Tweets analysis related to Covid-19

Wenyang wang (wenyang.wang@mail.mcgill.ca)
Yichen Huang (yichen.huang@mail.mcgill.ca)
Feiyu Guo (feiyu.guo@mail.mcgill.ca)

Introduction

In order to understand the discussion currently happening in social media, we collected 1100 tweets in English during 3 days, which mentioned the keywords related to COVID, and filtered out links and retweets.By observing 200 posts we developed 4 salient topics: and classified them in data annotation with setiment. After computing the top 10 words with highest tf-idf scores for each topic, we discussed the result and concluded insights about how positive the people responded to the vaccination and how negative the people responded to the COVID pandemic.

In conclusion, we found that people had an overall negative attitude to the epidemic situation because of the negative life impact and restrictions but they were positive to get vaccinated.

Data

In the first attempt, we setted the result type as: 'popular' , which helps us select the posts by considering the 'popular' factor of the posts by returning the top likes and retweets amount. We thought that helps us to collect salient topics that represent the opinions that most people discuss. However, since it may be impacted by the twitters' identity, such as the number of followers/ the social identity of the account owners etc,we realized these can lead the unrepresentative dataset to use, so we decided to change the method to collect tweets.

We filtered out all Retweets and duplicate tweets from the same user, such that each post we collected in our collection will be unique without repetition. Meanwhile, we filtered out tweets containing links because the keywords in the link could not be accessed directly in our dataset file, which makes their topics and opinion unclear.

The time zone is setted from 2021-11-26 00:00:00 to 2021-11-28 23:59:59 with any one of the five keywords (case-insensitive): "COVID" / "Vaccine" / "Vaccination"/ "Pifzer" /"Mordena"/"Sinovac"/"Sinopharm"/"AstraZeneca".

Under these restrictions, we collected 1100 tweets in the English language by random selection. Going through the data we get, we delete some tweets that have no meaning (with detailed description in the METHOD part). For example, some tweets contain only keywords or just use the keywords as nicknames,etc. After removing these meaningless tweets we get 1073 posts. At last, we picked 1000 posts again and formed our data set to continue with the following project.

Methods

After we got 1100 posts filtered by code, we randomly selected 200 tweets from the sample and each teammate classified these tweets into several topics separately.we annotated these 200 tweets from this data set and found following problems:

1.Searching with " vaccine" or "vaccination" as keywords is possible to collect a very small number of non-covid vaccines, such as chickenpox/hpv vaccination related.

2.Searching with "covid" as keyword will search a very small number of people using covid as a metaphor or the symbol of time. (e.g. ex is like covid./ Last time I went to ski is 2 years before covid)

3. The Keywords are contained as nicknames with no attitude or the whole content of the post is meaningless.

Considering some posts in these conditions may be irrelevant tweets for our topics and hard to filter out by code automatically, so we recollected 1100 posts using the same strategy and we would have tolerance to drop these unrelated tweets during the annotation part.

After removing these meaningless tweets we got 1073 posts,then we randomly dropped 73 tweets and got the final data set to continue with the following project.

During an open coding on 200 tweets, we try to summarize each tweet into one category as much as possible. Each teammate designed a set of topics with classification rules. After discussing and comparing our topics that were

developed by each person, we finally developed the above 4 topics with classification rules to category tweets, which makes each tweet belong to exactly one topic.

For the further annotation part, we re-sampled 200 tweets from a sample set, sorted them and coded each post for positive/neutral/negative sentiment (Each teammate did coding individually). We wrote a python script to find tweets that we labeled with different annotations and discussed our opinions on these examples. After preparing these challenging examples, we added more rules for categorizing topics and coding sentiment.

In conclusion, by summarizing and discussing 200 tweets, we designed 4 salient topics discussed around COVID with each topic primarily concerns following,which is also our code book preparation for annotation of the whole data set.

After we finished the topic design and annotation standard, we separated the remaining tweets and annotated the rest of the data set. When collecting data, we avoided retweets and links since they do not contain personal opinions of the author about COVID-19.

Secondly, we use a python program to remove symbols and we created a words list containing frequently used common vocabulary with no obvious attributes (similar to stopwords.txt used in HW8, with some thesis words 'COVID'), such as 'or' ,'and', 'with' ,'on', 'so' or some alphabets, because those words would not help us to do further topic analysis.

We sorted all the appearing words by computing their tf-idf scores, storing its dictionary as json format in descending order. However, we found out there still exist some meaningless words such as 'gonna', 'maybe', 'sounds' in our words dictionary. So we add those words to our stopwords.txt to improve our result for engagement analysis.

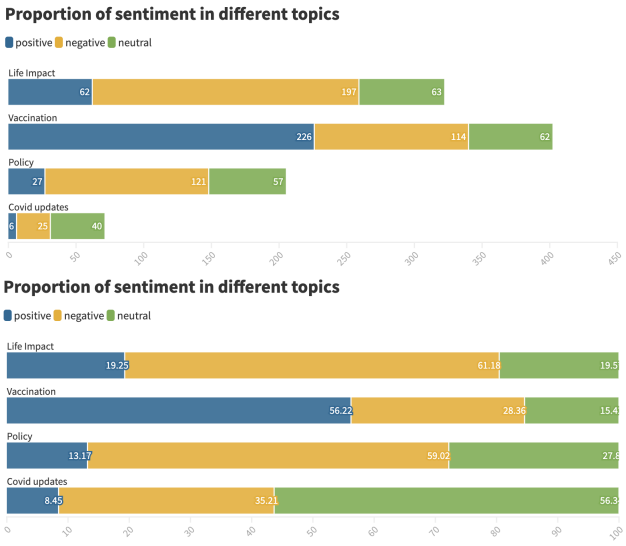
To find the main concerns under each topic, we kept only the Top 10 highest tf-idf scored words for each topic, because most words with smaller tf-idf scores were not really meaningful.

At last, we reached results and we made conclusions about these salient topics discussed around COVID and what each topic primarily concerns. By Top 10 highest tf-idf scored words, we figure out relative engagement with those topics.

And we visualized the results by bar charts.

Results

Overview of 4 topics



Topic 1 COVID Case: 71 (7.1 percent)
Topic 2 Policy: 205 (20.5 percent)
Topic 3 Vaccination: 402 (40.2 percent)
Topic 4 Life Impact: 322 (32.2 percent)

Positive: 321 (32.1 percent)
Negative: 457 (45.7 percent)
Neutral: 222 (22.2 percent)

Topic 1: COVID Updates

This topic is related to the discussion about COVID update reports about screening rates or case numbers.

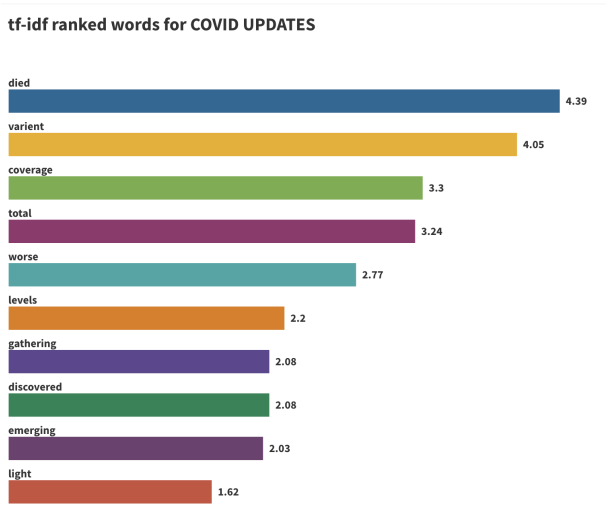
Positive:Statements imply the situation of COVID becomes better, such as the number of cases reduced/ more cured cases reported.

Negative:Statements imply the situation of COVID becomes worse, such as the number of cases increasing rapidly/ bad news such as new discovery of variant virus.

Neutral:The description for COVID situation with only case number updates without position trendings.

Total related posts of Topic 1 in 1000 posts: 71
Positive: 6 (8.45 percent);
Negative: 25 (35.21 percent);
Neutral: 40 (56.34 percent);

Top 10 words with highest tf-idf scores(in descending order):
"died","variant", "coverage", "total", "worse", "levels", "gathering", "discovered", "emerging", "light"



We could see the number of tweets with negative sentiment is about 4 times more than those with positive sentiment. In other words, people's response to covid update reports is generally negative.

The relative engagements from Top 10 words are most about:
The mortality of COVID ('died')/The emerging variant of virus / Total number of cases discovered/ Coverage of epidemic area/ Social mass gathering / The situation of COVID get worse/ The levels of COVID symptoms light/ Discovery of COVID

Topic 2: Policy

This topic is related to the special policy related to COVID, such as public masks policy, border restrictions, vaccine passport etc.

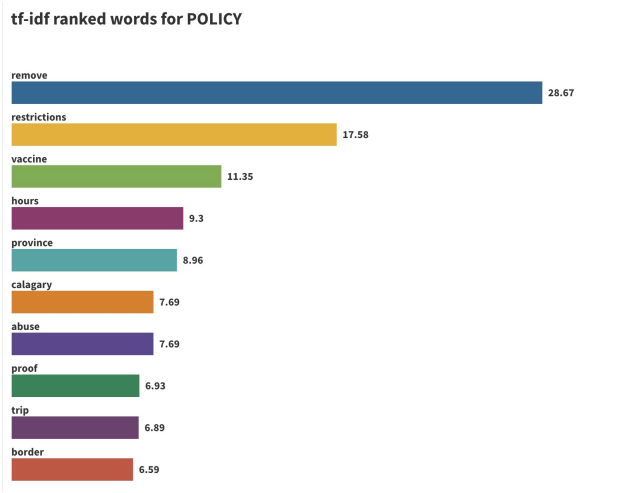
Positive: Statements that encourage or support the policy, posts expressed supporting related-policies under COVID with an optimistic attitude.

Negative:Statements that accept or discourage the policy, posts expressed rejecting related-policies under COVID with a negative attitude.

Neutral: Detailed announcements for related policies with no emotion.

Total related posts of Topic 2 in 1000 posts: 205
Positive: 27 (13.17 percent)
Negative: 1215 (59.02 percent)
Neutral: 57 (27.80 percent)

Top 10 words with highest tf-idf scores(in descending order):
"remove", "restrictions", "vaccine", "hours", "province", "calgary", "abuse", "proof", "trip", "border"



We could see that the number of tweets with negative sentiment is about 4 times more than those with positive sentiment. In other words, the response to policy is generally negative.

The relative engagements from Top 10 words are most about: The restrictions policy under COVID/ The quarantine time policy/ Province policy (Calgary especially) / Vaccine Policy/ Trip Policy include across border

Topic 3: Vaccination

This topic is related to vaccines and vaccinations, including people's attitude to get vaccinations and side-effect reports.

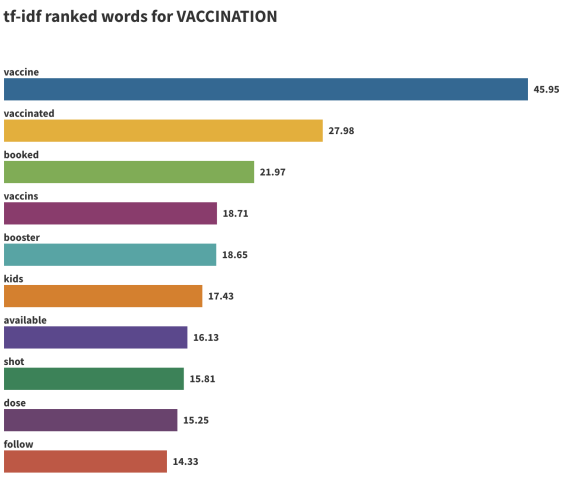
Positive: Supporting or encouraging people to get vaccinations, tweet posts with a positive attitude to get vaccinations of COVID.

Negative: Anti-vaccination or discouraging people to get vaccinations, tweet posts with a negative attitude to get vaccinations of COVID.

Neutral:Statement without clear position for supporting or rejecting vaccinations, such as related questions about injection age or address.

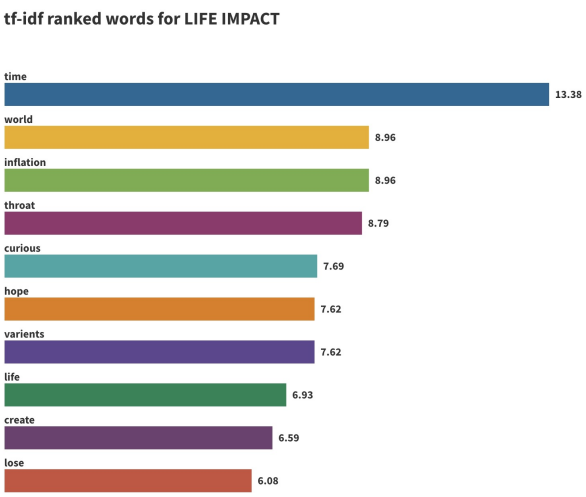
Total related posts of Topic 3 in 1000 posts: 402
Positive: 226 (56.22 percent)
Negative: 114 (28.36 percent)
Neutral: 62 (15.42 percent)

Top 10 words with highest tf-idf scores(in descending order):
"vaccine", "vaccinated", "booked", "vaccines", "booster", "kids", "available", "shot", "dose", "follow"



We could see that the positive sentiment for vaccinations, which refer to support getting vaccinated, is more than double the negative sentiment of vaccination.

The relative engagements from Top 10 words are most about: booking the vaccination/ vaccination for kids / vaccination age available/ vaccination available.



We could see that tweets with negative sentiment are about 3 times more than those with positive sentiment. In other words, the response to life impact is generally negative, which refers to depression and negative attitude during the pandemic.

The relative engagements from Top 10 words are most about: Time spending/ The journey in worldwide/ The inflation/ The throat symptoms/Interest and wishes (hope) / The creation and loss about people's life under COVID

Topic 4: Life Impact

This topic is related to the impact of COVID on people's life, including mental impact, COVID symptoms, influence on business activities and change of daily-life.

Positive: Posts expressing happiness or pleasure under the COVID,with an optimistic attitude to face these life impacts in different aspects.

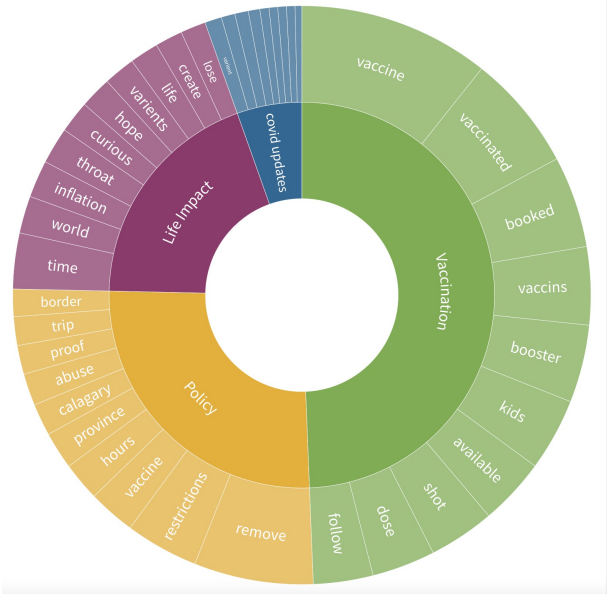
Negative:Posts expressing sadness or sorrow under the COVID, with a negative attitude to these life impacts.

Neutral: Statement of life impacts without emotional trending.

Total related posts of Topic 4 in 1000 posts: 322
positive: 62 (19.25 percent)
negative: 197 (61.18 percent)
neutral: 63 (19.57 percent)

Top 10 words with highest tf-idf scores(in descending order):
"time", "world", "inflation", "throat", "curious", "hope", "variants", "life", "create", "losing"

Overall visualization of 4 topics and corresponding top words



Discussion

Topic 1: COVID Updates

From the results of Topic 1, we could see that people were generally negative regarding the COVID condition. Top word is 'died' in this topic which refers to people paying the most attention to the mortality of COVID, the dead patients cases. People also talked about "variant" and "coverage" frequently as well, they used a lot of negative adjective words like "worse" to describe, which implies that people were worried and anxious about the current epidemic situation, that they consider the condition of the epidemic is getting worse.

Topic 2: Policy

From results of Topic 2, we could see that people are generally dissatisfied with the current policy under epidemic.
Most of sentiments in this topic related discussions are negative. Two most popular words under the topic of policy are "remove" and 'restrictions', indicating that people want to remove the current policies that restrict their freedom of life. We also noticed the word "abuse", which people use to comment on these policies, indicating their dissatisfaction. "Trip" and "border" is also a top related concern, people are not optimistic about the travel policy and border restriction, probably due to the emergence of new variants. ("variant" in the top 10 words of COVID updates and life impact because during the time period of our collection, new variants 'omicron' were just reported).

Topic 3: Vaccination

From results of Topic 3, we could see that people generally have a positive attitude to support vaccination.
People talked a lot about vaccination for kids and boosters, which implies that people were generally willing to get vaccination to protect themselves and their kids. Also, people used "booked"/"available" frequently, this indicates that they actually went for getting vaccinated. Therefore, we could conclude that people were most positive to get vaccination.

Topic 4: Life Impact

From results of Topic 4, we could see that the most discussed topics related to life impact are "time" and "world" and the majority of people's attitude is negative, we can tell that it is a difficult time We can tell that it is a difficult time to throughout the world. "Inflation" is also a topic with a high degree of participation, and we can see that the price increase has a great impact on people's lives."loss" is another word that indicates many people who have suffered setbacks in their lives and probably in their careers because of the pandemic. Overall the pandemic has had a negative impact on people's lives, and people tend to discuss aspects of their lives in a negative way. It is worth noting that the word "hope" appears in the engagement of life impact. Despite the negative attitude towards the impact of the pandemic on people's lives, there is still hope and expectation that the pandemic will end soon.

Overall

From Result we can see that in topics 1, 2 and 4,where people responding to reports, policy and life-impact, their proportion of negative sentiment are all several times larger than positive sentiment proportion, which reveals that people's attitude under COVID are impacted negatively in most aspects, which refers to depression and negative attitude during the pandemic.
Therefore, we conclude that People's response to the pandemic is most negative.

From Result about Topic3 and overall, comparing the positive sentiment and negative sentiment for vaccinations topic can obviously reveal that much more people are encouraged to get vaccinations.Also, vaccination is the only topic that people have an larger proportion of positive sentiment than negative sentiment, which implies that people feel pressure from other aspects through updates/policy/life impacts related to COVID, so they put more expectation on the vaccination. It reveals people are looking forward to ending the epidemic of COVID through vaccination.
Therefore, we conclude that People's response to vaccination is most positive.

Group member contributions

Yichen Huang:
Tweets collection (equally shared),data annotation (equally shared), topic design (equally shared), tf-idf score calculation, data visualization, report writing (data, results, discussion Topic1/Topic2/Conclusion)
Wenyang Wang:
Tweets collection (equally shared), data annotation (equally shared), topic design (equally shared), report writing (introduction, methods, discussion Topic3/Conclusion)
Feiyu Guo:
Tweets collection (equally shared), data annotation (equally shared), topic design (equally shared), report writing (introduction, discussion Topic4/Conclusion)

Reference

Graphs are made by website:
https://app.flourish.studio/