, which tells us the number of new cases of Coronavirus on a given day.

Twitter data is easily accessible using the Twitter API, and my initial approach to getting Twitter Data was to use the API to generate large dataset of vaccine related tweets to make the dataset rather than finding a dataset on the internet. There were two big issues which prevented me from doing so. The first issue I had, was that using just a normal Twitter developer account, the number of tweets we could get per month was not nearly enough to get the sense of scale we needed for this project. The rectification for this issue was to create a Twitter research license. The issue with getting a research license was that the processing time to get approval for the license is a few months, so that would have significantly cut into the time I had this semester to work on the project. The solution was to use a pre-created Twitter dataset. The dataset, created by the Lafayette researchers mentioned above, leverages said research license from Twitter and was found on Kaggle, a user driven site, where users can create and upload their own datasets about any topic as well as search and use datasets created by other users on the platform, where the Lafayette researchers had their dataset hosted.

**Preprocessing and Data Transformation**

A large portion of my time was spent building the dataset for analysis considering the dataset itself was very large. The dataset in total had 2.2 billion tweets so for me to do analysis upon it I had to filter the tweets down to a more manageable number. The way the Twitter data was stored was in CSVs, with one csv per hour for each day of the year. All this data is also separated into tweets details, tweet sentiment, and tweet hashtags. To combine the data into a master dataset, I had to iterate through each and preprocess pieces of the data and concatenate it to the master. The problem was that with doing this, I had to keep an eye on the RAM as creating the master couldn’t be held solely on the laptop.

The first iteration of the dataset I built was built using all the tweets in the dataset. I believed it would be possible potentially use a generator to pass data through to the analysis cells, however, I did have a little bit of an issue with finding a way to create graphs in that way, so the next idea I had was to sample the data. The idea was to use 25 to 50 sampled data points from each hour of each day and add it to the master dataset.

MLA

Lopez, Christian E, and Caleb Gallemore. “An augmented multilingual Twitter dataset for studying the COVID-19 infodemic.” *Social network analysis and mining* vol. 11,1 (2021): 102. doi:10.1007/s13278-021-00825-0

<https://pubmed.ncbi.nlm.nih.gov/34697560/>

Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian, Esteban Ortiz-Ospina and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]