A TF-IDF-based Topical and Sentiment Analysis on COVID Related Tweets

Sung Jun Lee, [1] Sagar Nandeshwar, [2] Namdar Kabolinejad [3]

McGill University[1,2,3]

sj.lee@mail.mcgill.ca,[1] sagar.nandeshwar@mail.mcgill.ca,[2] namdar.kabolinejad@mail.mcgill.ca[3]

Introduction

Using numerical statistics (Term Frequency – Inverse Document Frequency), we aim to discover the current trending topics on Canadian social media regarding the COVID-19 pandemic by collecting and analyzing 1000 Tweets from the popular social media platform Twitter. To gather the dataset required for this investigation, we utilize Twitter's API, which is readily available for public use. Using the API, we are able to create a dataset of English Tweets which mention COVID, the various vaccines as well as the overall responses to the vaccination mandate across the country. Then we manually annotate the one thousand Tweets based on their topics and sentiment into seven topics and three sentiment categories. To analyze our data we go through some minor preprocessing then compute the Term Frequency - Inverse Document Frequency (TF-IDF) scores for each word in each category and order them in decreasing order to generate a list of the top 10 keywords per category based on their TF-IDF scores. By figuring out the keywords per category, we aim to draw some big-data conclusions on the general public's sentiment on the pandemic as well as the government's policies. Our end-goal is to investigate further on the salient topics discussed around the Coronavirus pandemic and what each topic primarily concerns, the relative engagement around these topics as well as how positive, neutral or negative the responses to the pandemic/vaccination has been. We believe that the results from our analysis can provide broader and deeper insight around this topic and also potentially aid in any future investigations surrounding this topic.

Data

We used Tweepy, a Python library for accessing the Twitter API to collect our dataset. In order to collect 1000 tweets, we first created a Twitter developer account to generate consumer_key, consumer_secret, access_token, token_secret and bearer_token for authentication. We generated 26 requests to Twitter's API (each for 100 Tweets) to collect 2600 Tweets in total. We collected 900 Tweets on 30th November 2021, 900 Tweets on 1st December 2021, and 800 Tweets on 2nd December 2021 . We set the language parameter to "English" and used the following keywords (hashtags and words) to filter in COVID related Tweets - #covid (for 260 Tweets), #coronavirus (for 260 Tweets), #pfizer (for 260 Tweets), #astrazeneca (for 260 Tweets), #moderna (for 260 Tweets), #COVID (for 260 Tweets), #vaccination (for 260 Tweets), Covid-19 (for 260 Tweets), COVID (for 260 tweets), and Coronavirus (for 260 tweets). We then filter-out 1200 unique COVID related English Tweets. Out of this we then randomly selected 1000 Tweets for our analysis. We created a dictionary to save our data in a .json file where each record has a Tweet ID and text. We then manually read all 1000 Tweets for correctness.

Methods

The methodology for performing analyses on the collected Tweets were split into three main sections in the following

order of operation: (1) Annotation, (2) Preprocessing, and (3) TF-IDF Score Calculation.

Using the JSON data received from Twitter's API, a short Python script was created to convert the JSON data into a CSV file with the Tweet text, category, and sentiment columns. Then to generate annotations, an open coding was conducted on the first 200 Tweets in the dataset to determine dominant and relevant topics. Upon thorough investigation, it was determined that most of the Tweets fell into one of seven different categories (Table 1): Variant/Mutation (V), Illness (I), Travel Restrictions (T), Covid Measurements (M),Vaccination (N), Covid Cases (C), and Others (O).

Category	Description
Variant/Mutation (V)	Related to new COVID mutations or variants.
Illness (I)	Related to illnesses or deaths caused by COVID or the vaccines.
Travel Restrictions (T)	Related to the international travel restrictions put in place due to COVID.
Measurements against COVID (M)	Related to the measurements (e.g. masking, lockdown, ...) put in place to contain COVID cases.
Vaccination (N)	Refers to vaccines, booster shots, jabs, new vaccine designs
COVID Cases (C)	Related to existing and new number of cases
Others (O)	All other Tweets.

Table 1: Seven categories of Tweets

Furthermore, each Tweet was categorized into three additional groups based on their sentiments: Positive (POS), Neutral (NEU) and Negative (NEG). We believed that seven topics and three sentiments best represented the data and that the most informative conclusions could be drawn.

After annotation, we dedicated our efforts to preprocessing each Tweet's text to make them best suitable for analysis using Python. The following preprocessing decisions were applied:

1. Lowercasing
2. Newline Character Removal
3. Non-alphanumeric Character Removal
4. Stopword Removal (from A8)

Lowercasing was done to remove double counting of the same word with different casing. Newline character removal was necessary to prevent them from altering the TF-IDF calculations. Non-alphanumeric characters were removed using Regex to make it easier to isolate words, and to prevent double counting between a word and the same word with a hashtag in front. Lastly, stopwords were removed to eliminate "meaningless" but frequent words from the TF-IDF counts.

Following the light preprocessing procedure, within the same Python script, a procedure was written to compute the TF-IDF scores for each word per category. The specific formula for computing the TF-IDF scores per category were taken directly from Assignment 8 - COMP 598:

$$TFIDF(w, ct, sc) = tf(w, ct) \times idf(w, sc)$$

$$tf(w, ct) = \#\ of\ times\ word\ w\ was\ used\ in\ the\ category\ ct$$

$$idf(w, sc) = log(\ \#\ of\ ct\ /\ \#\ of\ ct\ that\ use\ w)$$

The impact of our preprocessing decisions were noticeable. Compared with blindly running the TF-IDF script on the original unprocessed Tweets, the generated top 10 words for every category were much more relevant, rid of meaningless words, and contained no duplicate words.

	Variant	Illness	Travel	Measures	Vaccine	Cases	Others	Positive	Neutral	Negative
1	moderna	iran	fly	mask	booster	analytics	cancelterm	booster	analytics	pfizer
2	variant	lung	pompeo	wear	pfizer	insights	pmoindia	moderna	insights	iran
3	dangerous	democraps	review	masks	dose	county	target	janssen	usafacts	poses
4	poses	players	special	shops	moderna	confirmed	price	delivered	cancelterm	emergency
5	reformulate	body	2022	transport	janssen	data science	iii	lowest	county	dangerous
6	rises	hospital	gold	wearing	vaccine	public health	con	rises	confirmed	moderna
7	current	lungs	press	lock	doses	observe	en	dose	moderna	trials
8	evade	shocking	able	airborne	shot	lowest	cancelboar dexams	pfizer	booster	evade
9	overall	adverse	airport	herd immunity	flu	curve	pfizer	book	pfizer	joke
10	emergency	death	probably	facemask	unite2fight corona	deaths	cycling	pcr	fly	f*ck

Table 2: Top 10 Words by TF-IDF Score from each Category in Descending Order (1 is highest TF-IDF)

Results

The results from our analyses were such that it varied greatly per category and sentiment. Table 2 illustrates the details of our findings.

First, upon the open coding of 200 Tweets, we discovered that the most frequent and salient topics were the seven chosen in Table 1. We discovered that the trending topics online were similar to the ones being mentioned offline, with a heavy emphasis on the new Omicron variant and vaccination. The primary concerns for the Variant category appeared to be the worries and dangers of new variants in the virus, as well as Moderna, whose vaccine may require a third shot for effective protection. The primary concerns for the Illness category can be characterized by people's worries on the lethal effects of COVID on our lungs as well as Iran and our medical infrastructure which have been taking a heavy toll from the virus recently. As for Travel, there was higher optimism for flight and future travel in 2022, as well as mentions about Mike Pompeo, the American politician responsible for decisions on America's response to COVID. As for the Measurements against COVID category, we were able to notice a clear pattern about masks, shops, public transport, and herd immunity - characterizations which seem to have not changed greatly since the beginning of the pandemic. The Vaccine topic had the greatest engagement for a single category (except Others), which frequently mentioned the need for a third booster shot from Pfizer, Moderna, and Janssen to combat the most recent variant. There were also a surprisingly low number of Tweets in the COVID Cases category, which as expected, showed data science keywords such as "analytics", "insights", and "curve". There was a fairly high amount of optimism for this category as analysts reported lower than previous numbers. Lastly, our findings for the Others category were ambiguous to draw

any solid conclusions from. Although there were trending mentions about school term cancellations, and stock prices, the miscellaneous data and some spam Tweets made the data difficult to interpret.

As for engagement in each topic category, we see in Table 3 that the most occurring topics were Others, Vaccination and Variant/Mutation. The high number of occurrences for the Others category can be explained by the fact that any other Tweet which did not fall under the six other categories were grouped into Others. There were also a surprisingly low number of Tweets about COVID cases and illnesses, which we speculated would be a dominant category. Further explanations are detailed in the Discussions section of this report.

Category Count	
Others (O)	343
Vaccination (N)	279
Variant/Mutation (V)	170
COVID Cases (C)	80
Measurements Against COVID (M)	45
Travel Restrictions (T)	42
Illness (I)	41

Table 3: Number of Tweets per Topic

We can also draw the conclusion that the overall sentiment towards COVID on Twitter is overwhelmingly and surprisingly neutral, where it was speculated that it would be dominantly negative. There are nearly three times as many neutral Tweets compared to negative Tweets and more than six times the number of positive Tweets, as shown in Table 4. This appears to be the case as there are a high number of official organizations and members of older demographics on Twitter. Further explanations are detailed in the Discussions section of the report.

Sentiment Count	
Neutral (NEU)	632
Negative (NEG)	266
Positive (POS)	102

Table 4: Number of Tweets per Sentiment

Discussion

Given the nature of the Coronavirus and the fact that the Omicron variant was just discovered and was on the rise in the period that we collected the data, we expected more negative Tweets than positive or neutral. However, the annotations made do not exactly follow the expected pattern. For one thing, the number of neutral Tweets is much more than the negative Tweets. The lack of positive and negative sentiment might be attributed to the fact that a large number of the Tweets were not opinionated posts but rather news and report posts or simply a statement which are in essence neutral, e.g. “Gujrat Govt authorizes police to enforce #COVID19 #vaccination” or “I got my first dose”. However, within the rather small number of opinion posts that can have a positive or negative sentiment, more than two thirds of them are negative.

Moreover, if we group the Tweets by category and count the number of each sentiment for each category we see that the number of neutral Tweets is much higher than negative ones (nearly double), and the number of positive Tweets are much lower than the negative ones. However, we have some categories that are negative in nature than the rest which have a higher number of negative Tweets than average and even less positive Tweets. For example the Illness (I) category has about the same number of neutral and positive Tweets but more than four times the number of negative Tweets. This is clearly due to the negative nature of illnesses and deaths which make it hard

to make any positive statements about them. Furthermore, the Variant/Mutation (V) topic has higher negative Tweet counts compared to other categories. This might be due to the fact that the new variant of Omicron was on the rise and was rather disappointing news to the population. People were already exhausted from COVID restrictions and the moment when countries relaxed them, they could not hold their frustration any longer after hearing about the emergence of a new, much stronger variant.

Group Member Contributions

The contributors of this project were Sung Jun Lee, Sagar Nandeshwar, and Namdar Kabolinejad.

Sung Jun Lee

- Annotated 500 Tweets
- Wrote Python script for Preprocessing and TF-IDF Calculations
- Wrote Introduction, Methods, Results, and Group Member Contributions for Report

Sagar Nandeshwar

- Wrote Python script for collecting Tweets
- Wrote Introduction and Data for Report

Namdar Kabolinejad

- Wrote JSON to CSV conversion Python script
- Annotated 500 Tweets
- Wrote Discussion for Report