Analyzing Opinions Towards COVID-19 on Twitter

Virasat Singh, Amal Koodoruth, Omar Sadek
COMP 598: Introduction to Data Science, McGill University
{virasat.singh, amal.koodoruth, omar.wagih}@mail.mcgill.ca

Abstract

This paper investigates the current discussions in North American social media around the COVID-19 pandemic and, more specifically, the different sentiments held towards vaccines. We go about this by analyzing a large number of tweets over a three-day period and grouping them into various salient topics. Those topics include vaccines and treatments, variants, cases and deaths, and politics, amongst others. Additionally, the sentiment of all tweets is manually recorded for later analysis. This dataset is then utilized to investigate relative engagement with different topics as well as present a conclusion regarding how positive and negative the general response to the pandemic and vaccines has been. We find that discussions regarding vaccines appear to be fairly balanced with slightly more negatives than positives. Nevertheless, opinions regarding the COVID-19 pandemic as a whole were overwhelmingly negative, especially regarding cases, deaths, politics, and conspiracies.

Introduction

COVID-19 has been a serious topic of discussion since it was officially declared a pandemic in March 2020. With the introduction of vaccine mandates, there has been a general divide between those in support of the vaccine, and those against it. By learning which major salient topics are discussed around COVID, we can get a better understanding of public perspective on the pandemic, as well as opinions on the vaccine. Since a lot of the discussion related to the pandemic occurs through social media platforms, for this project, we focused on analyzing tweets posted on Twitter.

After analyzing a collection of tweets, our key findings can be summarized as follows. A majority of the tweets were related to either Vaccination, Deaths/Cases, Politics or Conspiracies. The response to the vaccination has been only slightly more negative than positive. That is, general opinion on the vaccine is relatively balanced. Furthermore, the response towards COVID deaths and cases is significantly more negative than positive - the positive responses tended to be tweets about virus recovery and low COVID case numbers. A large majority of the political tweets were negative comments about recent political action in relation to COVID. Finally, the numerous conspiracy theory tweets that were analyzed were mostly negative in relation to the topic that the tweet was discussing.

Data

For this report we analyzed a collection of Twitter posts (tweets). This section will cover data collection and design decisions that had to be made when filtering the tweets. Initially, we had used the Twitter API to collect 1,000 tweets from the three days between and including November 29^{th} and December 1^{st}. We collected 333 tweets from two days and 334 from one day. We set filters such that all 1500 posts mentioned COVID-related keywords, which include common words related to the pandemic, commonly-used vaccine names and vaccine manufacturer names. The full list of the keywords used is in Table 2 in the Appendix section. For each keyword, we also searched for tweets containing the hashtag (that is the concatenation of the '#' and the keyword). Additionally, we ensured that all tweets were in English. We then proceeded with our open coding process which is discussed in further detail in Method section. However, we later found that many of the collected tweets were hard to categorize; and thus, would fall into an ”other” category. We wanted to avoid having such a category as it would provide no valuable insight into the topic at hand, so we had to repeat our data collection method, but with a different approach.

On our next trial, we decided to collect 1500 tweets from the same three-day window; 500 tweets from each day. Those tweets were collected using the same process as earlier, except this time we also collected its number of favorites and retweets, allowing us to measure the engagement with the tweet. We found that relative engagement would be best measured by looking into those metrics and as we had no previously considered collecting those in the first run, we found that it would be imperative to include them this time around. We also corrected an earlier mistake, which was that we forgot to confirm that we had no repeated tweets, so this time we filtered all the tweets collected to ensure that each post in our dataset was unique. We found that the number of tweets collected was reduced by about 10%, which meant that we would have been double-counting users who post the same thing more than once and users who simply directly copy other tweets. We then concluded that our dataset had been properly collected and we had enough tweets for each of the three-days to proceed with our annotations. During the annotation phase, the dataset was cut down to 1000 posts over the three-day window. This will be covered in the Method section, where we will discuss our categorization of tweets into 'major topics'.

Method

In our initial trial, we aimed to conduct an open coding on the 1,000 tweets we had collected in order to determine the major categories and salient topics discussed in our dataset. We generated a random sample of 200 of the 1,000 collected tweets and conducted an open coding on the sample. We came up with the categories listed in Table 1.

Category	**Description**
Drugs	Vaccines and treatments
Variants	Variants of the SARS-CoV-2 virus
Cases	New infections and deaths
Conspiracy	Conspiracy theories and fake news
Political	Related to politics
Regulations	Travel restrictions, lockdowns, vaccination passports, etc.
Precautions	Precautionary measures including testing, social distancing and wearing a mask

Table 1: Classes

When categorizing the tweets, we followed the following rule:

1. If the tweet belongs to only 1 class, assign that class to the tweet
2. If a tweet appears to belong to more than one class and one class is more specific than the other, the most specific class is assigned to the tweet
3. If a tweet appears to belong to more than one class and no class precisely describes the tweet, try finding another topic that includes it
4. If the tweet appears to be unrelated to COVID-19 entirely, classify it as ”other”

As we observed that about 30% of our 200 twitter posts did not belong to any specific category, and instead had to be labeled as ”other”, we decided that it would make sense for us to repeat our data collection phase as mentioned earlier. We extrapolated that if 30% of any set of tweets collected would fall into no specific category, then we would have to collect around 1,500 tweets to guarantee that we could annotate at least 1,000 tweets in total. Upon repeating the data collection phase as outlined in the Data section, we then repeated our open coding exercise. We decided that if we were able to generate the seven categories listed in Table 1 through our earlier sample of 200 tweets, then it would be a good idea for us to try and use those same categories on our new tweets. This made sense as we wanted to avoid the effect of sampling bias on our annotations.

We then proceeded by coding 1,000 tweets which distinctly belonged to one of the seven established categories (that is each of the 1,000 tweets belong to exactly one topic and none are labeled as ”other”). To maintain an even distribution of tweets over the three-day window, we coded 333 tweets from two days, and 334 from the remaining day. Additionally, we coded a sentiment for each tweet to get a general understanding of the public's responses to the seven established topics. The Sentimentality represents whether or not the twitter post has a positive, neutral or negative outlook in relation to the topic it was discussing.

Prior to the analysis phase, the following preprocessing was carried out on the data we collected and annotated:

1. We convert the tweets to lower case.
2. We remove all punctuation and words containing non-alphanumeric characters to avoid considering hyperlinks and hashtags.
3. We remove stopwords from the text, since they typically do not have significant importance in analysis.
4. We lemmatize the tweets. Lemmatization is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.

After concluding the preprocessing, we were then ready to proceed with our analysis. As the dataset had now been prepared, cleaned, and annotated sufficiently, we would now be able to begin our analysis and generate good-quality results that would not be skewed by the use of improper data. One important consideration we had to make was that since our dataset is not well-balanced, as seen in Figure 1, we decided that we would have to normalize our results per category against the number of posts in that category for a fair comparison across topics. This means that instead of solely analyzing absolute values for counts, we would consider their proportions instead. Our findings will be discussed in the Results section.

Results

For our analysis, we started out by using our topic annotations to get a better idea of the number of tweets within each of the categories and the relative sentimentality towards each of the categories. This allowed us to understand which of the topics are being discussed the most by users on Twitter, as well as their opinions on the topics. Furthermore, through looking at the number of favorites and retweets that a given post received, we were able to measure the engagement with the tweet. Additionally, we ran a Term Frequency-Inverse Document Frequency (TF-IDF) analysis on tweets within our categories to get a better idea of the sub-topics being discussed and the types of words being commonly used.

The number of tweets within each category was a strong indicator of which topics were most significantly discussed on Twitter. So first, we investigated the proportion of tweets that belong to each class. The results can be found in Figure 1.

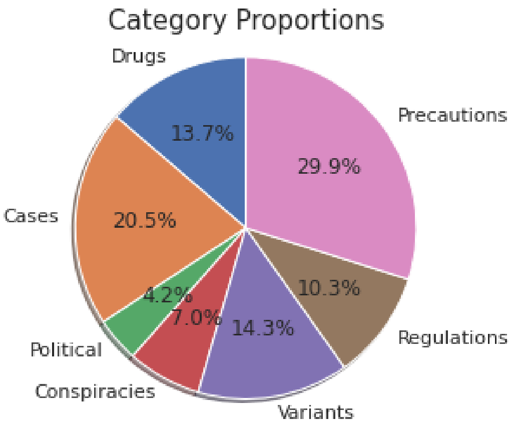

Figure 1: Dataset Proportion

Additionally, the relative sentimentality of tweets within each category allowed us to measure the opinions of the public in relation to our seven established topics. So, for a deeper analysis, we further separated posts belonging to each class by their sentiment. We calculated the proportion of the tweets that belonged to each sentiment within each category. Figure 2 shows the results.

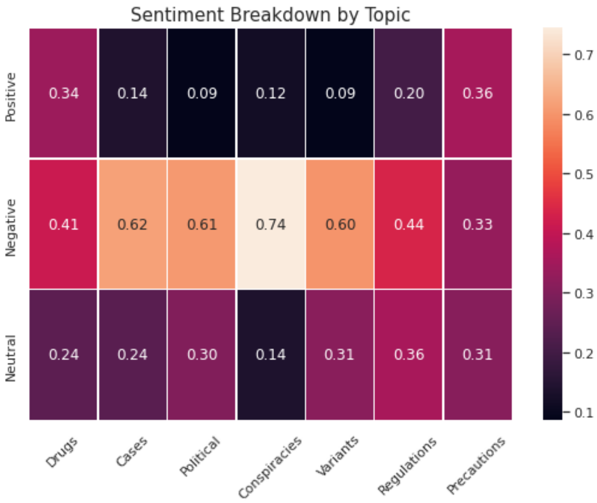

Figure 2: Dataset Breakdown

The relative engagement for each category was also a matter of interest. Figure 3 illustrates the average number of retweets and favourites per post belonging to each category.

Since we were also interested in the relative engagement by sentiment, we utilized two heatmaps, which illustrate the number of average retweets and favourites by category and by sentiment. These results are shown in Figures 4 and 5.

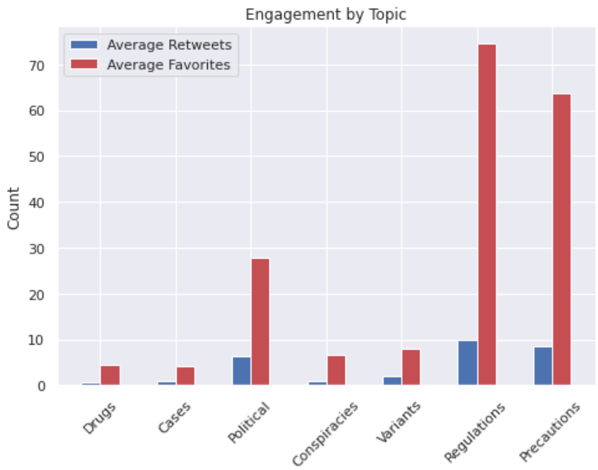

Figure 3: Engagement by Topic

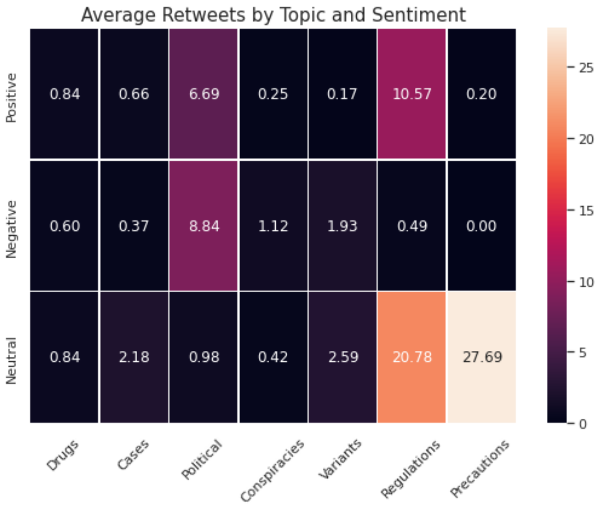

Figure 4: Retweets by Topic by Sentiment

Finally, we conducted a TF-IDF analysis, which allowed us to understand which keywords tweets were using when discussing a specific topic. These frequently used words distinguish these tweets from the rest and thus, provide us with a more informed understanding of the definition of each of our seven categories. We recall that TF-IDF is calculated as follows:

$$tf - idf = tf(t,d) \times idf(d) \tag{1}$$

t = token
d = document
tf(t,d) is the number of times a token t appears in a tweet d.

IDF is given by:

$$idf = log \frac{n_d}{1 + df(d,t)} \tag{2}$$

n_d = number of tweets
$df(d,t)$ = number of tweets d that contain term t

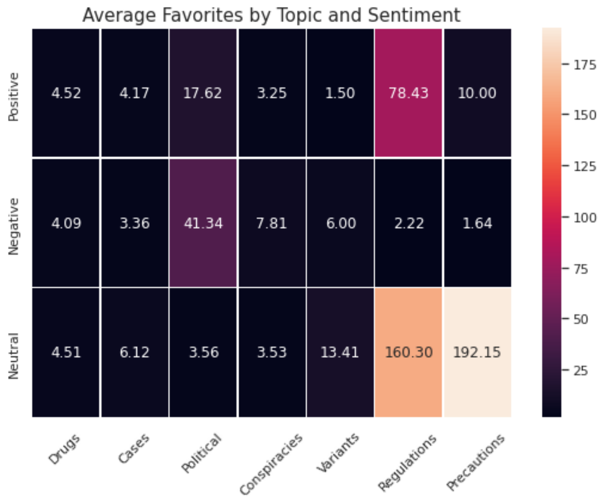

Figure 5: Favourites by Topic by Sentiment

Discussion

We observe in Figure 1 that the data collected was not evenly distributed across the classes. This might suggest that classes that contained too many tweets could have been further broken down, and that those that contained much fewer tweets could have been simply eliminated or better grouped. However, we kept these categories because they followed our coding rule described in Section Method, with the exception of the final ”other” categorization.

The sentiment analysis results as outlined in Figure 2 illustrate how sentimentality is distributed within each category of tweets related to COVID-19. We observe that the proportion of tweets that have neutral sentiment is almost about a third, except for the *Conspiracies* class. We note that a negative sentiment is dominant in all categories, except in *Precautions*. The negative sentiment is significantly present in the *Conspiracies* class and that can be attributed to most of those tweets being strongly opinionated against science and government efforts to curb spread. These people, in an attempt to be convincing, seem to be aggressive or deeply affected, which further explains the strong negative sentiment. The class with the most positive tweets was *Precautions*. This can be explained by the fact that most of those tweeting about the precautions they are taking to curb spread, want to share their relatively good actions and possibly inspire others to partake in similar behavior. We also note the significant negative sentiments (more than 60%) in the remaining classes.

In the next part of our analysis, we looked at the relative engagement within each category. Figure 3 illustrates how many times on average a post belonging to a particular category is retweeted and marked as favorite. These numbers are obtained by dividing the total number of retweets (or favourites) for a category by the total number of posts in that category. We can clearly see that the topics in which people were more engaged are *Politics, Regulations* and *Precautions*. One of the reasons explaining this is that during the three-day period we analyzed, news about ex-United States President Donald Trump contracting COVID and still participating in the presidential debate was trending. Many countries also imposed travel restrictions in parts of the world where the Omicron variant was detected, and those conversations were often very popular. Another regulation that was being discussed was the implementation of vaccination passports, which was often met with positive sentiment. A lot of people were for that proposition, since it brought positive outcomes in states and countries where it was being implemeted. Mask mandates were also a hot topic at that time, explaining the high engagement in both *Regulations* and *Precautions*.

We also explored how the relative engagement per topic was distributed across the sentiments. From Figures 4 and 5, we observe that neutral tweets in *Regulations* and *Precautions* classes were most retweeted marked as favourites. This can be attributed to us finding that most such tweets were objective in nature as they were from reliable news sources simply discussing newly introduced regulations and scientifically recommended precautionary measures. The large values for these cells can be justified by the fact that the those news source have a much greater reach on Twitter than the other randomly collected accounts; some online press accounts might have millions of followers, whereas very few individuals have such outreach. So it is expected that news spread by the press receives more ”retweets” and ”favourites”.

We observe that political tweets with negative sentiment have relatively high numbers of average retweets and favourites. This might be because people wanted to share the news that Donald Trump attended the presidential debate while being sick. People were clearly unhappy about this, and this is shown by the high numbers of retweets and favourites in that category.

Finally, we looked at the top 10 TF-IDF words in each category. The full list of the top ten words by TF-IDF score for each category can be found under TF-IDF in the Appendix section. We note that some words observed are typically used to convey negative messages, including: *dying, suspending, fake, avoid, false, neglect*, etc. We also note the presence of many words that are strongly associated to the pandemic and that in this context, typically convey negative messages, including: *outbreaks, spike, foster, restrictions*, etc.

Contributions

For this project, all three members of the group contributed equally and fairly. We first planned the sequence of work. While Omar was in charge of fetching the tweets from the API, Virasat and Amal had to come up with class names for the open coding. We then agreed on what should be kept or rejected. We also had regular meetings on social media where we would discuss new ideas. For the report, Virasat worked on the introduction and on parts of the data and of the methodology, while Omar and Amal completed these sections and were in charge of visualization and analysis.

Appendix

Filters

Table 2 shows the filters that were used in data collection.

Type	Filter
Common Names	COVID Vaccine Vaccination Jab Pfizer AstraZeneca Moderna
Laboratory Name	Pfizer-BioNTech Comirnaty Vaxzevria Spikevax Janssen

Table 2: Filter Names

TF-IDF

```
{
  "Vaccines/Treatments": [
    "outbreaks",
    "surging",
    "spike",
    "statewide",
    "driven",
    "effects",
    "merck",
    "religious",
    "foster",
    "yourself"
  ],
  "Deaths/Cases": [
    "analytics",
    "players",
    "dying",
    "seen",
    "marcus",
    "november",
    "lamb",
    "bron",
    "lebron",
    "team"
  ],
  "Political": [
    "judge",
    "trump",
    "workers",
    "administration",
    "abiding",
    "suspending",
    "louisiana",
    "occupational",
    "politics",
    "republicans"
  ],
  "Conspiracies": [
    "governments",
    "fake",
    "humans",
    "send",
    "conspiracy",
    "jabs",
    "avoid",
    "false",
    "means",
    "causes"
  ],
  "Variants": [
    "alberta",
    "variants",
    "mutations",
    "japan",
    "sentiment",
    "weigh",
    "mutated",
    "wave",
    "cbcedmonton",
    "hunt"
  ],
  "Regulations": [
    "ny",
    "flight",
    "restrictions",
    "workers",
    "flights",
    "aids",
    "judge",
    "ban",
    "employees",
    "passports"
  ],
  "Precautions": [
    "michigan",
    "neglect",
    "mitigation",
    "virtual",
    "learning",
    "fourseasons",
    "wash",
    "book",
    "precautions",
    "beloved"
  ]
}
```