**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 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 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 focused 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 the 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 made it difficult to do so. The first issue encountered, was that using just a normal Twitter developer account, the number of tweets that could be retrieved per month was not nearly enough to get the sense of scale 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 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 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 analysis to be done upon it the tweets had to be filtered 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, each piece of data had to be preprocessed and concatenate it to the master. The problem was that with doing this, attention had to be paid to the RAM as creating the master would overwhelm the laptop if the data wasn’t properly stored.

The first iteration of the dataset was built using all the tweets in the dataset. The original idea was to use a generator to pass data through to the analysis cells, however, there was an issue with finding a way to create graphs in that way, so the next idea was to sample the data. To start, the folders had to be iterated with the sentiment then read in each hour csv in separately. The read dataframes were then sampled and added it to the master dataframe. With this iteration, after a little bit of initial analysis it was realized that this iteration of the dataset wouldn’t be helpful as there was no way to make sure that the data points being looked at were from the United States as a lot of the datapoints were coming from different countries upon further inspection.

The next iteration was focused on making sure that most of the tweets we’re looking at are coming from the United States. The first method was to maximize the number of tweets from the United States was to filter language to only English. The other method incorporated into the program was created by doing research into tweeting in the United States. The peak tweet times for the United States is from 11am to 1pm according to Twitter. So, the idea was filter to only tweets from the peak hours in the United States. The way this was accomplished was by breaking the file names into pieces and only sampling from the peak hours.

The final iteration of the dataset was focused on making sure that the tweets that gotten were related to the vaccine rather than just the coronavirus in general. Coronavirus tweets in general wouldn’t give the correct sentiments that was being looked for as people would likely speak about corona in a negative sentiment. To combat this, it was made sure that the hashtags that were being used were related to the coronavirus vaccine. This brought up another issue as the hashtags and the sentiment data for the tweets were stored in separate file directories. The issue being that a way had to be found to first verify that a tweet had a hashtag related to the vaccine, so to do that the hashtag dataset was used to create a csv containing all the data tweets from the peak times that were related to the covid vaccine using keyworks such as ‘vax’, and ‘vacc’. Once that csv had been created, a set was used to create the group of all tweet IDs which fit the criteria. The reason a set was used is due to the underlying structure in sets. Sets use a hash table to store the data making it easy for quick access. It had originally been tried inner joining the hashtag and the sentiment dataset; however, this ballooned the runtime making it take over three days for the code to run on the computer. With the set implementation the runtime was cut down to just under 24 hours as sets are optimal for quick access to see if a value exists within it. After the dataset was created using the sentiment and hashtags, which were then by day easily to build the final dataset, which was done by grabbing a maximum of 100 tweets per hour as a sample size. In the end, the final dataset had over 800,000 data points.

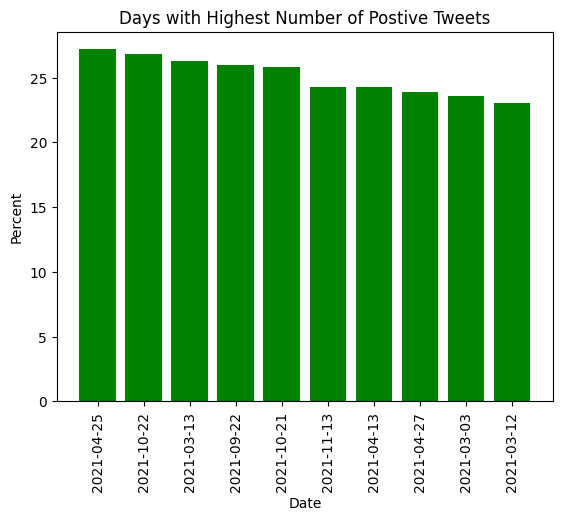
**Analysis**

­­­­

The dataset created had each tweet as its own row, this wouldn’t be very helpful as the best way to analyze the tweets over time would be to have them grouped by date with the negative and positive counts. To transform the dataset to this format, the groupby function in pandas had to be used. This function creates groups for each of dates holding all the data for each group. With these created groups, the number of each category of sentiment was summed up (negative, neutral, or positive). With the counts, the percentages also had to be converted as not all days would have the full number of tweets. Meaning that were days that did not have the 100 tweets related to the coronavirus vaccine every hour, so in those cases the function took the maximum number of tweets that fit that criterion. By using the percentage of each category, the imbalance in days could be combatted.

Chart

Description automatically generated

The first set of data investigated was the sentiments by day. As you can see in the above graph, there was the neutral, positive, and negative precents for each day of the year. As hypothesized, there is a definite inverse correlation between the positive and negative lines. There are multiple points circled above when spikes can clearly be seen in the positive resulting in negative downwards spikes and vice versa. What this tells us is that there is there are events that are causing these spikes as not only is there a spike in negative sentiment, but the positive is being affected and the spike is not just being compensated by the neutral sentiment.

Next, the days with the highest values in positive sentiment were investigated. The goal with doing this was to hopefully identify some key causes of this jump in sentiment. As can be seen from the graph above the days presented are those with the highest number of positive tweets. These days were investigated to hopefully gain some insight to what caused the spike. The first day with the highest number of positive tweets sampled is April 25th, 2021. After some research and analysis of the trending hashtags the can be seen was one of the most talked about events that day was the American movie awards, the Oscars. The Oscars was a massively publicized event and they had coronavirus protocols in effect for the event. They had masks required for certain parts of the event. This massive viewing likely brought in a lot of new discussion on the vaccine as this was the first Oscars that occurred during the pandemic.

Chart, bar chart

Description automatically generated The next analysis done was on the days with the highest negative sentiment sample of tweets. Something interesting found about this graph, as seen above, is that the highest negative day has a far more significant uptick than the highest positive day, so that day was investigated first. The day with the highest percentage of negative tweets sampled was September 9th, 2021. The best indication found which may have led to this large uptick in negative sentiment was using the twitter trending hashtags once again. On September 9th, 2021, the top trending hashtag was #ARMYvaccinatedtoo, after further research it was found that this hashtag is in reference to a speech given at the United Nations by the Korean Pop Group called BTS where they encouraged vaccinations.

The next analysis went into was looking into the biggest changes in sentiment from day to day. The way this was done was by looking at the difference between the sentiment of a day and the day before it. When investigating these values, they were graphed and charted, but after looking more into the dates, there was no new noticeable insights from the data gotten from the highest positive and negative days, so the findings from the changes were omitted in this report.

Square

Description automatically generated

To see the correlations between the sets of data, a correlation matrix was created to see how each of the variables how they affect each other. A focus was placed on a few of the correlations that were interesting and worth pointing out. The positive percent and the new vaccination rate had a correlation of around .4, which is as hypothesized. Having a higher correlation would mean that as the positive sentiment towards vaccines increase the number of new vaccinations increases. The next point of interest was the correlation between negative percent and new cases these two had a correlation of around .3, which was quite high, but made sense as well considering new cases occur when people are not vaccinated. The opposite is true for negative percent and new vaccinations as they are indeed negatively correlated with a correlation of around -.3. This also makes sense considering that the more negative sentiment there is towards the vaccinations, the less vaccinations there will be.

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]