Analysis of Twitter publications concerning COVID-19

Mohamed Elenany[1], Paul Bergeron[2], Alex Liu[3]
McGill University[1,2,3]

Introduction

In this paper, we collect 1000 covid-related tweets using specific keywords. Through open coding and manual annotation, we categorize them into 8 topics: travel, vaccine, measures, opinion, politics, advertisement, research and statistics. We also annotate each tweet with their perceived sentiment (negative, positive or neutral). By computing the TF-IDF score of each word, we then produce a top-10 list of most relevant words for each topic. Key findings in this respect include texas-related discussions topping the politics topic, boosters being a main focus of the vaccine topic, and mutations for the research topic. Furthermore, we found a sharp incline in negative sentiment around the same time the omicron variant was publicised. Overall, positivity was still quite low. Using TF-IDF scores for each sentiment, we found that politics were most relevant for negative tweets, while a theme of thankfulness was the counterpart in positive tweets. Relevant words for neutral tweets were mainly centered around research.

Data

Our dataset consists of 1000 tweets that were carefully collected with the Twitter API and the Tweepy python library. The tweets collected were posts published by users using the English language from the 25th November to the 27th November of 2021. We have chosen to collect tweets based on two criterias. The tweets could not be retweets and each and every tweet had to contain one word from a list of Covid related words, which we will now call “filters” in the rest of this discussion. We first decided to apply filters that were simple words for the textual content of tweets, but then decided to only use filters with hashtags, since the first method led to us collecting a lot of tweets unrelated to our subject. The list of words we used to collect the tweets was (all of them had the hashtag symbol in our code) : SARSCoV2, COVID, Vaccinated, COVID19, vaccine, moderna, pfizer, Astrazeneca, antibodies and antibody. These words are written here with capital letters as they are in our code, but the Tweepy API makes case insensitive queries, thus case-sensitivity was not an issue in the data collection process.

We started by collecting 1500 tweets, consisting of 500 tweets per day mentioned above. Then, we randomly chose 333 tweets of the first two days and did the same but for 334 tweets of november 27. This was done in order to reduce the impact of outlier tweets. We then had to clean our data, hence we went through it once to ensure there were no unrelated tweets or repeated ones. After that, we ran a Python script on our dataset to clean all of the tweets from irrelevant characters, hashtags, emojis, mentions and stop words in order to later compute the tf-idf on those posts.

Methods

We had to make some decisions in the data collection phase in order to orient our project the way we wanted it to be. First, we selected the period of time from November 25th to November 27th because we didn’t want our project to be overflown by the appearance of the Omnicron variant and the resulting wave of tweets related to it. We wanted to have a portrait of the current discussion about Covid as it was before this relatively bad news. In the data collection phase, we also decided to randomize the time of the tweets we collected since the Tweepy API collects tweets in the order at which posts have been published and we didn’t want to only get tweets posted at midnight or just a little before that time.

With our clean data at hand, after carefully examining 200 of the 1000 tweets we collected, we came up with 8 distinct topics to categorize the different subject of discussion people have had in our tweets..We then computed, for each of these topics, which are going to be specified later, the count for each word encountered in the cleaned tweets. From that, we only kept the words that occurred more than 5 times across the dataset. Then, we computed the term frequency-inverse document frequency of all terms we found in the tweets we collected. We did so by multiplying the number of occurrences of the selected word in the tweets of a specified topic with the logarithm of the number of topics divided by the number of topics that used the selected word.

Results

The names of these 8 topics are: advertisement, measures, politics, public opinions, research, statistics, travel, vaccine. **Advertisement** is pretty self explanatory. As we went through the 200 tweets, we saw that some businesses did advertise products using Covid as a way to attract some attention, thus resulting in this topic as one of our categories. These tweets include a link to a product or a service that can be bought and they do not bring any relevant information about the Covid situation or the perception of it. **Measures** is a topic designed to include any post that treats the subject of measures taken in order to control the propagation of the virus. These include tests, masks, social distancing, limitations on the number of people that can be present in a certain location at a certain time, the deployment of vaccination, etc. Here, it is important to distinguish the efforts put in the deployment of vaccination, as in the logistic sense, from the topic of vaccination, which will be covered later. **Politics** includes any post that discusses politics. As most measures are taken by people of power, these two categories can be confusing. Any post discussing the measures including a political take towards those is considered a political post, thus ending up in the "politics" category, where posts discussing measures without a political stance are labeled as "measures". Also, any post not discussing measures but including a political take or only discussing politics, without taking position, was put in this category. **Public Opinions** is a rather broad category which contains every tweet in which people simply voice their opinion towards a particular subject that is not politics. In the case of our dataset, it is mostly covid, but some people simply voiced their opinion, which had a very small link to the covid situation, and their tweets ended up in this category. **Research** is a category for every tweet that discusses a new research advance or that simply states the actual progress of a certain field of research related to Covid. **Statistics** includes every post that simply states statistics concerning the covid situation or a related subject. If a post included statistics about a research as well as an interpretation for those, it would have been considered as a research post. **Travel** is a topic for tweets that discusses the impact of covid on travelling, may it be the bans, the restriction or just missing travelling. **Vaccine** is the category we came up with to label all tweets that were about the subject of vaccines. Tweets about the logistics of vaccination were not labeled under this category, as they fell in the "measures" one.

After completing the computation described above, we obtained results for the words with the highest TF-IDF scores for each topic, which we have listed below, in decreasing order of TF-IDF score.

- **Advertisement**: Tauranga, womens, available, project, happy, pm, recovery, strong, investigation, booking.
- This category represents 10.2% of the tweets collected for a total of 102 tweets and the sentiment distribution is 6 negative, 86 neutral and 10 positive.
- **Measures**: Measures, masks, wear, season, mask, holiday, mandate, wearing, tests, protect.
- This category represents 6.5% of the tweets collected for a total of 65 tweets and the sentiment distribution is 20 negative, 28 neutral and 17 positive.
- **Politics**: Texans, workers, women, trying, government, criminals, Biden, mrna, ban, Trump.
- This category represents 10.9% of the tweets collected for a total of 109 tweets and the sentiment distribution is 68 negative, 32 neutral and 9 positive.
- **Public Opinions**: Lt, fake, family, sick, rates, pretty, wrong, truth, cdc, hope.
- This category represents 28.0% of the tweets collected for a total of 280 tweets and the sentiment distribution is 150 negative, 95 neutral and 35 positive.
- **Research**: Mutations, dr, delta, worse, b11529, warn, SARSCoV2, concern, scientists, potential.
- This category represents 9.1% of the tweets collected for a total of 91 tweets and the sentiment distribution is 27 negative, 54 neutral and 10 positive.
- **Statistics**: Analytics, USAfacts, insights, county, team, distribution, confirmed, 1k, bc, growth.
- This category represents 13.6% of the tweets collected for a total of 136 tweets and the sentiment distribution is 18 negative, 117 neutral and 1 positive.
- **Travel**: Ban, flights, travel, southern, holidays, monday, authorities, restrictions, countries, situation.
- This category represents 3.2% of the tweets collected for a total of 32 tweets and the sentiment distribution is 15 negative, 16 neutral and 1 positive.
- **Vaccine**: Booster, mrna, jab, vax, vaccinated, dose, shot, vaccination, vaccines, im (for I'm).
- This category represents 18.5% of the tweets collected for a total of 185 tweets and the sentiment distribution is 64 negative, 50 neutral and 71 positive.

We also ran our TF-IDF computation by sentiment, using our sentiment classification as the "topic" in the computation, which yielded the results following.

- **Positive**: Thankful, thank, thread, local, flu, helping, yourself, month, life, China.
- **Neutral**: Analytics, USAfacts, distribution, team, insights, 1k, confirmed, growth, 7d, tauranga.

- **Negative**: Govt, human, pretty, caused, rights, texans, insist, accept, inactive, viron.

Discussion

From the results we acquired, there are many interesting observations that can be drawn out. The first one is that the general tendency of people that posted about Covid19 just before the apparition of the Omnicron variant is people voicing negative opinions about the current situation. Indeed, we can compute from the data that in total, 368 posts that we collected were considered to have a negative sentiment to them, where only 154 posts seemed to be positive. Also, there were 478 tweets that were labelled as neutral as a lot of publications simply stated facts or tried to advertise a product. We also can see from the results that 28% of posts were people voicing their personal “non-political” opinion about the situation and that constitutes the largest category of them all.

In the advertisement category, which does not give any insight about the Covid situation, we can observe that there is a relatively large amount of businesses using Covid to draw attention to what they are selling since 10.2% of the data we collected was part of this category. We can also observe that since advertisements generally do not take a stance on anything apart from the product sold, these posts were considered as neutral by our team. We can advance that since Covid is still the talk of the hour, there are a lot of people trying to make money out of this situation. Also, there are some very obvious words we found in our most frequently used terms list as well as some bizarre anomalies. Words such as available, project, happy and booking give a clear direction towards the commercialisation of a certain product, but words such as Tauranga, recovery and investigation are a little confusing, as they do not appear to be very marketable choices of words.

When looking at the results we got from the measures category, we can first observe that the list of most important words in terms of TF-IDF score indicate a trend of people discussing masks in particular. Indeed, we can find in this list two versions of the word “mask” as well as two different versions of the verb “to wear” which suggest that trend. Also, since there are 20 posts about measures that are negative for 17 positive posts, it is safe to say that the general opinion about the different measures in place is mitigated and that there is an ongoing debate about those online.

As stated earlier, the most important topic in terms of number of posts is the public opinion category, which indicates that there is a trend of people using twitter as the place to speak up about what they think of the Covid19 situation as a whole. Since there are 150 negative such posts, for 95 neutral and 35 positive, the general attitude of people posting their opinion about Covid seems to us as negative, which is not surprising as this situation is far from being fun and it seems as if we are not about to get out of it. The words that came out of the TF-IDF computation does not seem to indicate any kind of trend, but the presence of the words “fake”, “sick”, “wrong” are coherent with the large negative number of tweets of this category. Those words generally show how the public opinion are really questioning decisions and not believing news things that are being said about Covid.

For the politics topic, the word list is very clearly related to the subject. The presence of “texans” in this list is no surprise. Indeed, the state of Texas is a boiling political scene, with very polarized views of what should be done about this Covid situation. Since 68 out of these 109 posts are negative and words such as “criminals” can be found in the most frequent words of this category, it is possible to observe a high amount of dissatisfaction in the general public towards the political class, especially the American one as Biden and Trump are named in the most frequently used words list.

The research topic is a rather neutral one, as most posts simply stated advances in the different researches currently being conducted about Covid and treatments for it. But, there is a clear indication that there is some concern in the scientific community towards variants, as the words “delta”, “b11529”, “mutations”, “concern” and “worse” are all present in the word list. Even though our dataset was collected before the wave of Omnicron variant posts, the variant was still already discovered, which explains the presence of the “b11529” word in the list. Also, there were still 27 negative posts for 10 positive ones in this category, so it seems as if the opinion of what is currently developing in the research field is either neutral or negative and that negativity would be explained by the expressed concern of the community towards variants.

We found that there were a lot of posts strictly about statistics and that those were mostly neutral, as 117 out of the 136 statistics related posts were categorized as neutral. Nothing surprising was found in the list of most frequent words, as they are all very related to statistics.

It was clear from the data that we analyzed that the perception of traveling in the current Covid setting was negative or neutral. When looking at the words the TF-IDF computations gave us for the topic of ‘travel’, we observe that the first word is “ban”. We can also find the word “restrictions” in the list, which, combined with the fact that the sentiment distribution of this category is 15 negatives, 16 neutral and 1 positive, indicates that there is a negative reaction of the public towards the restrictions and the unavailability of travelling. Also, since there were only 32 posts of this category in the 1000 tweets we used, it seems as if it is not a very important subject right now, as people have a lot of other Covid related things to discuss more than this subject.

Vaccination has a very different sentiment distribution than the other categories. In this topic, 64 of the posts we collected were of negative nature, 50 were neutral and 71 were positive. This topic and the topic of advertisement are the only two to feature more positive posts than negative ones.When looking at the most important words of vocabulary used in this topic, we observe that there are a lot of words that simply describe vaccines such as “booster”, “jab”, “vax”, “dose”, “shot”, “vaccines”. The fact that there are more positive posts than negative ones indicates that there is a trend of people posting to declare that they received the vaccine. Also, from the sole sentiment distribution, we can observe that the subject is very mitigated, but it is a very mitigated subject in the real world in which we live, so there is nothing surprising there, only the confirmation that the debate about vaccination is also taking place on Twitter.

Group Members Contributions

- Paul Bergeron: Took part in the data annotation, wrote this final report.
- Mohamed Elenany: Collected and filtered Tweets, took part in the data annotation step, performed cleaning on the text of all tweets, and took part in writing some of the report.
- Alex Liu: Took part in the data annotation, stop words removal, and wrote some of the report.