Covid and Vaccine Related Tweet Analysis
Final Project Report

Hamza Usmani, Himanshu Ranka, Yufeng Wang

Introduction

In this project we wanted to get an understanding of the discussions currently happening around COVID in English Speaking Twitter with a special focus on the causes of vaccine hesitancy. Our dataset consisted of COVID related tweets extracted through the twitter API using specific filters. We spent special attention on extracting salient topics to annotate our dataset. We made sure the topics are broad enough to encapsulate all of our dataset, yet narrow enough that they give a concrete impression of what kinds of discussions are being had while making sure each tweet successfully fits into a unique category.

We then used various means of analysis to judge what the data says. Our results show the relative engagement with each topic, the sentiments within each topic and the top tfidf words of each topic. Each of these helps to better understand the current discussions happening on social media as well as allow our client to address the right areas when it comes to addressing vaccine hesitancy.

Data

We collected 1000 tweets over the course of 3 days; 200 on the first day, 500 on the second, and 300 on the third. We chose to look at only tweets containing the words “covid”, “coronavirus”, “vaccine”, “vaccination”, or the name of a vaccine brand such as “Moderna” or “Pfizer”. To minimize redundancy and avoid large threads of discussion, we excluded both retweets, quotes, and replies. We also discarded any tweet that was not in English because translation tools are often inaccurate and unreliable.

We carried out the collection using Python to access Twitter’s stream API, through which we could receive new tweets in real-time. We set the appropriate filters and terminated the program upon having collected the desired number of tweets. Before annotation, we chose to encode the dataset in ASCII and then decode it back to plaintext. This allowed us to remove problematic characters such as emojis and symbols that would become corrupted when parsed by editors such as VS Code or Excel.

Methods

Typology

As we were especially concerned with vaccine hesitancy we specifically designed our typology with categories that capture all possible topics of discussion around Vaccines. More specifically these categories were medical vaccine (mv), political vaccine (pv), and general vaccine (gv) Within these categories we measured sentiment as whether the tweet could be construed as an argument in favor of vaccines (positive), against them (negative), or neither (neutral). We felt that this would be the best way to capture the cases of vaccine hesitancy as well as understand the reasoning behind them. Additionally we believed this split would help us provide the best results to our client.

For every other category measuring sentiment as a stance against a particular topic wouldn’t be particularly useful as the categories are broad enough to contain

different types of opinions. E.g., a tweet about the negative side-effects of covid but against masking mandates can't be easily classified. Furthermore, subdividing the categories into more specific topics would quickly lead to exponentially many categories and would lead to overlap.

Thus, we decided to measure sentiment via tonality, i.e., whether the tweet leaves the reader with a positive, negative or neutral feeling. We believe this to be useful to understand the current climate of public discourse, as these have an impact on the psyche of the individual as well as society as a whole. Further more data split in this way would help us gain insights into the general sentiment behind covid discussion (within a category or as a whole).

Lastly since no method of data collection is perfect, inevitably there were some tweets that matched our filters but weren't related to covid, even tangentially. Thus the 'other' category was used. These tweets were in a foreign language, used our sought hashtags such as #covid or #vaccination only to gain visibility, or simply made no sense.

Annotation and Analysis

We went through three iterations of open coding on 200 tweets until we settled on our final typology. Our first iteration resulted in categories that were too broad. This made it difficult to sort tweets into individual categories uniformly. Our second typography was too narrow, and too many tweets did not fit into any category. We then found a middle point between the 2 that encapsulated all the important information we wanted to collect while keeping out of topic tweets to less than 1% in open coding and less than 5% in the final annotation.

Then each of us used the typology to annotate all 1000 tweets individually in an attempt to weed out any individual bias or errors that we may have made.

To create a set of final annotations, we then wrote a script that used a simple majority voting on each tweet to pick the best category. Any tweet where all 3 of us disagreed was then annotated again with further discussion among the group giving us a rather accurate annotation of the tweets. Analysing our annotations we noticed a category that we expected of being of more value (health h) had less than 25 tweets. This was a fault of our initial typology that was corrected, and the tweets redistributed into existing categories.

We then ran Cohen's kappa on all 3 annotations, and then on each individual annotation set and the final set. The results of the same are captured in Table 1. As observed the final set both matched the annotations of each individual much better while standardizing the data increasing our confidence in our annotations. The raw kappa for all 3 sets on the other hand was much lower. The reason for this variation in annotation among each individual resulted from each person's perception of each category and tweet which we would naturally expect to vary.

Annotations compared	κ
Final x Hamza	0.69
Final x Himanshu	0.56
Final x Yufeng	0.57
Hamza x Himanshu x Yufeng	0.27

Table 1: Kappa scores

We finally computed the TF-IDF scores of each category while disregarding stop words, words that appeared too infrequently, or words that contained irregular characters with the intention of recognizing the most prominent topics of discussion in each category. The top words by TF-IDF gave us a good indication of the most prominent topics for discourse in each category and each sentiment.

Results

The typology we developed along with its reasoning is as follows:

- **Medical Vaccine discussion – mv**
  - Discussions on effectiveness of the vaccines
  - Discussion on side effects of vaccines
  - Valid reason for hesitancy (People who have disorders, asthma, natural immunization etc.
- **Political Vaccine discussion - pv**
  - Debatable reasoning (my body my choice, freedom, vaccine mandates, ideological differences)

- **General Vaccine Discussion - gv**
  - Anything that is not medical or political relating to the vaccine
  - Ex: 'I like pfizer over moderna', 'I just heard someone talk about the vaccine for an hour'
- **Political Covid/Non-Vaccine related discussion - pc**
  - Non-Vaccine related politics
  - Masks/Distancing etc
  - Ideological debates
- **Covid discourse - cd**
  - General discourse regarding covid, that doesn't take any particular stances
  - Vaccine categories take precedence in case of overlap
- **Covid and vaccination Statistics and developments – cs**
  - Recent news and policies regarding covid
  - Plain stats about covid
  - How covid affects the economy and the individual
  - Gives us an impression of what kind of data the public is being exposed to
- **Comedy/Satire - c**
  - Comedic discussion that does not add value to the conversion/ share any opinions. In other words, comedy for the sake of comedy
  - If tweet is comedic but shares a point of view, is added to 'covid discourse' instead
- **Other – ot**
  - Any tweet not relating to covid but that got past the filters.

The results of our annotations, sentiment analysis and TF-IDF scripts are presented in Figure 1 and Figure 2 respectively in the Discussion section of the report along with the insights we have gained from them.

Discussion

Firstly we can deduce the relative engagement of each topic by measuring their relative counts. 'Covid discourse and stats' (cs) had the most engagement representing 24 percent of all tweets followed by 'Covid developments' (cd) at 21 percent. This makes sense as they were the most general of the categories and captured a vast array of tweets. This just tells us that around half of the tweets people see is just general talk and news. In addition, the positive and negative sentiment is almost equally divided and the neutral sentiment is around 50 percent for both telling us that there's no prevailing sentiment in the general covid discourse currently which was a rather surprising result considering the general negative sentiment we all hold towards covid individually.

In addition around 30 percent of all tweets were related to the vaccine (mv , gv and pv) which tells us that vaccinations are a hot topic of discussion among the populace currently. Furthermore, Political Vaccine discussion was the most prominent category out of these three at 13 percent of all tweets, even more than General vaccine discussion with 10. One possible reason for this is as we were collecting the data, news of the novel Omicron variant was being published, a lot of it relating to how the old vaccinations were ineffective against this variant, leading to a lot of debates on the usefulness of vaccines and necessity of vaccine mandates. Furthermore A lot of governments are in the process of making a decision on the viability of vaccine mandates, the results of which also led to a lot of political discussion on the same amongst the general populace. This also explains why the majority sentiment is negative in the political and medical vaccine categories, representing the frustration and lack of trust the public is feeling at this particular moment about vaccinations and authority figures making vaccine related decisions.

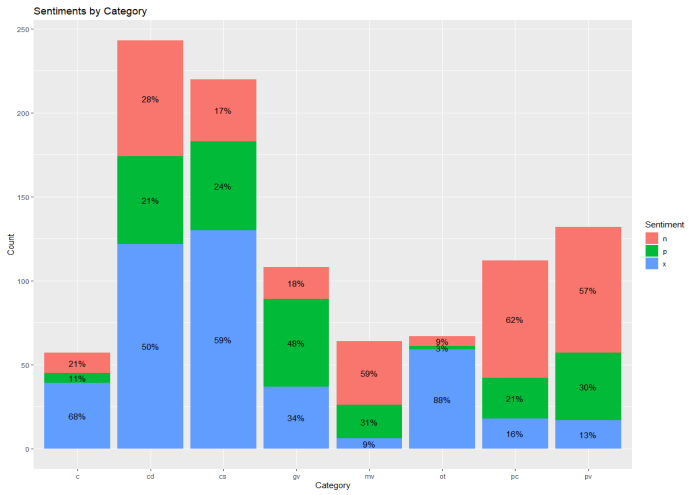

Figure 1: The total counts of each category and the relative percentages of sentiments. n is "negative", p is "positive", and x is "neutral".

The 'Political Covid' (pc) also showed similar negative sentiment, which could also be a result of the developing omicron situation, which could represent that at this point in the pandemic the public is starting to doubt the effectiveness of covid policies such as lockdown and

distancing, as none of that has prevented the covid situation from developing and continually getting worse. This also shows the frustration individuals are experiencing with the cycle of strict and relaxed policies and the lack of normalcy in everyday life.

It should be noted however that these were only results of tweets collected over a three day window, which corresponded to the exact time news of the omicron variant was being published. Thus it would be irresponsible to say that the sentiments and engagement shown by our result accurately reflect the discussions in general.

We also had a comedy category that mainly consisted of jokes about the pandemic. Determining the tonality of these tweets was difficult, and they didn't contribute much to the discussion. Given more time to refine the system, we would have removed this category.

According to the TF-IDF results, many users discussing the medical aspects of vaccines expressed doubt about the vaccine manufacturers, especially their CEOs and relentless promotion of their boosters. This suggests the intrinsic suspicion of the manufacturers exploiting the pandemic to profit from their products. The majority of these tweets (59%) were critical of the vaccines by citing their adverse effects and inefficacy especially in regards to the Omicron variant.

Users discussing the political aspects of vaccines were interested in the policies of South Africa and particularly of its president, Ramaphosa, following the Omicron outbreak originally reported in this region. The majority of tweets in this category were against vaccine mandates (57%), citing the violation of personal freedoms and even claiming that negative effects of the vaccines were intentional and well orchestrated.

Users discussing the statistics such as case numbers and deaths were keenly interested in the situation in South Africa and the UK. People were concerned about the outbreak in South Africa and its impacts on other countries such as the UK, which under Boris Johnson has been criticized for refusing to implement stricter travel restrictions to combat the new strain. Some applauded China's pledge to offer 1 billion vaccine doses to Africa and others claimed that China would use these vaccines to spawn a new variant. The political nature of discussion around covid did seem to discuss inequity among countries with regards to the fact that the Omicron outbreak was initially observed in Southern Africa as well as the discrepancy by the UN in naming it a variant of concern even though symptoms were reported to be milder than most preceding variants.

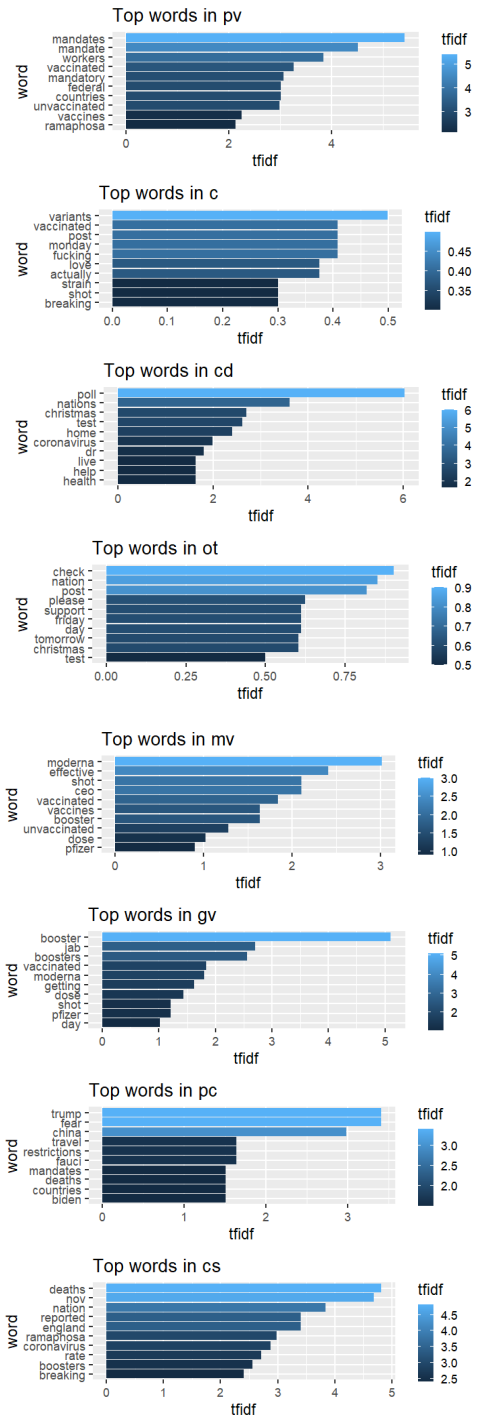

Figure 2: TF-IDF results per category

Improvements

Working on this project was a great learning experience for the whole group. The iterative process of coming up with a typology as well as the analysis of the results provided great insight into areas of improvement and had time permitted, we could have made several improvements to the project.

After annotation and analysis we observed that there was scope for tightening the categories chosen in the typology giving more accurate insights into data.

There was also great scope for a recursive analysis of results especially in the case of TF-IDF. Here, extremely common words without consequence or lexically different words that have similar meaning could be removed or grouped together giving us a more diverse insight into the topics being discussed. Further a TF-IDF conducted on n-grams may have given us insight into topics that were lost by considering only single words.

Furthermore an annotation with subcategories would have allowed us to group tweets and topics of discussion more robustly giving us a tree of discourse as well areas of interest a lot more accurately.

Contributions

- Hamza Usmani (260858598)
    - Wrote the intro and parts of the discussion, results and methods.
    - Helped in Data Collection/annotation
    - Wrote scripts to collect relative sentiment/ categories data
    - Helped in creating the typology
    - Contributed to every part of the project

- Yufeng Wang (260855204)
    - Decided on filters to use
    - Wrote script to collect tweets
    - Wrote script to clean and standardize tweets
    - Wrote TFIDF script, analyzed TFIDF results
    - Created TFIDF plot
    - Created stacked bar plot of sentiments by category
    - Wrote “Data” section of report

- Himanshu Ranka (260849076)
    - Wrote the script for running Cohen’s Kappa on the annotations
    - Worked with the group to come up with a good typology
    - Helped in annotating data and cleaning annotated data
    - Worked on several parts of the report, as well as on editing the report to increase coherence and flow
    - Made sure the report fit the AAAI formats