COMP 598: Data Science Final Project
COVID and Vaccination Hesitancy as Viewed by Social Media

Wenjia Liu, Sophearah Suy-Puth, Ran Lu

McGill University

wenjia.liu@mail.mcgill.ca, sophearah.suy-puth@mail.mcgill.ca, ranlu@mail.mcgill.ca

Introduction

This is the COMP 598 Topics in Computer Science 1 final project report from McGill University. Using Twitter data, the study attempts to measure and comprehend public sentiment over the COVID-19 outbreak. COVID-19 is an infectious disease. The disease has spread to all continents and was declared a pandemic by WHO (World Health Organization) on March 11, 2020. It not only causes symptoms and complications, but it also leaves our emotional and mental health at risk. Our lifestyle has changed dramatically because of the epidemic. The public has different viewpoints on the policies and vaccinations that have been introduced. Twitter is one of the world's most popular social media networks. On the site, a vast number of individuals voice their opinions and thoughts, expressing popular opinion on pandemic-related issues.

This report is an analysis of the current discussions surrounding COVID-19 on social media. It mainly includes the procedure to collect data from Twitter, prominent themes discussed around COVID-19, and data analysis on the focus of each topic and people's sentiments toward the pandemic and vaccination.

Next, we will go through our dataset, methods, and observed results in detail below, before discussing our findings and conclusions.

Data

We used the tweepy library in Python to search for tweets through Twitter's official standard v1.1 API. We collected 1300 tweets within a 3-day window to account for some tweets being unrelated to COVID or not in English although we have set filters appropriately. The filter we used are as follows:

- language is "en"
- contains case-insensitively at least one keyword in "COVID", "vaccination", "pfizer", "moderna", "johnson&johnson", or "astrazeneca"
- is not a retweet

First, we set the language to English. Second, we chose the set of keywords above since they are either directly related to the topic of discussion ("COVID", "vaccination") or a brand name of a COVID vaccine ("pfizer", "moderna", "johnson&johnson", "astrazeneca"). Third, we filter out retweets so that among an original tweet and its retweets, only the original tweet will be kept, thus only collecting unique tweets. We allow replies and quoted tweets since they contain new information written by the user replying or quoting.

We saved search results to a JSON file in the original format returned by the Twitter API. Then, we ran another Python script to validate the following and discard invalid tweets:

- every pair of tweets have different content
- every tweet has language set to "en"
- no tweet is a retweet

The script then writes valid tweets to another JSON file containing the following fields:

- row (index among all collected tweets)
- id (tweet ID)
- created_at (creation timestamp)
- full_text (full content; for replies and quoted tweets, text of the original tweet is appended to aid annotation)
- category (for annotation)
- sentiment (for annotation)

Finally, we manually checked each processed tweet and eliminated tweets unrelated to COVID or vaccination, and tweets that have language set to "en" but is in fact in another language). This resulted in 1014 valid tweets from which we sampled 1000 for annotation.

Methods

Before sentiment analysis, we manually annotated and preprocessed the dataset to include a category column that will indicate the topic associated with each tweet. The themes are **1. political, 2. health, 3. vaccination, 4. variant, 5. social**. We decided the categories after open coding on 200 tweets and discussion among the team members. Specific reasons for choosing the categories will be discussed below. Then, we manually annotated the data set with the category and the emotion of each tweet: **positive, neutral,** or **negative**. Before any further analysis, each group member cross validated the manual annotations.

First, the COVID-19 epidemic has caused tremendous changes in people's perceptions. So, we choose a category that is **social**. It includes the impact of people on travel, study, work, and life after COVID-19. Security needs to be given unprecedented attention. Facing the possible harm caused by the spread of COVID-19 has caused people to think about themselves. In the face of COVID-19, people's physical needs such as food, clothing, shelter, transportation, and safety needs such as preventing epidemics are highly valued, while social needs, respect needs, and self-realization needs are temporarily put aside. These changes in demand are quietly changing people's work and lifestyle and triggering people's repositioning and thinking about life and their own value.

Second, we chose another category **political**. These include each country's policy on public protection and the policy formulation of each country's entry and exit management. People who travel will need to meet specific and evolving travel restrictions. People will discuss more about them on Twitter due to the concerns about those policies.

Third, we chose another category **health**. It contains the impact of COVID-19 on health. Although medical professionals have better treatments for the current new coronavirus infection, its long-term health effects are still uncertain. Therefore, people's panic about COVID-19 continues to rise. There are many opinions that COVID-19 will cause permanent lung damage, as well as hearing the heart, vascular system, and permanent nervous system damage. Including in the incubation period of the new coronavirus in the early stage, people may experience headaches and fever, poor taste, dizziness, and coughing and other cold-like symptoms are also hot topics that have caused widespread discussion.

Fourth, the category we choose is **vaccination**. As of November 15, 2021, WHO has evaluated the following COVID-19 vaccines that meet the necessary safety and effectiveness standards: AstraZeneca/Oxford Vaccine/Jansen Vaccine/Modena Vaccine/Pfizer-Biotech Vaccine/Sinopharm Vaccine/Kexing Vaccine/COVAXIN. Therefore, people's hot topics include but are not limited to topics related to vaccines. And we need to report the reflection about vaccination.

Fifth, the category we chose is **variants**. Because we think health, vaccination, and variants have different focus, we decided to divide them into three categories. With the gradual development of the COVID-19 era, variants of different viruses have received strong attention from the public. Before the emergence of new vaccines, the world will continue to be affected by the COVID-19 variants. The following are the current main variants: 1. Delta, 2. Lambda, 3. Mu, 4. Omicron. Because the virus keeps mutation, and the vaccine cannot catch up with the speed of virus mutation in time, the topic of mutation is also eagerly discussed.

We made sure that any duplicate tweets and unconfirmed tweets are removed before we analyze the data. Tweets with duplicate text content, as well as tweets containing English and another language, are deleted. We also eliminated any hyperlinks, user references, and URLs from the post content because they had nothing to do with our research.

Results

Results Overview

After annotating 1000 tweets with categories and sentiments, we ran several analysis scripts in Python to produce the following JSON files:

- word counts per category (removed punctuation and stop words)
- top 10 word with their tf-idf scores per category
- sentiment counts per category

The formula for calculating the tf-idf score of a word is:

$$
\begin{aligned}
& tf - idf(w, category, script) \\
& = \ tf(w, category) * idf(w, script) \\
& tf(w, category) \\
& = \ the\ number\ of\ times\ category\ includes\ the\ word\ w \\
& idf(w, script) \\
& = \ log\ [\ (total\ number\ of\ categories) \\
& /\ (number\ of\ categories\ that\ use\ the\ word\ w)\ ]
\end{aligned}
$$

The top 10 word with their tf-idf scores per category (case-insensitive) is summarized in the table below.

vaccination	**social**	**health**	**political**	**variants**
pfizer	pic	analytics	johnson	botswana
booster	concert	brain	ban	mutations
sore	park	conditions	nhs	delta
moderna	thanksgiving	passed	borders	detects
az	retweet	pain	macron	mutating
vaccinating	allah	hiv	tory	discovered
unvaxd	attending	nhill	border	investigating
arm	white	vitamin	mps	heavily
anti	situation	maybe	temporarily	spike
efficacy	event	ivermectin	minister	version

Table 1: Top 10 words by tf-idf score in each category

Then we used the matplotlib library to produce bar plots of the 10 words in each category with the highest tf-idf score, as well as the proportions of sentiments per category as a stacked bar plot.

Top Words by TF-IDF Score

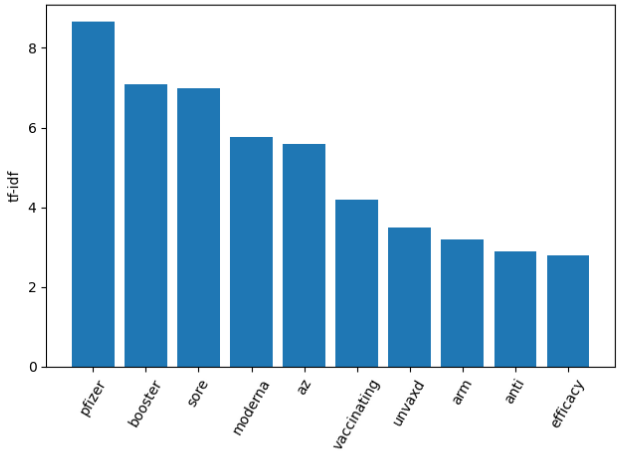

Figure 1: Top 10 words and tf-idf score in "vaccination"

In the **vaccination** category, the top 10 words by tf-idf score includes certain brands of COVID vaccines ("pfizer", "moderna", "az" short for AstraZeneca), the vaccine policies ("booster", "efficacy"), the vaccine effects ("sore", "arm"), as well as discussions about the unvaccinated ("unvaxd" meaning unvaccinated, "anti").

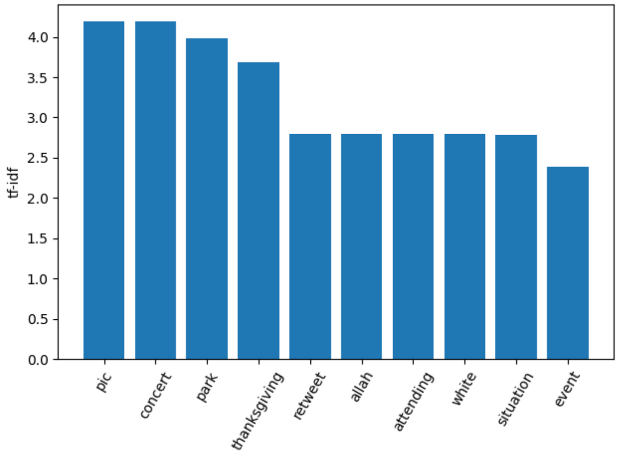

Figure 2: Top 10 words and tf-idf score in "social"

In the **social** category, the top 10 words by tf-idf score includes event discussions ("concert", "park", "thanksgiving", "attending", "event") and socializing on Twitter ("pic" short for picture, "retweet").

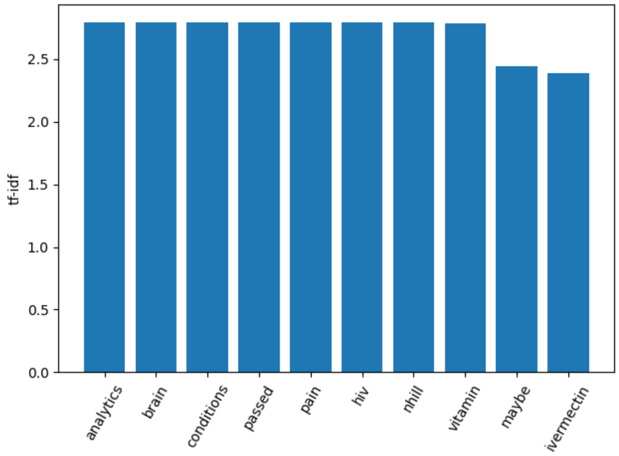

Figure 3: Top 10 words and tf-idf score in "health"

In the **health** category, the top 10 words by tf-idf score includes COVID's health effects ("brain", "conditions" "passed", "pain") and medicine ("vitamin", "ivermectin"), health studies ("analytics") and comparison with another disease ("hiv").

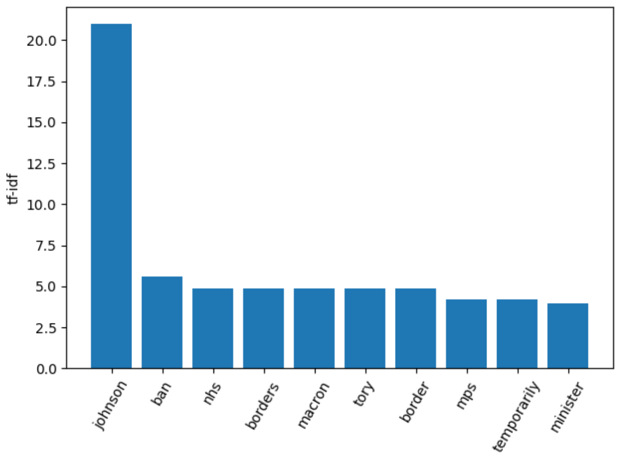

Figure 4: Top 10 words and tf-idf score in "political"

In the **political** category, the top 10 words by tf-idf score includes names of politicians and parties (most notably "johnson" with a particularly high tf-idf, also "macron", "tory" meaning the Conservative Party, "mps" short for Members of Parliament). Other topics include policies relating to COVID ("ban", "borders", "border", "temporarily") and a health system ("nhs" short for National Health Service).

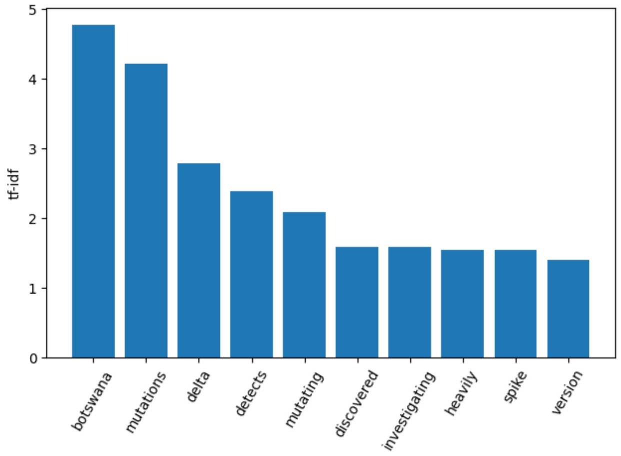

Figure 5: Top 10 words and tf-idf score in "variants"

In the **variants** category, the top 10 words by tf-idf score includes the most recent Omicron variant ("botswana", "detects", "discovered", "investigating"), the dominant delta variant ("delta"), and characteristics of the variants ("mutations", mutating", "heavily", "spike").

Sentiments by Categories

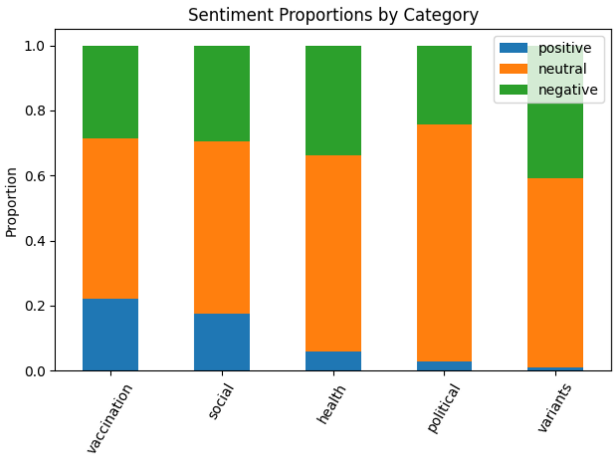

Figure 6: Sentiment proportions by category

The proportion of positive sentiment decreases in the sequence of vaccination, social, health, political, variants. People tend to see more hope and joy when talking about **vaccination** and **social** than **health**, **political**, and **variants**, which brings more uncertainty and dissatisfaction.
The proportion of negative sentiments also shows this trend as it is lower in **vaccination** and **social** and higher in **political** and **variants**. **Political** appears to be an exception since it has the lowest proportion of negative sentiment. This is because a lot of tweets in the **political** category are news and facts, which results in a considerable proportion of neutral sentiment.

Discussion

From the data collected and analyzed, we discovered a lot about the discussions surrounding COVID on Twitter. Since we split up the discussions into specific topics (social, political, health, vaccination, and variants) it was easier to discover the publics perspective on each of these different categories.

The goal of the **social** category was to see how the public thought about how COVID impacted people's lifestyles, including, but not limited to work, study, travel, and relationships. Based off the top words in this category by tf-idf, COVID greatly impacted how people are able to come together by frequency of words like "concert", "park", "thanksgiving", "attending", and "event". Additionally in this category words that indicate socialization on Twitter were frequently used like "pic", and "retweet". By sentiment, this was one of the most polarizing topics as many people were positive and negative about COVID's impact on people's lifestyle. During annotation, it's noted that many of negative reactions were because people were disappointed events are being limited or even canceled because of COVID. The positive reactions to this subject were because people are thankful about being able to see the people they are close to despite COVID as part of Thanksgiving.

The target of the **political** category was to understand the public's perspective on various policies put in place because of COVID. Based off the top words in this category by tf-idf, discussions in this category mostly surrounded names of politicians and parties (most notably “johnson”, “macron”, “tory”, “mps”) and policies related to COVID (“ban”, “borders”, “border”, “temporarily”) and the health system ‘nhs’. The sentiment was more neutral than the other categories because most of the tweets in this category were for the purpose of sharing news about COVID policies. The second most frequent sentiment in this category was negative because most people were dissatisfied about COVID related policies.

The primary function of the **health** category was to understand the public’s perspective on individual and public health. The tweets in this category are much more varied than most other categories that were analyzed. From the annotation there were tweets about stats behind how widely spread COVID was around different areas “analytics”, people’s reactions on how COVID impacted them and people they are close to “passed”, “pain”, effects of COVID “brain”, “conditions”, “pain”, or other theories how people thought that helps COVID “vitamin”, “ivermectin”. There were far more negative tweets surrounding health than positive because most people went on twitter to complain about the negative effects of COVID had on them. The few positive tweets were about successful recoveries.

The goal of the **variants** category was to understand the public’s perspective on the different COVID variants. Based on the top 10 words by tf-idf score, vast majority of the conversation on twitter had to do with news about the specific variants like “Botswana”, “delta” and finding more information about it “discovered”, “investigating” and characteristics of these variants “mutating”, “heavily”, “spike”. The sentiment of the variants category was overwhelmingly negative compared to positive because most people showed concern over these new variants or neutral which was simply reporting news about these new variants.

The main objective of the **vaccination** category was to understand the public’s perspective on vaccination, specifically how positive or negative the response to the vaccination has been. Based on the top 10 words by tf-idf score, most of the conversation revolved around specific brands of the vaccine “pfizer”, “moderna”, “az”, vaccine policies “booster”, “efficacy”, the effects “sore”, “arm” and the unvaccinated, “unvaxd” and “anti”. Based on the sentiment, it appears as though the public is more negative than positive towards the vaccine, but instead it is just skewed negative because even when people take the vaccination they complain about the effects ‘sore’. It is more likely that people’s views of vaccination are very split between positive and negative towards the vaccine, with slightly more negative sentiments than positive. Much of the neutral response to vaccination were simply news and facts about how the various brands of vaccines are doing.

Relative engagement with each of these topics can be further discussed by calculating the average tf-idf values of the top 10 words in each category by tf-idf. These values are social with average of 3.24, vaccination with average of 5.07, political with average of 6.34, health with average of 2.72, and variants with average of 2.39. Based off these numbers, the political and vaccination categories had highest levels of engagement followed by social, health, and variants. This means, the most frequent topics of interest of the tweets analyzed were about politics and vaccination.

Limitations & Future Work

In this experiment, the size of the dataset we adopted was only about 1,000. The conclusions drawn are not of statistical significance. Therefore, if we need to improve the experiment, we should increase the size of the dataset. At the same time, the data set can be filtered according to the attributes of different regions according to the proportion of the population. In this way, the dataset we get is more extensive. The size of the dataset will directly affect our judgment of emotions. In the classification process, we only distinguished five categories. But for many vague comments, it is possible that different people use different wording that we misinterpreted, although we tried to reduce such error by cross validation. Therefore, with a large dataset and analysis techniques such as machine learning, the emotional judgment may actually be more accurate and definitely quicker than manual annotation. This way we can achieve automation, since as long as a model is trained and updated in time, it can be used continuously on the type of problem and analysis. This way will greatly improve the efficiency of our annotation. In the sub-categories, we can refine the emotional objects and emotional judgments. The conclusion drawn in this way may be completer and more correct.

There are some consequences to the interpretation as a result of how this research was conducted. Based off of SproutSocial’s “Social media demographics to inform your brand’s strategy in 2021”, posted on March 9th, 2021, Twitter’s most active users demographic is very far from unbiased. According to that site, 44% of users are between the age of 30 and 49 and 68% of its userbase is male, meaning our conclusions best represent the perspectives of this type of demographic. In addition, 80% of tweets on the platform come from the top 10% most active accounts, meaning our results may be skewed towards this vocal minority. For simplicity, we only had to analyze people’s perspective via the text-based content in the twitter posts, not images. By removing images, we remove the context and losing information. During annotation it seemed to most frequently be in the form of posts in the social category, where the images are jokes or to be reacted to in some way such that the text by itself no longer made sense and had to be removed. Because Twitter is not as much of an anonymous platform, people may not feel comfortable voicing their true, but controversial opinions. The data that was collected during this study was also done very close to a major American holiday (Thanksgiving) so there are definitely more social related posts than usual as a result of people gathering together for this holiday and posting about how COVID impacted their Thanksgiving. Overall, most of the sentiment was negative, this is likely because many users use Twitter as a platform to voice their opinions and rant about what they are upset about.

Group Member Contributions

Wenjia Liu contributed to collecting tweets, processing collected tweets, and calculating and producing plots for tf-idf scores and sentiment counts from annotated tweets, as well as writing the Data and Results sections of the report.

Ran Lu contributed to writing the Introduction and Methods section of report.

Sophearah Suy-Puth contributed to annotating all tweets and writing the Discussion section of the report.

References

SproutSocial. 2021. Social media demographics to inform your brand’s strategy in 2021. https://sproutsocial.com/insights/new-social-media-demographics/. Accessed: 2021-12-12.