Analyzing COVID-19 Sentiments on Twitter in Several Selected Areas
COMP 598 Final Project

Marek Borik, Seung Yon Kim, Wendy Wu

marek.borik@mcgill.ca, seung.yon.kim@mail.mcgill.ca, ying.y.wu@mail.mcgill.ca

Introduction

Overview

This paper analyzes the sentiment on the COVID-19 pandemic on the social network Twitter. Our team randomly sampled tweets in the English language over the span of several days. We used a custom annotating script and manually annotated data to find out the sentiment and feelings of the general public. In this process we also discovered the presence of bots and malicious accounts with the sole purpose of creating misinformation, promoting extremism, anger, hatred and polarizing the society.

Key Findings

After collecting and annotating data, we used a customized tokenizer to tokenize all the tweets and counted the word frequencies within each topic. We then used the TF-IDF metric to calculate the most frequent words for each salient topic. Our analysis suggests an overall neutral feeling towards the COVID-19 vaccines on average, which is in part due to many severely negative and severely positive tweets. (Figure 1). Our analysis also shows the most common topic on Twitter with "#covidvaccine" is `Other`, and the least common topic is `Comparison` (Figure 2).

Data

This section explains in detail how we collected data and how we annotated them.

Data Collection

All the data used for this project are collected from Twitter, which is a well-known online micro-blogging service founded in 2006.

In order to access to Twitter from the back-end, we need to use its API. The Twitter official standard v1.1 endpoints were launched in 2012, enabling programmers to post, interact and retrieve data for resources (TWITTERDEV 2020). However, this version has stopped accepting new developer accounts. Thus we had to create a developer account for Twitter API v2 with essential access level. This access level is free and granted us the Search Tweets feature with some limitations:

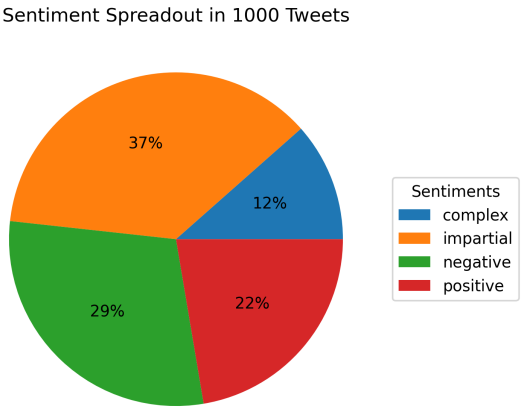

Figure 1: Sentiment from 1000 tweets

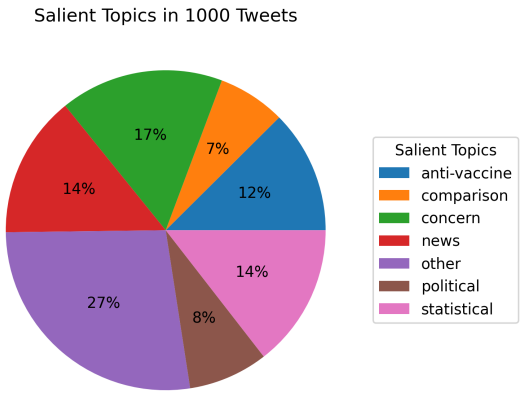

Figure 2: Topic Engagement from 1000 tweets

Figure 3: The terminal result returned by `os.listdir()`. All the CSV files are generated from data collection process and are named as "collect_date_timefrom_to_timeto.csv". Each file containing exactly 90 tweets.

- It only provides a seven-day search endpoint
- This endpoint has a maximum of 100 results and a minimum of 10 results per response.
- The rate limit is 180 requests per 15 minutes per user.

To study the data well and to obtain a more accurate result, we went to the actual Twitter smartphone app and inspected the "COVID-19" tab. We found out the hashtag "#covidvaccine" was a popular trend. Therefore, following the documentation on the Twitter developer platform, we built a query to pull all non-retweeted tweets containing "#covidvaccine" in English using the Request library. To respect the rate limit on the API, we pulled 90 tweets starting from `2021-11-20T06:00:00Z` to `2021-11-24T18:00:00Z` for every 6 hours and outputed each request to a seperate .cvs file (Figure 3).

To better organize the files and facilitate for future data annotator, we re-arranged the data. First, we used the glob library to fetch all the .cvs file pointers. Then by calling the pandas library, we were able to merge and re-index these tweets into a large dafaframe object. We also did a sanity check that filtered out any duplicate contents, regardless of whether or not they have different tweet ids. The final combined .csv file yields approximately 1600 tweets.

Data Annotation

We approached the Data Annotation phase from a unique perspective. We could have just opened the collected data in Excel and annotate them manually, but we opted for a better way. Since there was enough practice in the course, we decided to write a python script that presents the annotator with a nice user interface and simplifies the process (Borik, Marek 2021). The script takes a .csv file as an argument, which was helpful during the earlier testing phase and very crucially saves the results into another .csv file after every annotation. That way the potential loss of work is drastically minimized. The script also speeds up the process, because the person annotating needs to only type one character per topic and one character per sentiment and doesn't need to spell the whole word each time or try to copy/paste. It also supports the ability to skip tweets that were not really meaningful for our discussion and seeking through the file, so there is no need to annotate everything at once.

Methods

Data Analysis: Word Tokenizer

Despite the fact that there are already many libraries that implement a word tokenizer method, most of them are not made specific for analysing tweet data, resulting in incorrect tokenizion of the URL link and the hastags. Since tokenizer is a fundamental tool for processing any data, we decided to implement our own word tokenizer for this project.

The method `tokenizer` takes a string as input and returns a list of tokenized words. We used regular expression `re.findall` to match all the desired pattern. This includes:

- Preserve the URL link. For instance, the string *https://t.co/4t6mgbaG2C* won't be tokenized.
- Separate composite words with different part-of-speech tags. For instance, the word *wouldn't* will be tokenized to *would* and *n't*; the word *you're* will be tokenized to *you* and *'re*.
- Preserve hashtag symbol '#' and at symbol '@'. For instance, the string *#Toronto* will be kept as a single word.
- Preserve mathematical terms. For example, *500,000* will be kept as a single word.
- Preserve special terminologies. For instance, the term *Covid-19* will be kept as a single word instead of being separated to three terms: *Covid*, -, *19*

To post-process the tokenized tweets, we also wrote another method `clean_a_tweet` that takes a list of words as input and returns a list of cleaned words. The main purpose is to lower-case all the words, anonymize all the usernames to *@USER*, and replace all the URL links to the string *URL* by using the regular expression `re.sub`. Punctuation is also removed in this step.

Data Analysis: TF-IDF

Similar to what we have done in class, we have built our own methods that calculate the tf-idfs.
The definition of TF (Term Frequency) used in our approach is:

$$\texttt{tf(w,topic)} = \text{number of times word } w \text{ appears in topic} \tag{1}$$

The definition of IDF (Inverse Document Frequency) used in our approach is:

$$\texttt{idf(w,tweets)} = log\left(\frac{\text{total number of topics}}{\text{number of topics used word } w}\right) \tag{2}$$

The 7 categories that we decided to calculate the tf-idf values were: "anti-vaccine", "concern", "news", "comparison", "statistical", "political", and "other". The method `generate_tf_idf_list` receives as input the category list, the result of word counts as json dictionaries and an integer which determines the number of words in each category with the highest tf-idf scores (10 in our case), and outputs a json dictionary of the top 10 words for each category that has the highest tf-idf score.

Results

Topics Selected and Their Definitions

As mentioned above, there were 7 topics selected for analysis. It is worth noting that many tweets would fit into several categories. We did our best in the Data Annotation part to choose the most dominant category, but presumably a different annotator could feel that some other category was more dominant. We tried to control our personal biases, but it is impossible to objectively put each tweet into a concrete category. The topics and their definitions are as follows:

- `Anti-Vaccine` : Tweets that show hostility towards vaccination or vaccinated people, or tweets related to the topic of anti-vaccination in general. This category usually included most of the negative sentiments around the COVID-19 pandemic and in general had the most extremist opinions.
- `Concern` : Tweets that show any COVID-19 related concerns, but not strictly anti-vaccine sentiments. Tweets that fall into this category were, but not limited to: people or their relatives experiencing negative side effects of vaccines, conclusions drawn from an increased number of infant deaths from vaccinated mothers but also concerns governments not mandating stricter restrictions.
- `News` : Tweets that deliver any news, headlines or short paragraphs regarding COVID-19, or tweets that are an excerpt of the news. Entries in this category usually contained a shortened link to the article which we didn't follow to determine the validity and correctness of the news.
- `Comparison` : Tweets that compare two or more COVID-19 related subjects or themes. Data that falls into this category was, but not limited to: people comparing how different governments handle the pandemic and/or restrictions, comparisons drawn on the efficacy and side effects of different vaccine brands and comparing groups of people with different vaccination beliefs.

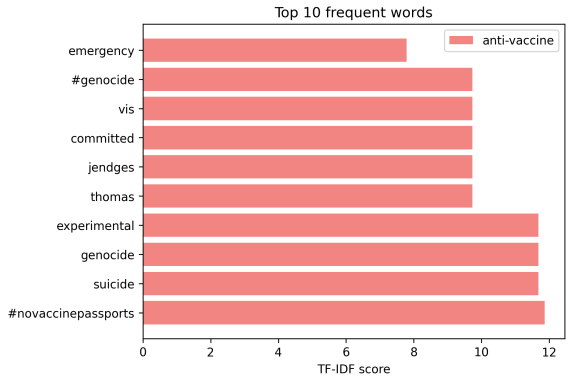

Figure 4: Top 10 frequent words under topic "Anti-Vaccine"

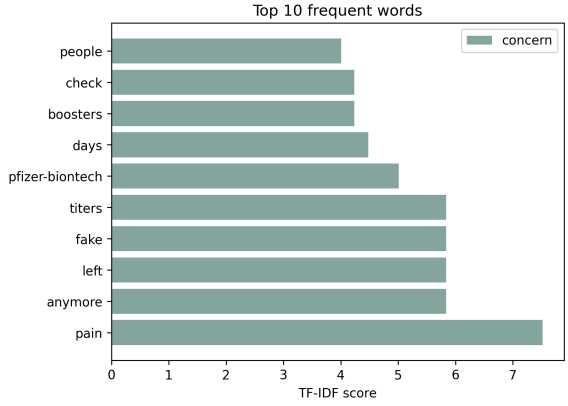

Figure 5: Top 10 frequent words under topic "Concern"

- `Statistical` : Tweets that contain any COVID-19 related statistics. Many of the tweets in this category seemed to be tweeted without a human interaction and were usually depicting the number of slots or vaccines left in a hospital in some region. This category also included statistics about infected people, death rates and vaccine efficiencies, if the tweet didn't contain any subjective opinion and only talked about the numbers. We did not verify the correctness of this information.
- `Political` : COVID-19 related tweets that are used for the purpose of appealing, directly or indirectly, for votes or for financial or other support in any election campaign, as well as tweets directly aimed at politicians or governing organizations.
- `Other` : Any other topics that didn't fit into the previous categories. Tweets in this category were mostly personal beliefs, showing off their vaccination status and encouraging others to get vaccinated, among other things.

Topic Characterization

Anti-Vaccine

- The top 10 words with the highest tf-idf scores (in descending order) are "#novaccinepassports", "suicide",

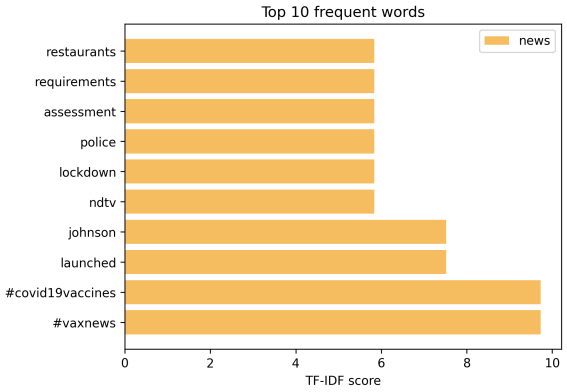

Figure 6: Top 10 frequent words under topic “News”

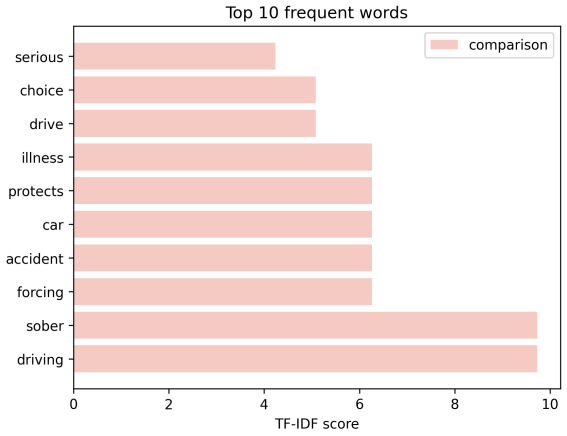

Figure 7: Top 10 frequent words under topic “Comparison”

“genocide”, “experimental”, “thomas”, “jendges”, “committed”, “vis”, “genocide”, and “emergency”. (Figure 4).

Concern

- The top 10 words with the highest tf-idf scores (in descending order) are “pain”, “anymore”, “left”, “fake”, “titers”, “pfizer-biontech”, “days”, “boosters”, “check”, and “people”. (Figure 5).

News

- The top 10 words with the highest tf-idf scores (in descending order) are “#vaxnews”, “#covid19vaccines”, “launched”, “johnson”, “ndtv”, “lockdown”, “police”, “assessment”, “requirements”, and “restaurants”. (Figure 6).

Comparison

- The top 10 words with the highest tf-idf scores (in descending order) are “driving”, “sober”, “forcing”, “accident”, “car”, “protects”, “illness”, “drive”, “choice”, and “serious”. (Figure 7).

Statistical

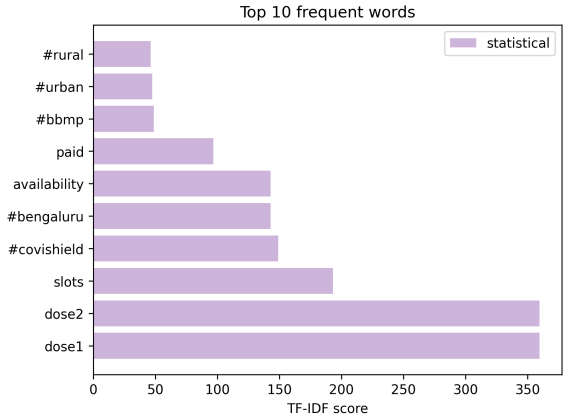

Figure 8: Top 10 frequent words under topic “Statistical”

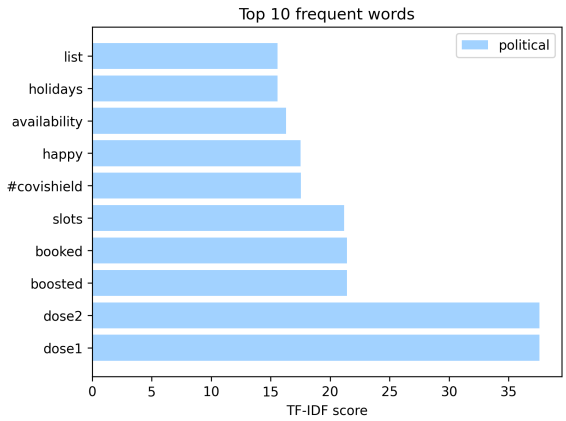

Figure 9: Top 10 frequent words under topic “Political”

- The top 10 words with the highest tf-idf scores (in descending order) are “dose1”, “dose2”, “slots”, “#covishield”, “#bengaluru”, “availability”, “paid”, “#bbmp”, “urban”, and “rural”. (Figure 8).

Political

- The top 10 words with the highest tf-idf scores (in descending order) are “blablabla”, “#pfizergate”, “legal”, “`covidscam”, “#crookedpoliticians”, “apparent”, “#toronto”, “#topoli”, “#profits”, and “#coercionisnotconsent”. (Figure 9).

Other

- The top 10 words with the highest tf-idf scores (in descending order) are “dose1”, “dose2”, “boosted”, “booked”, “slots”, “#covishield”, “happy”, “availability”, “holidays”, and “list”. (Figure 10).

Topic Engagement

The sentiment spread out under each topic is as follows:

Anti-Vaccine Figure 11 shows the proportion of sentiments under topic ‘Anti-Vaccine’. As the name ‘anti-vaccine’ suggests, negative tweets dominate this topic

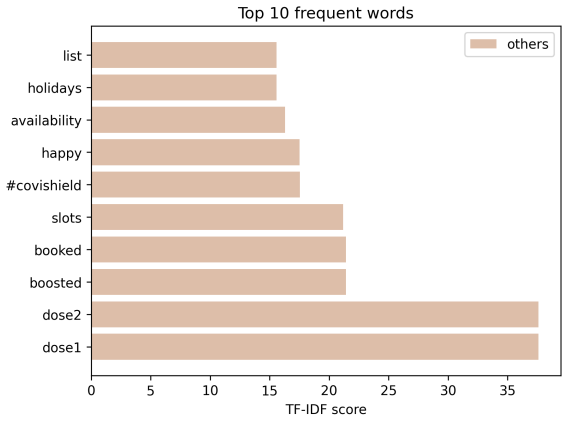

Figure 10: Top 10 frequent words under topic "Other"

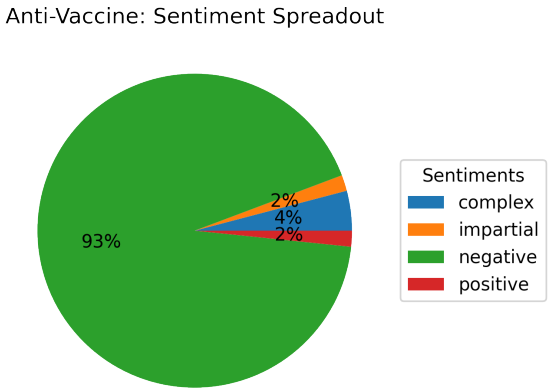

Figure 11: Sentiment analysis under 'Anti-Vaccine'

(93%), followed by complex (4%), impartial (2%) and positive (2%).

Concern Figure 12 shows the proportion of sentiments under topic 'Concern'. Negative tweets takes up half the percentage, followed by complex (30%), impartial (15%) and positive (5%).

News Figure 13 shows the proportion of sentiments under topic 'News'. Impartial tweets dominate this topic (54%), suggesting more informative points of view rather than emotional content that may affect social judgements. Positive tweets took up 24%, followed by negative tweets (13%) and complex tweets (9%).

Comparison Figure 14 shows the proportion of sentiments under topic 'Comparison'. Positive tweets take up the most space (45%). Negative tweets have the lowest proportion (9%). Complex and impartial comments occupy about the same amount, recording 25% and 21% respectively.

Statistical Figure 15 shows the proportion of sentiments under topic 'Statistical'. Impartial sentiment overpowers the other three, recording 96%. Positive and negative tweets

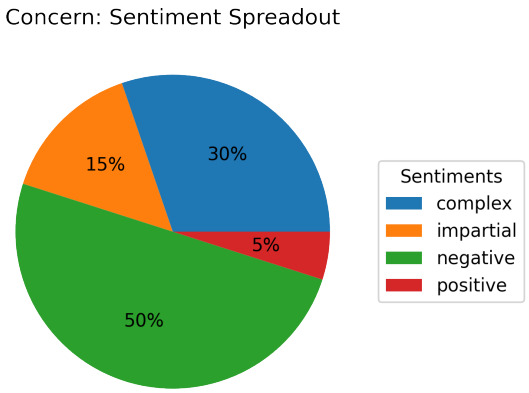

Figure 12: Sentiment analysis under 'Concern'

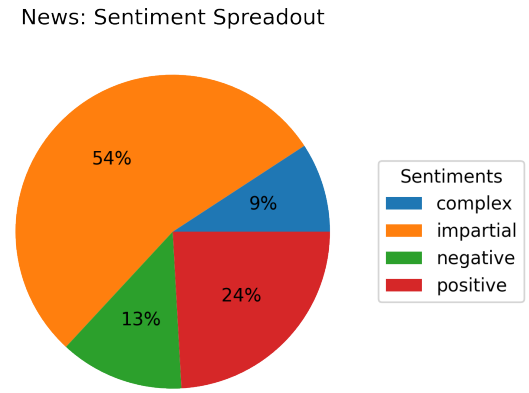

Figure 13: Sentiment analysis under 'News'

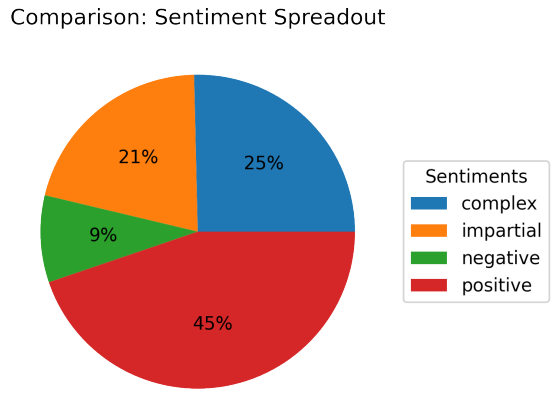

Figure 14: Sentiment analysis under 'Comparison'

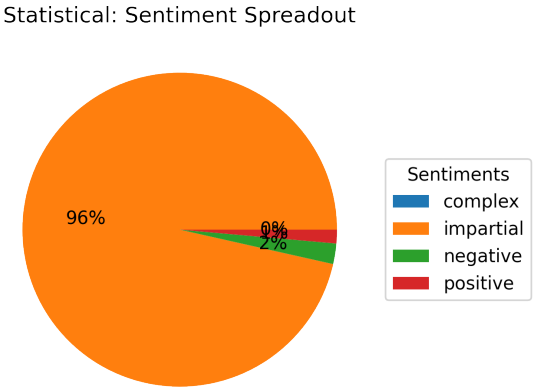

Figure 15: Sentiment analysis under 'Statistical'

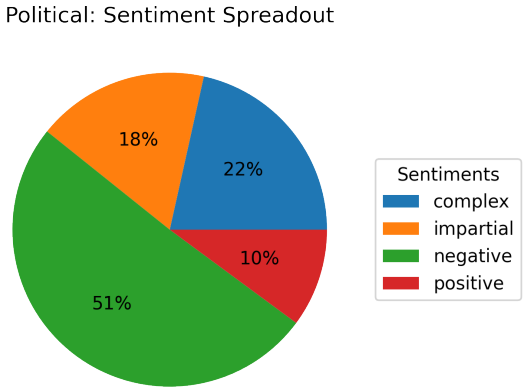

Figure 16: Sentiment analysis under 'Political'

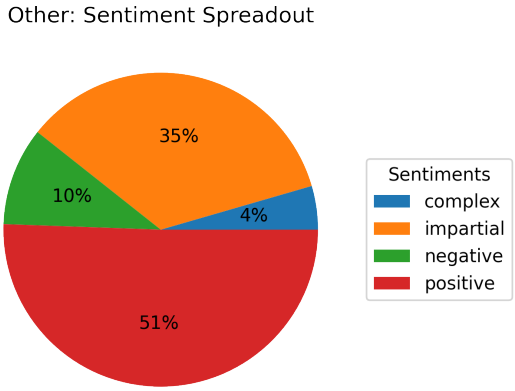

Figure 17: Sentiment analysis under 'Other'

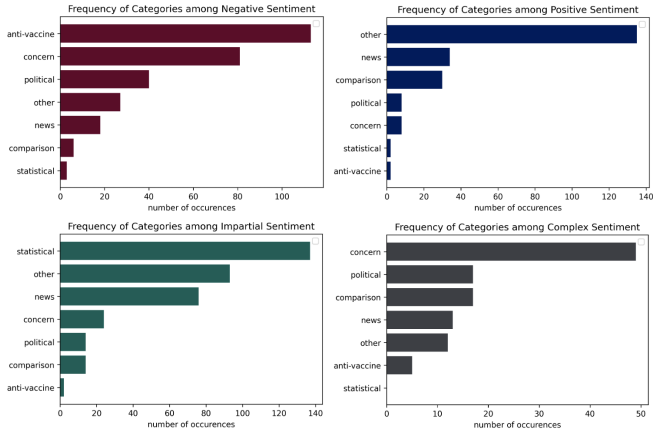

Figure 18: Frequency of Categories among Different Sentiments

both score 2%, and there are no complex tweets (0%).

Political Figure 16 shows the proportion of sentiments under topic 'Political'. Negative tweets takes up about half the percentage (51%), followed by complex (22%), impartial (18%) and positive (10%).

Other Figure 17 shows the proportion of sentiments under topic 'Other'. Positive tweets takes up about half the percentage (51%), followed by impartial (35%), negative (10%) and complex (4%).

Discussion

General Findings

It is perhaps not surprising that we encountered many extremist beliefs on both sides of the problem. The most negative tweets and the most hate speech were in the anti-vaccine category, closely followed by the concern and the political category. This is exactly what we saw in the data - the tweets showing the most hateful speech were in those categories. The most positive sentiments were from the people being proud on getting vaccinated, showing off which dose they received and promoting vaccination. As mentioned before, we put that into the 'Other' category. The most common impartial sentiment for the statistical category, since we encountered lots of statistical tweets, and most of them were simply informative. For the complex sentiment we found many of people reporting side effects of vaccines for them or for someone they know.

Suspicious activities

We believe we have a pretty strong evidence for malicious actors on the Twitter social network with the sole purpose of radicalizing and polarizing the society as well as spreading skewed and out-of-context information. We found several occurrences of word-to-word identical tweets with (usually both) extremely anti-vaccine and anti-democrat tweets that were tweeted several hours apart from each other. There were also several instances of tweets presenting data that

would indicate an increase in child mortalities in a specific hospital. However, we found that the same data and the same numbers were used in several word-to-word identical tweets, where the only change was the name of the hospital or the city. Keep in mind that our sample was consisting of only about 1000 tweets, so this issue must be wide spread when we discovered it in a relatively small sample. We hope that the website would do more to prevent vulnerable groups from seeing such false information.

Possible Improvements

In this section, we will discuss some possible areas for future improvement.

Implementation of TF-IDF For this project, we used the following implementation from one of the past assignment:

$$\texttt{tf(w,topic)} = \text{number of times word } w \text{ appears in topic} \tag{3}$$

$$\texttt{idf(w,tweets)} = log\left(\frac{\text{total number of topics}}{\text{number of topics used word } w}\right) \tag{4}$$

However, this implementation causes many words to share the same tf-idf score due to the fact that the English vocabulary is huge and we only pulled a small amount of data, so most of the words appear one or two times. Therefore, to avoid such problem, we can re-define the term frequency function to be adjusted for document length, the improved tf-idf implementation is as follows:

$$\texttt{tf(w,topic)} = \frac{\text{number of times word } w \text{ appears in topic}}{\text{total number of words in the topic}} \tag{5}$$

$$\texttt{idf(w,tweets)} = log\left(\frac{\text{total number of topics}}{\text{number of topics used the word } w}\right) \tag{6}$$

Stemming and Lemmatization

Looking at results shown in the above section, we have noted that many words, especially those with hashtags, convey the same means but differ in a few characters. For example, the word with the highest tf-idf score under 'Anti-Vaccine' is '#novaccinepassports'. However, there also exist hashtags such as '#novaccinepassport', '#novaccinepassportanywhere', and '#novaccinepassportanywhere'. Since the value of tf-idf scores not only depends on the word frequency of a word but also its uniqueness, devising a way to minimize this effect (i.e. stemming or lemmatization) may increase the accuracy of the result.

Group Member contributions

- Marek wrote the annotating script and did all the data annotation for consistency. He also helped to choose the topics, interpreted the results and analyzed the bot / malicious accounts.
- Wendy did the data collection including creating the developer account, writing the script to fetch tweets while respecting the rate limit, arranging and combining the data. She also participate in the open coding session for deciding salient topics. For the data analysis, she implemented the customized tokenizer and made suggestions on improving the tf-idf method.
- Yona wrote scripts for analyzing word counts of the tweets under each topic, and calculating the tf-idf scores. She also participated in analyzing the tf-idf method, the Results section of the report and added some suggestions for improvements under the discussion section.

References

Borik, Marek. 2021. Ultra Annotator - A python script that helps with annotating CSV data for COMP598 Final Project. https://git.io/JDCHG. Accessed: 2021-12.

TWITTERDEV. 2020. Building queries for Search Tweets. https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query. Accessed: 2021-12-12.