COMP 598 Final Project
Data Science for COVID

Esteban Mendez, Yu Rong, Hanying Shao
McGill University
esteban.mendez2@mail.mcgill.ca, yu.rong@mail.mcgill.ca, hanying.shao@mail.mcgill.ca

Introduction

The coronavirus pandemic has spread to almost every country in the world. As governments work hard to adopt new policies to deal with the spread of the virus, its spread has still influenced people's daily life. Therefore, many people are highly concerned about the latest news about COVID-19, participate in related topics discussed on social media and wonder what the recovery will look like.

The project's primary goal is to understand the discussion currently around COVID in social media and determine the salient topics. We focused on the discussion on Twitter, one of the most popular social media with 211 million daily active users (Jill 2021). Briefly, we collected 1000 tweets within a 3-day window, categorized the tweets into five topics and conducted some analysis based on the contents. In our findings, we discovered the salient topics discussed around COVID are focusing on "Opinion" and "Vaccinate," and there have been more discussions about "Omicron" recently. We also found there are more negative posts to the pandemic and vaccination. Therefore, we reasonably speculate that people show more pessimism about pandemics based on the results. At the same time, the data we collected might be biased and influence the conclusion.

Data Collection

The dataset used for the project is collected from Twitter using the Twitter API. We utilize the library "tweepy" that manages the connection to the API given the "bearer_token" and query information. For this, we wanted to keep the search as broad as possible and reduce any potential bias. We first query for covid, in which we got multiple covid related tweets, from vaccination to the new variant. To ensure the language was tagged as English and the data was not a retweet, we included "lang:en" and "-is:retweet" as part of the query. Then, we collected 1000 tweets each day from Nov. 29th to Dec. 1st in 2021, giving us in total 3000 tweets of raw data. After that, we randomly sampled 1000 of the 3000 tweets containing more than ten words to reduce the bias based on date and ensure the tweets contained valuable information. In the end, we have 1000 posts in a range of three days mentioning all subjects concerning the current COVID situation as the raw data. After all this, we save the data as a TSV file for easy annotation, in which the data were then cleaned by eliminating the invalid and duplicate tweets.

To better understand the data we had collected, a word cloud was generated to visualize the contents of the data.

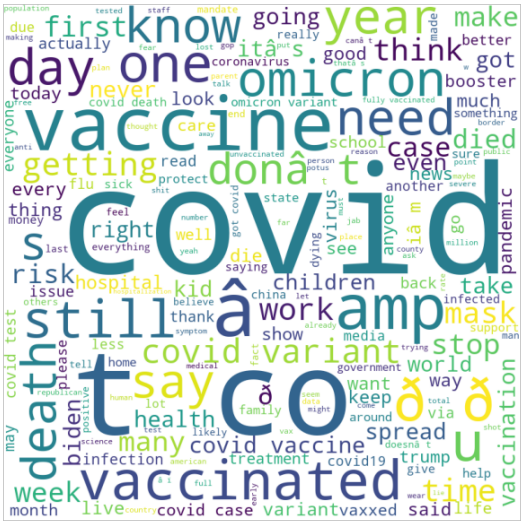

Figure1: Word Cloud plot for all collected tweets.

As shown in Figure 1, we can see that the most frequent words in posts are COVID, Vaccine, Omicron, Deaths, etc. This picture can help us better understand the content of tweets and inspire us how to classify topics later.

Methods

This section introduces the design decisions we made that might impact the results.

Cleaning Data

In cleaning up tweets, we keep the account name followed by "@" and the link in the tweet's content. The reason is that we found that this information is sometimes helpful for our annotation. For instance, the contents of some tweets seemed irrelevant with COVID. However, if we enter the link of tweets, the link usually will redirect us to a page with a report or news related to COVID. In this case, we still regard such tweets as related to COVID topics and include them as part of the dataset. In order to interpret the meaning of tweets correctly, we keep the link and account information for now and remove them when finishing annotation.

Moreover, we found that there are some tweets with highly similar content. For instance, somebody posts the same content twice with one more word the second time. Technically, it will not be regarded as duplicated tweets since they are indeed different. However, such tweets cannot bring us more useful information about popular topics. Therefore, we decided to remove them by marking them differently during annotation and extracting them later. After removing, we fill the datasets with some new tweets randomly selected in the rest of the raw data stored before.

Annotation Process

As we discussed during the lecture of this course, there is no absolute objectivity in coding (i.e., the process of annotation). As coders, we might be biased towards finding certain information and categorize tweets based on our understanding.

We conducted an open coding on randomly selected 200 tweets and developed the topics as "Omicron," "Vaccinate," "News," "Influence," and "Opinion." The details of categorization are illustrated in the next section. As we need only the single annotation approach for this project, the topic annotations were given priority, precisely in the previously mentioned order. In other words, if a tweet can be categorized as both the "Omicron" and "Vaccine," it will be annotated as the "Omicron" topic in our case. Setting priority leads the annotation process easier and more precise and minimizes confusion simultaneously.

To minimize subjectivity further, we split the dataset into three sub-datasets, and each of our team members is responsible for annotating one of them. A sentiment tag is also given to each tweet during the annotation process to reduce the repetitive works. If somebody is puzzled by categorizing specific tweets, they can provide their annotation and mark "?" in an extra column. After that, the other two members will check for such tweets and annotate them. If all the members annotate the tweet differently, we will annotate after careful discussion with unanimous agreement.

Modification of STOP-word list

To characterize the topics, we need to compute the top ten words in each category with the highest TF-IDF scores. For the stop-words list, we modified the stop-words list provided in Assignment 8 to fit our dataset better. Then we removed some words from the lists such as "work" and added some new terms such as "Covid," "vaccination," "dose," "via," etc. The modification allows the words with the highest TF-IDF scores to bring more useful information about the salient topics around COVID.

Result

In this section, we would like to share the findings of the project, including the selection and definitions of the topic, topic characterization, topic engagement and then visualizing the result using corresponding graphs.

Topics Category and Definitions

After removing the invalid data, we randomly selected 200 tweets from 1000 tweets. Then we conducted open coding on the 200 tweets to develop our topics. By manually looking over the contents of tweets, we found that tweets can be roughly categorized as third-party tweets and personal tweets. The third-party tweets are mainly about summarizing the latest news or report with one or two sentences, usually written in formal and objective words. We denoted such tweets as part of the "News" category to differentiate them from other tweets. If personal tweets are concerned about the news, the tweets are also categorized as "News".

For personal tweets, the contents cover many different topics. To capture the trend of the salient topic, we selected three keywords from the contents that can best describe each tweet and added them to the separate column. Then we count the frequency of each keyword. According to the counting results, there are about 30 tweets that mentioned the new variant of coronavirus, including Omicron. Hence, we set "Omicron" as a separate category. Meanwhile, regarding vaccine hesitancy, we separate "Vaccination" from other topics as a single category, and there are about 45 tweets mentioning vaccination. For the rest of the tweets, the contents mentioned various topics. Hence, we roughly divide them into two categories, "Influence" and "Opinion." The tweets talking about things that happen in real life are categorized as "Influence," including their own experience or

things that happened around them during COVID-period. The "Opinion" tweets are about personal argument and opinion of COVID, which includes future prediction and emotional comments. Most of the things mentioned in "Opinion" have not really happened yet. These topics are broad enough to capture most of the posts gathered and provide useful information around COVID to be analyzed later.

These are the five selected topics with the corresponding definitions.

- "**Omicron**": Tweets are categorized as "O" if they are concerned about the new variant of COVID, including Omicron.
- "**Vaccine**": Tweets are categorized as "V" if they are concerned about anything related to vaccination, including vaccine passport, new policies, boosters, etc.
- "**News**": Tweets are categorized as "N" if the contents are about the latest policies, news, or reports.
- "**Influence**": Tweets are categorized as "I" if the people talk about how COVID influences their daily life and their experience related to COVID.
- "**Opinions**": Tweets are categorized as "P" if people generally post their opinions regarding COVID or related topics. It also includes emotional comments.

Then we followed the annotation process as we described in the Method section. A sentiment tag is also given to each post during the annotation process.

Topic Characterization with TF-IDF

To characterize our topics and better understand the concerns of each topic, we compute the top 10 words in each category with the highest TF-IDF scores and form a table. We use all 1000 annotated posts to calculate the inverse document frequency. Besides, we modified the STOP-words list as described in the method section.

	Omicron	Vaccine	News	Influences	Opinion
1	detected	mandate	count	mom	fake
2	sheba	prevent	Tuesday	wife	created
3	jerusalem	mandatory	confirmed	worry	effect
4	confirmed	decision	source	close	bunch
5	plans	received	merck	collection	scared
6	southern	tired	insights	fairly	stop
7	stock	chance	analytics	sharing	survival
8	market	mrna	usafacts	husband	question
9	strain	dies	minister	hit	harm
10	concerns	judges	air	recently	freedom

Table 1: Pie chart of topics related to COVID

From the above table, it is easy to observe that the words with high TF-IDF scores are highly correlated to the topic, proving that the annotations and methods we used are adequate. For instance, we categorized tweets as "Influences" topics if people mention the things around them or describe how their daily lives are influenced during the COVID-period. As we can see, words such as "mom," "wife," "husband," "worry" appear on the top 10 TF-IDF scores words list in the "Influence" category. It proves the accuracy of our annotation to a certain extent.

Topic Characterization and Engagement

After the annotation process, some pie plots were generated to visualize the topics and sentiments' distribution in the tweets we collected.

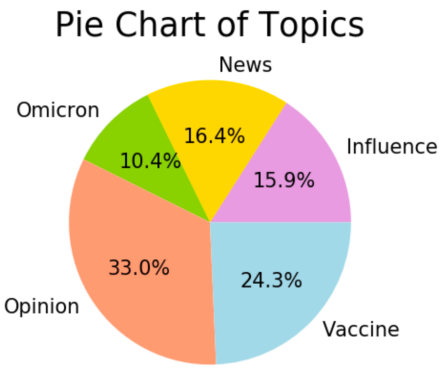

Figure 2: Pie chart of topics for collected tweets

Figure 2 illustrates the proportions for different topics in all collected tweets. It can be seen that the most common topic is "Opinion," which accounts for 33% of all tweets. The second popular topic is "Vaccine," which accounted for 24.3% of total data. This observation means most tweets we collected from Twitter are users' personal opinions on the current COVID situation. The proportion of "News" and "Influence" is quite close, accounting for 16.4% and 15.9%, respectively. The least mentioned topic is "Omicron," which accounted for 10.4% of all tweets. It is reasonable that the number of discussions about "Omicron" is relatively small since it is a new topic.

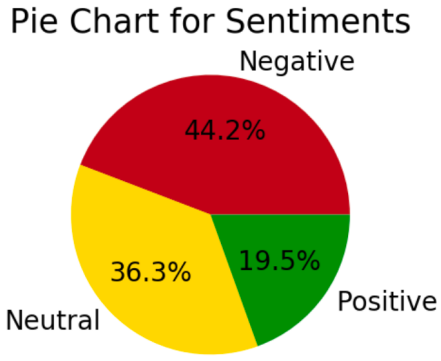

Figure 3: Pie chart of sentiments for collected tweets

Figure 3 shows the overall sentiment tendency of all the collected tweets. It seems that negative tweets are the most common, accounting for 40% of total tweets. There are 36.3% neutral tweets, and only 19.5% of the tweets are positive. In order to know more about the proportion of different sentiments tweets inside each topic, we plotted the counts of tweets by grouping by topic and sentiments as a bar plot.

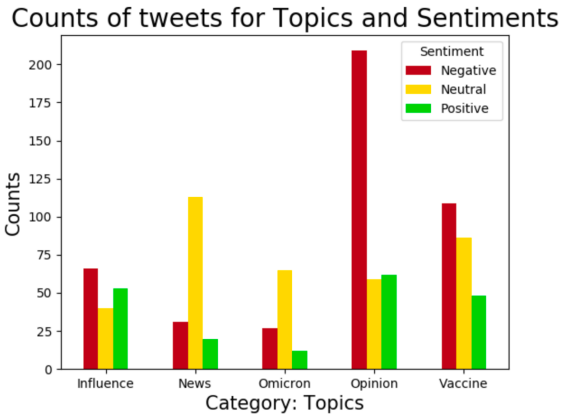

Figure 4: Pie chart of topics related to COVID

Figure 4 shows the counts of different topics and the sentiment tweets number inside each topic. From Figure 2, we can find that more than 200 tweets are categorized as "Negative Opinion," which dominates other sub-categories. Besides, less than 25 tweets are classified as "Positive Omicron." Given that the number of tweets for "Omicron" is the smallest among the other topics, the result is reasonable.

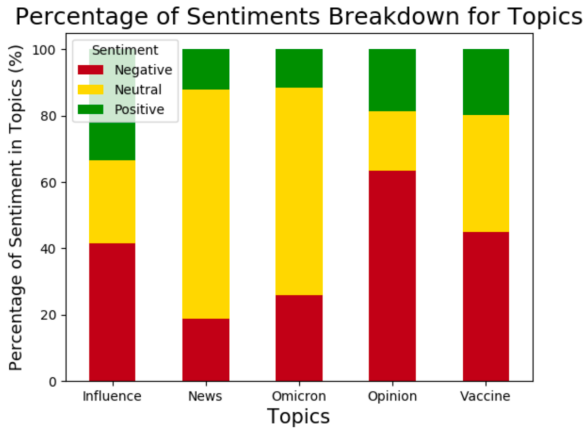

Figure 5: Stacked bar chart for different topics.

Figure 5 is the stacked bar chart for different topics. It allows us to understand the proportion of each sentiment inside various topics, given that each topic has a different number of tweets. We found that most posts regarding "Omicron" or "News" are neutral from the plot, which is reasonable because news is predominantly neutral. Except for the tweets from the "Omicron" and "News" categories, the negative sentiment is dominant in all other categories, suggesting that the general population's reaction to the current COVID situation is negative. As we went through the tweets for the annotations, we also observed a significant number of serious doubts about the vaccines, the statistics of COVID cases reported on the News, and the government's policies concerning the pandemic. In other words, there are a considerable number of people expressing negative sentiments about COVID on Twitter, according to the results we acquired from the project.

Discussion

"Influence" of COVID

The global outbreak of COVID-19 has led the governments to propose different policies, such as lockdowns, in an attempt to minimize the spread of the virus (Charlie Roberts 2020). It dramatically influences people's normal daily life and routines. In this case, many people have to change their lifestyle, spend more time at home, and live with their family members. Of the respondents, 58% indicated they lived with the family during the lockdown (Charlie Roberts 2020). Hence, it is not surprising that words such as "mon," "wife," and "husband" appear on the top 10 TF-IDF scores in the category "Influence." In other words, when people talk about the influence brought by COVID, they are highly likely to mention their family members and things happening around them. Moreover, the distribution of sentiments for "Influence" seems to be the most evenly compared with other categories. The positive tweets have the highest proportion within the "Influence" category. For example, one of the "Positive Influence" tweets from our dataset are that "I work at home since COVID. now I have more time to do the things I like and spend time with my family." It suggests that people would like to share their daily life no matter good or bad things happen.

Concerned with Vaccine Hesitancy

According to the word frequency and top 10 TF-IDF score of the word list in "Vaccine," words such as "prevent," "dies," "mrna," and "mandatory" are strong evidence to show the correlation of contents and our topic. It can seem as an evidence that verified the reliability of our annotation.

For the "Vaccine" topic, there are about 45% of tweets are negative, 36% of the tweets are neutral, and 19% of the tweets are positive. Most people can objectively regard the principle, effects, and side effects of vaccines. These people are willing to get a vaccine. People who post negative tweets online about vaccines are more likely to show vaccine hesitancy. Their main concerns focus on the vaccine's effectiveness and safety.

The emergence of Omicron and Related Topics

Considering the recent emergence of Omicron, the 10% proportion of tweets related to the topic "Omicron" reflect its trend and people's concern about the new variant of coronavirus. It is interesting that "Tuesday" appears as the word with the second top TF-IDF score on the "News" category. The tweets were collected from Nov. 29th to Dec.1st in terms of day in a week, Monday, Tuesday, and Wednesday. Therefore, it indicates something significant happened on Tuesday, which could be a piece of breaking news or the latest report. Considering that, we tried to find a reasonable explanation for that and began to search the latest news online. According to NCIRD 2021, the United States designated Omicron as a Variant of Concern on Nov. 30th (i.e., Tuesday), and on Dec.1st, the first confirmed U.S. case of Omicron was identified. It explains the appearance of "Tuesday" on the top TF-IDF score list on the "News" category and the relatively large proportion of tweets mentioned "Omicron." From another perspective, it shows that people are highly concerned about COVID-related news, especially the latest report related to the new variant. In addition, the information also proves the credibility of our data collection and the accuracy of annotation.

Analysis of the attitude of People toward Pandemic

According to our annotation priority order, we set "Omicron" as the highest priority and "Opinion" as the least priority. In this case, the result shows that "Opinion" is the most common topic, indicating that people tend to express their general opinions or future predictions about "COVID" instead of focusing on a specific aspect. Meanwhile, more than half of the opinions are negative. Moreover, about 44.2% of the collected tweets are negative. Based on the results, we reasonably guess people show more pessimism about pandemics and COVID-related topics.

Some reports show a similar opinion with our hypothesis. According to the report from Brandwatch and Ditch the Label, in 2021, there was an increase of 22% of violent threats online during the COVID period, specifically in ethnicity-based and racist hate speech. Such comments would be categorized under this report's "Opinion" topic. Therefore, the findings in the report are consistent with our result, with the fact of "Opinion" having the highest number of negative tweets out of all categories by a significant margin. It usually happens in conversations or debates between people from both sides of the political spectrum. For future projects, we could take more effort to look more into, to clarify the origin and the characteristics of the negative comments.

Potential Problems

Some potential problems might influence the conclusion that we found in the project, and we would like to discuss them here. For data collection, even if we collected 1,000 tweets every day from Twitter, we still could not contain all the tweets. The dataset depends on the collect order and the API's limitation. Usually, the time range of data collected from API cannot cover the whole day and thus, it isn't very objective towards a specific period.

Compared with topics, categorization of the sentiment seemed to be more subjective during the annotation process. It is more difficult to differentiate the semantic category of a tweet than distinguish the topic category without a clear definition and standard. As non-native speakers (i.e., English as a second language), it is hard for us to distinguish ironic sentences from positive sentences. For example, one of the tweets is, "*We cannot worry anything except covid since it is too important now. Trust your government. They know what's best.*" It seems a positive tweet at first glance. However, if you click on the corresponding link with the tweets, it will redirect to a recent crime report news page. Considered that, the tweet should be tagged with negative sentiment. Hence, the judgmental bias might lead to mistakes in the categorization during annotation and thus influence the category distribution.

Group Member Contributions

Esteban Mendez

Contributed to design decisions, wrote the data collection and preprocessing script, participated in the definition of topics, annotated tweets, wrote the computing TF-IDF script, cleaned code on GitHub, participated in writing the data collection and discussion section and participated in report revisions.

Yu Rong

Contributed to design decisions, developed topics by conducting open coding on 200 tweets, annotated tweets, modified plots and created tables, wrote most of the report and report revisions.

Hanying Shao

Contributed to design decisions, participated in the definition of topics, annotated tweets, analyzed data, wrote the script for plots on the notebook, participated in writing the data collection and the result section and participated in report revisions.

References

Brandwatch and Ditch the Label. 2021. “Uncovered: Online Hate Speech in the Covid Era” https://www.ditchthelabel.org/research-papers/hate-speech-report-2021/

Charlie Roberts, Nicholas Gill, Stacy Sims. 2020. The Infleunce of COVID-19 LockDown. Frotiers in Nutrition https://doi.org/ 10/3389 /fnut. 2020.589737

Jill Goldsmith. 2021 “Twitter Hits 211 Million Daily Active Users IN Q3; U.S.Ad Sales Jump 51%, Profits hit By Litigation Charge.” https://deadline.com/2021/10/twitter-jack-dorsey-1234862906/

NCIRD (National Center for Immunization and Respiratory Diseases). 2021. “Omicron Variant: What You Need to Know” https://www.cdc.gov/coronavirus/2019-ncov/variants/omicron-varant.html