IEOR 243 - Module 1 Report

Group 10

Haolan Mai, Jiaqing Li, Yucong Li, Ruoxin Xu, Makie Maekawa, Bennett Cohen

February 14, 2023

1 Project Description and Objectives

Our primary project objective is to provide insights into public sentiment toward COVID-19 and how it has evolved over time, as well as to empower decision-makers with the ability to explore this information in a meaningful way. Specifically, we can break down our objectives for each of the three modules as follows.

Module 1: Data Collection and Exploration

- To obtain the COVID-19 tweets dataset either through Kaggle and by using an API to scrape more recent data.
- To describe the components of the data, and its limitations such as regional bias.
- To perform data validation such as checking for missing data.
- To perform data processing and cleaning steps to make sure the textual data is suitable for analysis.
- To create visualizations that highlight interesting aspects of the data, such as the most common words in tweets.

Module 2: Model Development and Performance Evaluation

- To perform sentiment analysis on the COVID-19 tweets dataset.
- To select and apply appropriate models and methodologies to achieve the sentiment analysis goal.
- To choose the best models based on performance and validation metrics.
- To explain why the selected models were chosen and what their limitations are.
- To tune the models to optimize their performance.
- To evaluate the performance of the models and measure the results using appropriate metrics.

Module 3:Interactive Visualization and Conclusions

- To create a high-level, concise, and coherent summary of the final model.
- To provide an interactive tool such as a web visualization using Flask that allows the target audience (such as public health officials and policymakers who are interested in understanding public sentiment towards COVID-19) to explore the consequences of different decisions related to COVID-19.
- To explain to the target audience how their decisions will impact the sentiment analysis results.
- To communicate the effects of the decisions in a clear and understandable way to the target audience.

2 Data Description and Source

In this project, we collected tweets data from Kaggle and the Twitter API via Python, and we extracted tweets that contained the hashtag `#covid19`. Because our goal is to compare tweets during and after the pandemic, we filter our results into two separate tables with identical schema. To access the Twitter API, we modified a script found here: https://github.com/gabrielpreda/COVID-19-tweets.

- `COVID2020.csv` contains tweets between July 25th, 2020 and August 29th, 2020 (179,108 rows).
- `COVID2023.csv` contains tweets between January 29th, 2023 and February 8th, 2023 (24,352 rows).

Column Name	Description
user_name	account user name
user_location	account user location
user_description	account user description
user_created	when the account was created
user_followers	# of followers of the account
user_friends	# of friends of the user
user_favorites	# of favorites of the user
user_verifed	boolean if the user is verified
data	date tweeted
text	text of the tweet
hashtags	list of strings of hashtags
source	source of tweet
is_retweet	boolean if tweet is retweet or original

Table 1: Data source colum descriptions.

2.1 Dataset Limitations

It would imprudent to dive into data analysis without first contextualizing our datasets. There are various limitations that we see in the data we've collected:

1. User Location/Regional Bias: People in certain regions, such as China, do not have access to Twitter without using a VPN. Therefore, we could not obtain opinions of COVID-19 from people in these areas.
2. Time Zone Bias: Twitter users come from different time zones, and we collect data uniformly from a single time zone, so we might introduce bias here.
3. Data Bias: The data we pulled directly from Kaggle is already quite cleaning, so certain steps may have been taken to remove outliers, impute missing data, etc.
4. Social Media Bias: Many people use other social media platforms instead of Twitter (Instagram, Facebook, Snapchat, etc).

We also must note that due to the access level of the Twitter developer account, we can only crawl the recent-7-day of tweets, so the size of the data is much smaller compared to the 2020 dataset.

3 Data Loading and Inspection

We will refer to the tables `COVID2020` and `COVID2023` to distinguish between the two tables when necessary. Upon loading the data in, we can see the distribution of the numerical statistics using Pandas `describe` method.

	user_followers	**user_friends**	**user_favourites**
count	179108.00	179108.00	179108.00
mean	109055.53	2121.70	14444.11
std	841467.00	9162.55	44522.70
min	0.00	0.00	0.00
25%	172.00	148.00	206.00
50%	992.00	542.00	1791.00
75%	5284.00	1725.25	9388.00
max	49442559.00	497363.00	2047197.00

Table 2: Distribution of `COVID2020` numerical features.

	user_followers	**user_friends**	**user_favourites**
count	24352.00	24352.00	24352.00
mean	70776.39	2506.54	20891.31
std	715123.12	18074.44	56820.15
min	0.00	0.00	0.00
25%	208.00	127.00	289.00
50%	1125.00	577.00	2975.00
75%	3892.00	1870.00	14439.75
max	16216538.00	477573.00	1342610.00

Table 3: Distribution of `COVID2023` numerical features.

Upon inspection, we see there are many missing values in our data. The table below shows the raw count of missing values, along with the relative amount.

	Total	Percent	Data Types
user_name	0	0.0	object
user_location	6152	25.262812	object
user_description	1256	5.157687	object
user_created	0	0.0	object
user_followers	0	0.0	int64
user_friends	0	0.0	int64
user_favourites	0	0.0	int64
user_verified	0	0.0	bool
date	0	0.0	object
text	0	0.0	object
hashtags	6724	27.611695	object
source	0	0.0	object
is_retweet	0	0.0	bool

Table 5: Distribution of missing values in `COVID2023`

	Total	Percent	Data Types
user_name	0	0.0	object
user_location	36771	20.5300	object
user_description	10286	5.7429	object
user_created	0	0.0	object
user_followers	0	0.0	int64
user_friends	0	0.0	int64
user_favourites	0	0.0	int64
user_verified	0	0.0	bool
date	0	0.0	object
text	0	0.0	object
hashtags	51334	28.660	object
source	77	0.0429	object
is_retweet	0	0.0	bool

Table 4: Distribution of missing values in `COVID2020`

4 Data Cleaning

For both of the `COVID2020` and `COVID2023` tables, we apply the following steps to clean the data, along with a brief rationale.

1. Drop duplicate columns, defined as directly identical tweet `text` column. We combined the Kaggle data with our own, so there was an overlap.
2. Drop content from users with no follows (i.e. user_follower = 0) because a user without followers does not produce content that spreads to other users. Some people may have created accounts simply to complain, so it may not be valuable in understanding the distribution of sentiment.

3. Drop rows with N/A values. Because we have many rows of data, but only a few columns, all of which are vital to describing a tweet, it is fine to drop any N/A rows.
4. Process the tweet text data (`text` column) using the most common text cleaning methods covered.
    - Convert all letters to lowercase
    - Removed URL
    - Deleted any numbers in the text column
    - Deleted punctuation, stop words from NLTK, emoticons/emojis, and abbreviations.
    - Stripped leading/trailing and extra spaces.
5. Removed emojis from `user_name`, `user_location`, and `user_description`

5 Exploratory Data Analysis (EDA) - Pandemic vs Post-Pandemic

5.1 Bar Plot of Most Common Words in Tweets

Our natural first step was to explore the frequency distribution of the most common words in the tweet text body. We visualized this by displaying the top 20 most frequently used words (excluding stop words as they already have been purged). Unsurprisingly, the word "COVID" was used far more frequently than any other word. The second and third most frequent words are "cases" and "coronavirus".

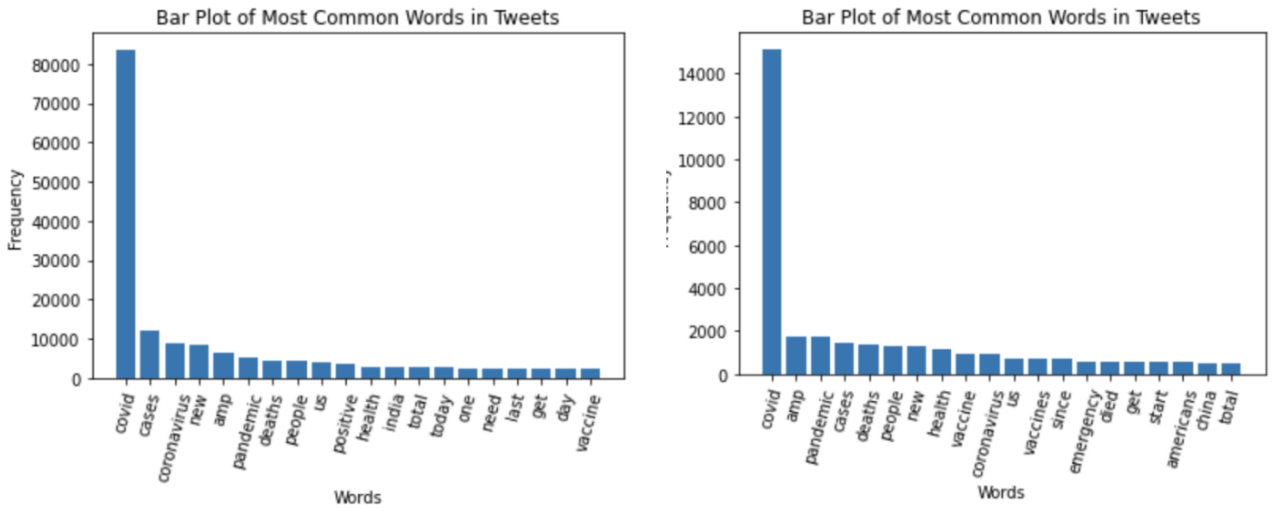

Figure 1: Bar plots of the frequency of 12 of the 20 most common words in the `text` column of our dataset.

5.2 Distribution of Length of Tweets

We create a histogram to see the characters of tweet lengths. This is a left-skewed distribution, according to the image, because it is a distribution with the tail on the left. If the data distribution is skewed to the left, the mean is less than the median, which is usually less than the mode. In this image, most tweets are around 70-90 characters.

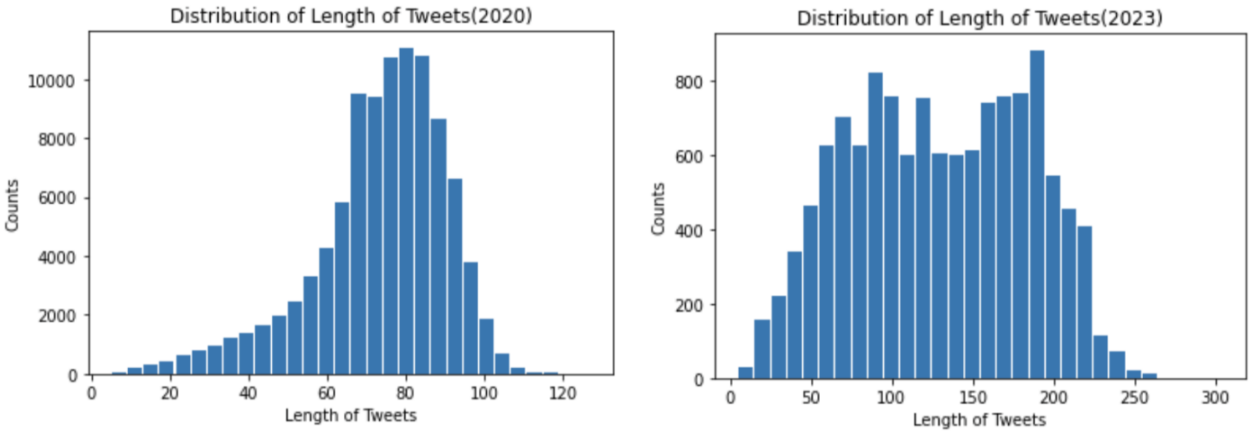

Figure 2: Histogram of the distributions of tweet lengths in 2020 and 2023.

5.3 User Location Analysis

Before creating the word cloud, we clean the `user_location` column. Then, we draw the word cloud for the `user_location` column after cleaning. Based on the word cloud we have mapped, we can see that there are many regions that appear on the word cloud. The larger the word cloud displays, the higher the frequency of the word, and therefore the higher the frequency of the region in the data we obtain. We can clearly see that the users we collected are often distributed in some countries or cities, such as New York, Delhi, Washington DC, Mumbai, South Africa, the United Kingdom, Australia, etc.

User_location for covid2020 User_locationin for covid2023

Figure 3: Word cloud of the tweet location distribution.

5.4 Text Word Clouds

Here, we want to find the common words in tweets about COVID-19 and investigate if there is a difference in keywords between 2020 and 2023 data. From the user location analysis above, we obtained some locations where users are concentrated. Without losing generality, we generated word clouds for different representative locations, such as India, the United States, the United Kingdom, Canada, South Africa, and Switzerland.

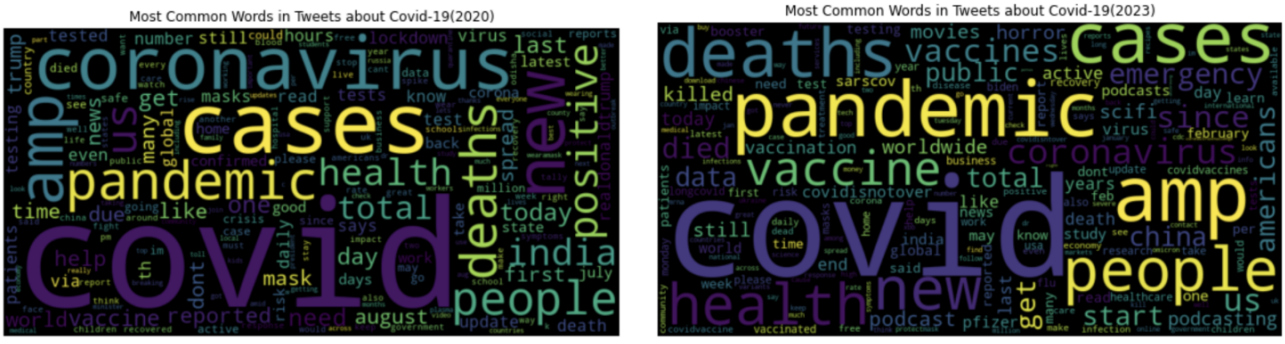

Figure 4: Word cloud of the most common words in all tweets.

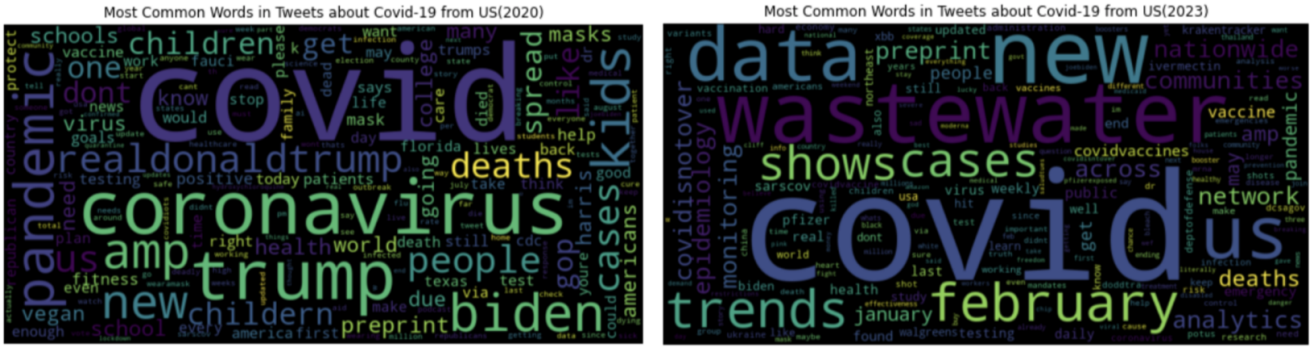

Figure 5: Word cloud of the most common words in tweets from the United States.

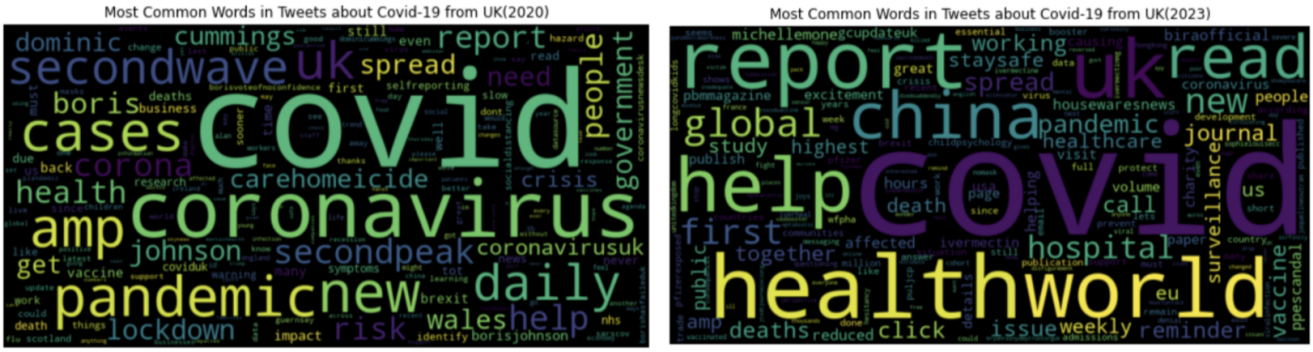

Figure 6: Word cloud of the most common words in tweets from the United Kingdom.

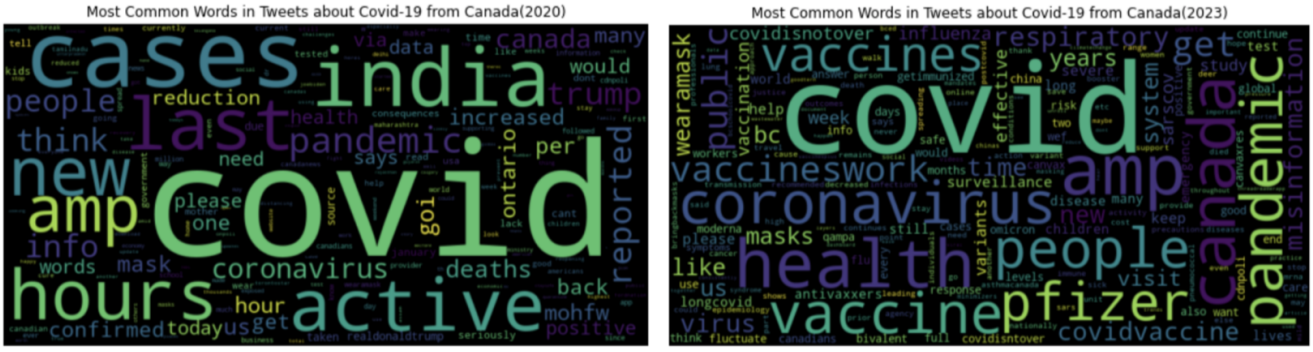

Figure 7: Word cloud of the most common words in tweets from Canada.

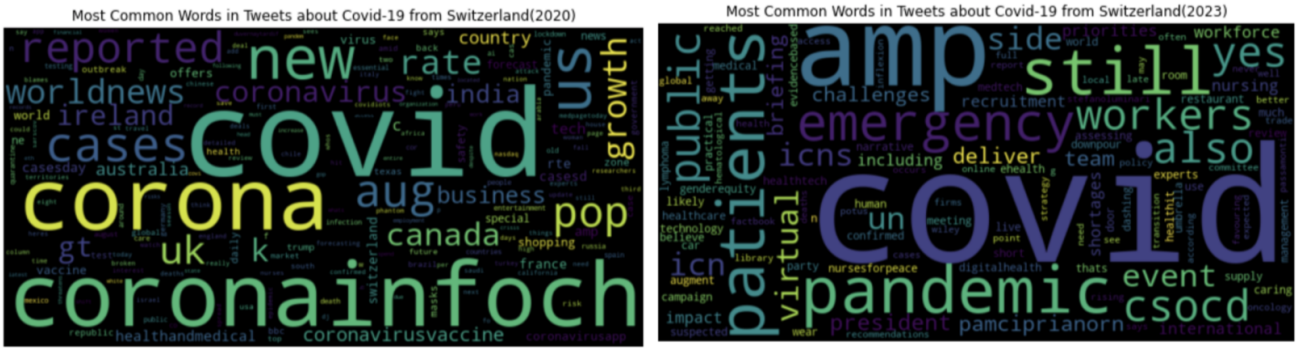

Figure 8: Word cloud of the most common words in tweets from Switzerland.

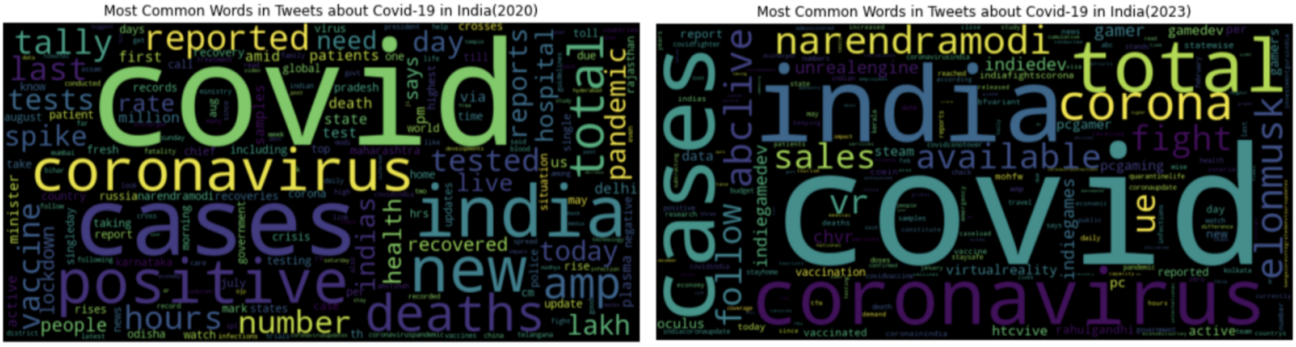

Figure 9: Word cloud of the most common words in tweets from India.

These plots show that the most frequent word in tweets published in these locations in both the 2020 and 2023 datasets is "COVID". This states that the discussion of the COVID has been very high among users in the past three years. This also gives us a positive signal that our 2020 and 2023 datasets are both COVID-related, ensuring that our future epidemic sentiment analysis will not go off-topic.

More specifically, although they are all related to COVID, a closer look brings that there are some light differences between the common words of 2020 and 2023. In 2020, in addition to "COVID", the more common words in these locations in tweets were "death", "growth", and "lockdown". But in 2023, in addition to "COVID", the more common words in tweets become "vaccine" and "health". From the changes in common words, we can predict that people's attitude toward COVID-19 in 2020 was extremely negative, but now in 2023, due to the development of vaccines, people have become less anxious about the COVID.

5.5 Hashtag Analysis

In this analysis, we are trying to investigate the number of hashtags per tweet and the most common hashtag. The number of hashtags per tweet is a simple way to determine whether or not the tweets we have collected are analytically valuable. Usually, tweets should have one to two hashtags to make the content of the tweet more specific.

Similar to the previous analysis, we plotted the number of hashtags per tweet and the common hashtags in 2020 and 2023 respectively.

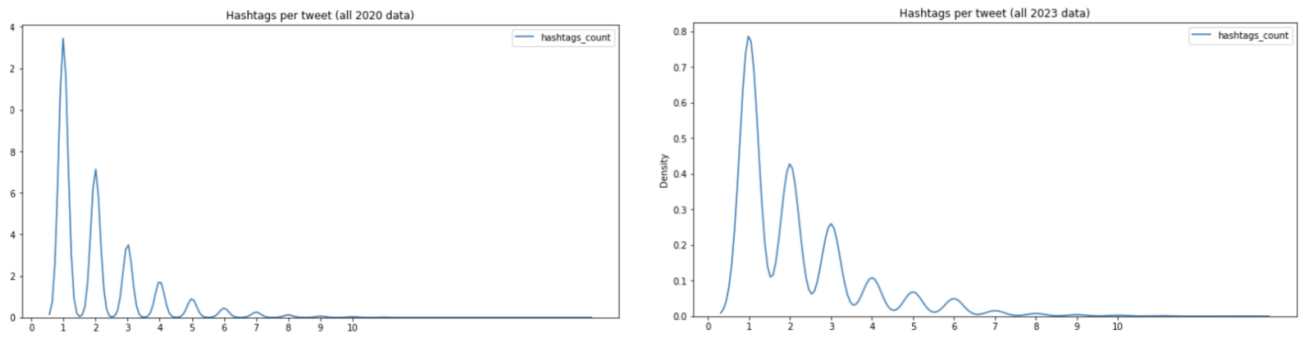

Figure 10: Distribution of the hashtag count in 2020 and 2023.

The plot of the `hashtags_count` distribution shows us the frequency of the number of hashtags used in a tweet. And the Word Clouds below express the most frequent words in hashtags.

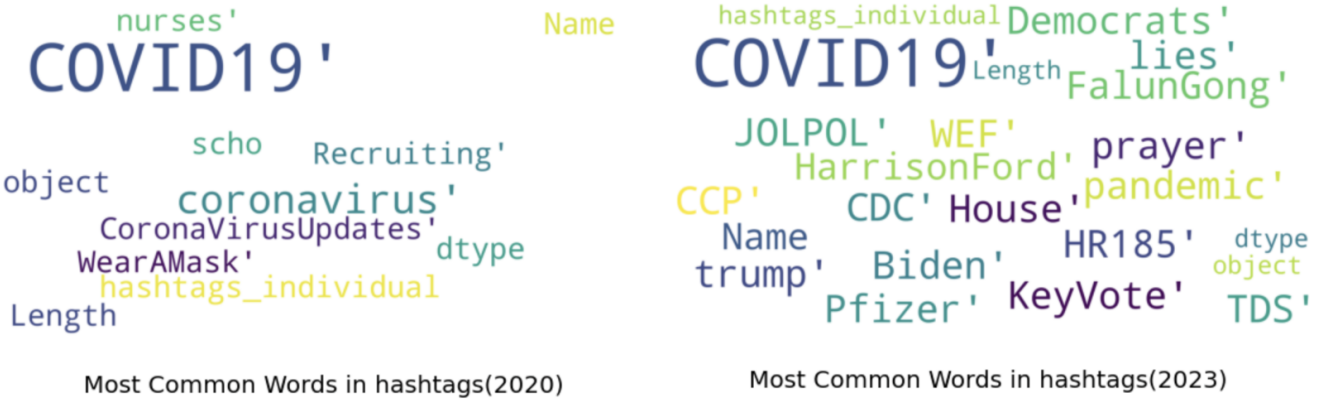

Figure 11: Word cloud of the most common hashtags in tweets in 2020 and 2023.

These plots exhibit that most of the tweets we collected in 2020 and 2023 contained one or two hashtags, and the most frequent hashtag was “covid19”, proving once again that our datasets are analyzable.

5.6 Verified Account Analysis

We are curious to see how many tweets about COVID-19 are coming from credible sources. From the pie charts below, we could observe that in 2020, only 16.1% of the accounts are verified, and even lower percentage, 13.1% in 2023. The majority of accounts are not verified. Non-verified accounts tend to post more biased tweets since they could pretend to be an ”anonymous stranger” online.

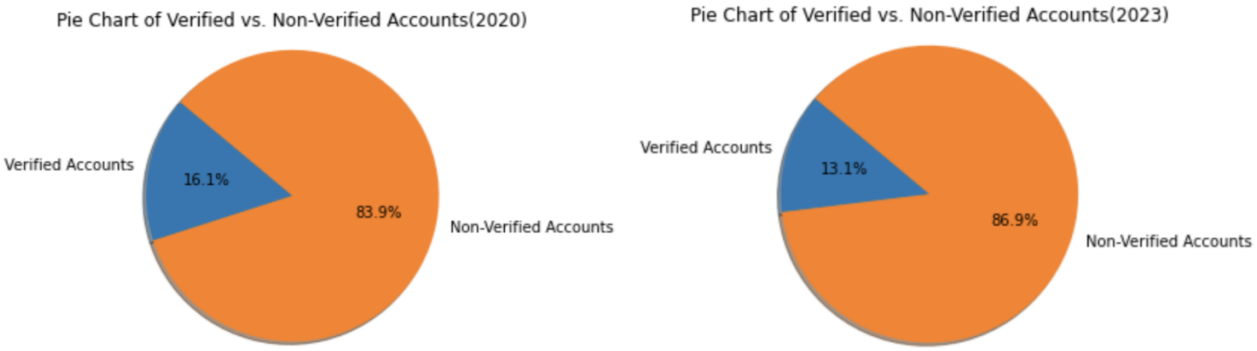

Figure 12: Pie chart of the distribution of verified accounts.

5.7 Scatter plot to see the relationship between the number of followers and the number of friends

We created scatter plots to check if there is a relationship between the number of friends and the number of followers. By letting the x-axis represent the number of followers and the y-axis represent the number of friends, we could see that there does not seem to have a relationship between the two variables for either 2020 or 2023.

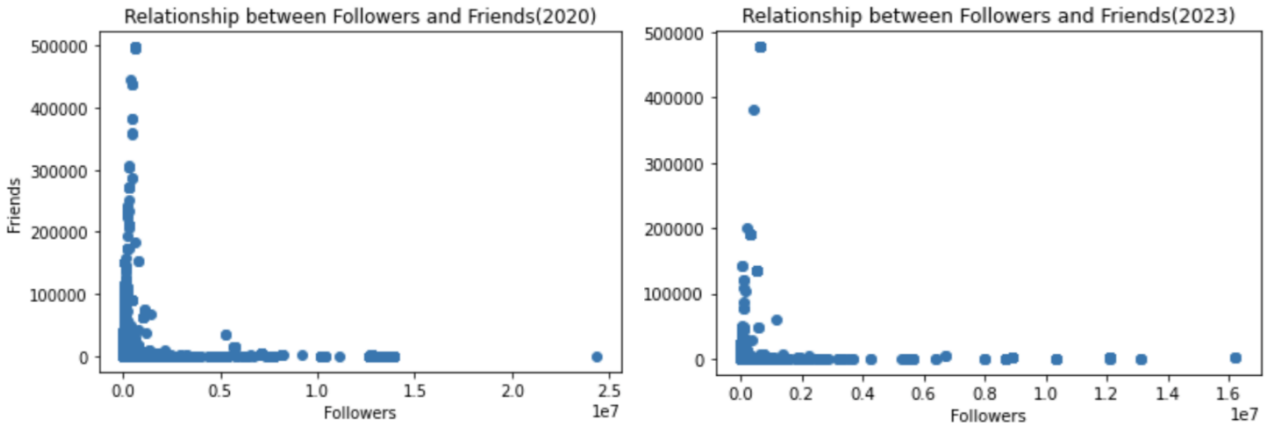

Figure 13: Scatter plot of follower count and friends count.

From checking the timeline posted by the Centers for Disease Control and Prevention (CDC), we know that on July 16th, 2020, ”many states, including California, Michigan, and Indiana postpone re-opening plans as COVID-19 case numbers rise.” On that day, the U.S. reports a record of 75,600 new cases. On July 22nd, 2020, ”the Department of Defense (DOD) and HHS reach a deal with Pfizer BioNTech for the delivery and distribution of 100 million doses of the Pfizer BioNTech COVID-19 vaccine candidate in December 2020, upon confirmation that the vaccine is safe and effective.” On July 23rd, 2020, ”CDC releases resources for school administrators, teachers, parents, guardians, and caregivers to help build appropriate public health strategies to slow the spread of COVID-19 in a school environment.” There are no special events or festivals on July 25th, 2020.

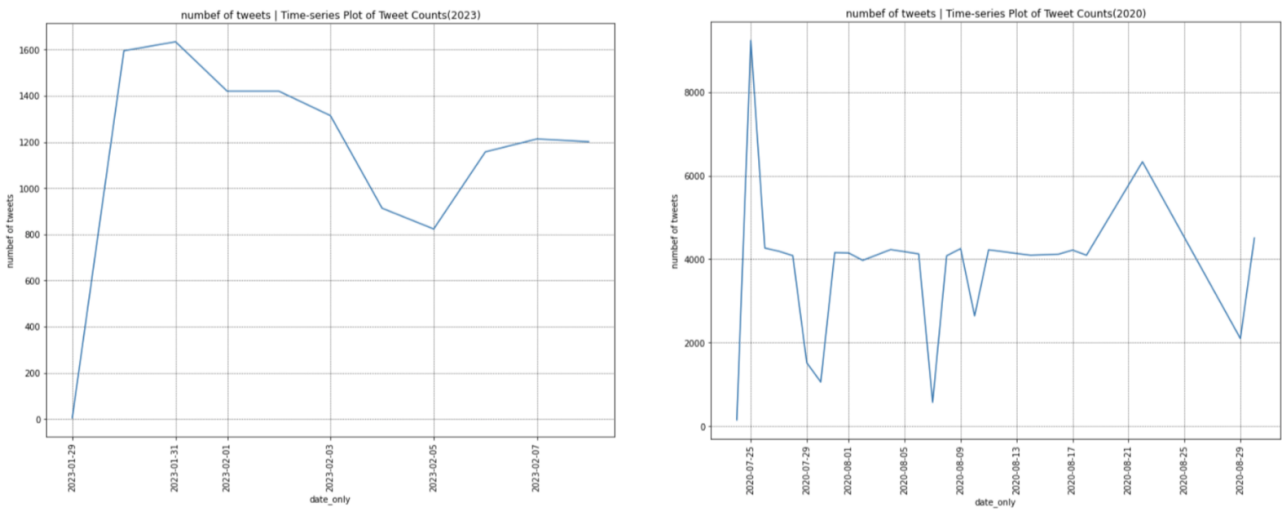

Figure 14: Plot of the number of tweets over time in 2020 and 2023.

On 2020-8-19, ”after CDC studies show that American Indians and Alaska Natives are among the racial and ethnic minority group at higher risk for severe COVID-19 outcomes, CDC provides more than 200 million in COVID-19 funding to Indian Country.” On 2020-8-22, ”a study published by the Journal of the American Medical Association calls into question the clinical benefits of the anti-viral drug Remdesivir being used to treat patients hospitalized with COVID-19.” These events could be the reason for the second peak.

Because we only have ten days of records, the plot does not show any obvious peak. It only shows a lower point on 2023-2-5, but there does not seem to have any big event on that day globally.

6 Summary and Future Work

In this module, we use Kaggle’s 2020 data and 2023 data obtained by ourselves to clean, analyze and compare them to help us conduct sentiment analysis later. We have repeatedly demonstrated the availability of our data. In the process, we discover that the most frequently tweeted word in the 2020 and 2023 data sets was ”COVID.” This indicates a very high level of user discussion on COVID-19 over the past three years, i.e. our 2020 and 2023 data sets are both COVID related, so this positive signal ensures that our future pandemic sentiment analysis stays on topic.

Beyond that, we also found some subtle differences between common words used in 2020 and 2023. In 2020, other than ”COVID-19,” the more common words in tweets for those positions were ”death,” ”growth,” and ”lockdown.” But by 2023, in addition to ”COVID-19,” the more common words in tweets had become ”vaccine” and ”health.” People are moving from fearing COVID-19 to fighting it with a vaccine. As attitudes improved and death rates dropped with the advent of vaccines, the topics and keywords of discussion changed. Later, we also tried to investigate the number of hashtags per tweet and the most common hashtags. We speculate that tweets with one or two hashtags will be more specific in their content. At the same time, we also noticed that about 13 percent of users were unwilling to authenticate their Twitter account in either 2020 or 2023. Finally, we also identified the peaks and troughs of COVID tweets in July and August 2020 and mid-February 2023, respectively, and we used the news at that time to guess the reasons for the peaks and troughs in comments.

While there are some limitations to our data, such as differences in user location and time zone, and the fact that Twitter does not cover all users and regions, ignoring these factors allows us to gain insight into public perceptions of COVID-19 and its evolution over time, and enable policymakers to explore this information in meaningful ways.

In the next step, we hope to perform sentiment analysis on COVID-19 tweet data set based on our exploration, and select and apply appropriate models and methods to achieve the goal of sentiment analysis. In addition, we will select the best model based on performance and validation metrics, and explain why the models were chosen and their limitations. Finally, we will adjust the model to optimize its performance as well as evaluate the performance of the model and measure the results using appropriate metrics.