Analyzing Online Discourse on COVID-19 Vaccines: Evidence from Twitter

Garda Ramadhito, Jianing (Jonathan) Wang, Leo Liu, Zhengze Hou

QMSSG5067: Natural Language Processing for the Social Sciences

Columbia University

December 17, 2021

Abstract

To understand the public's opinion and perceptions on COVID-19 vaccines, natural language processing is utilized to assess COVID-19 All Vaccines Tweets dataset. The procedure includes text data preprocessing, word cloud generation, time series analysis, comparison using data visualization. The methods include sentiment analysis through TextBlob and VADER , word clouds, TF-IDF and time series. We find that Western-made and mRNA vaccines tend to induce tweets with higher positive sentiment than non-Western, non-mRNA ones. The Chinese vaccines have a slightly lower sentiment score on Twitter than non-Chinese vaccines. Ultimately, we can prove that there are significant differences in online conversations between different types of vaccines.

Introduction

In a world of which assets and information are ubiquitously digitized, people express their opinion through social media all the time. Approximately over 5,000 tweets are sent out per second everyday. With such a rapid generation rate, social media information is gradually becoming a significant index for decision-making processes. Ever since its emergence in March, 2020, COVID-19 has become the central topic on all kinds of social media platforms.

As COVID-19 continues to greatly impact people in all aspects of life, administrators and researchers have been implementing public health prevention measures to control the spread of this devastatingly contagious pandemic, including enforcing COVID-19 vaccination requirements. However, the public's opinion and perceptions on COVID-19 vaccines are varied drastically especially on social media platforms. Understanding the variation in people's opinion on COVID-19 can be insightful and useful in implementing vaccination enforcement strategies.

The purpose of this study is to utilize natural language processing to explore and assess the overall sentiment and attitudes toward different types of vaccines. We will be mainly using the tweet posts on COVID-19 posts as the dataset, since tweet posts are generally shorter in length and more opinion-driven. This specific study will be mainly focusing on answering the following research questions:

1. How do discussions about the vaccines differ by different kinds of COVID-19 vaccines?
2. What are the differences in sentiment between Chinese vaccines and non-Chinese vaccines?
3. How do discussions over vaccines vary over time?
4. Does discussion on mRNA vaccines differ from discussion on non-mRNA vaccines?

The analysis procedure will include text data preprocessing, word cloud generation, time series analysis, comparison using data visualization. Hopefully the study can provide some insights on the deviation in people's attitudes towards different types of vaccines, and thus contribute to the future implementation of vaccine policies.

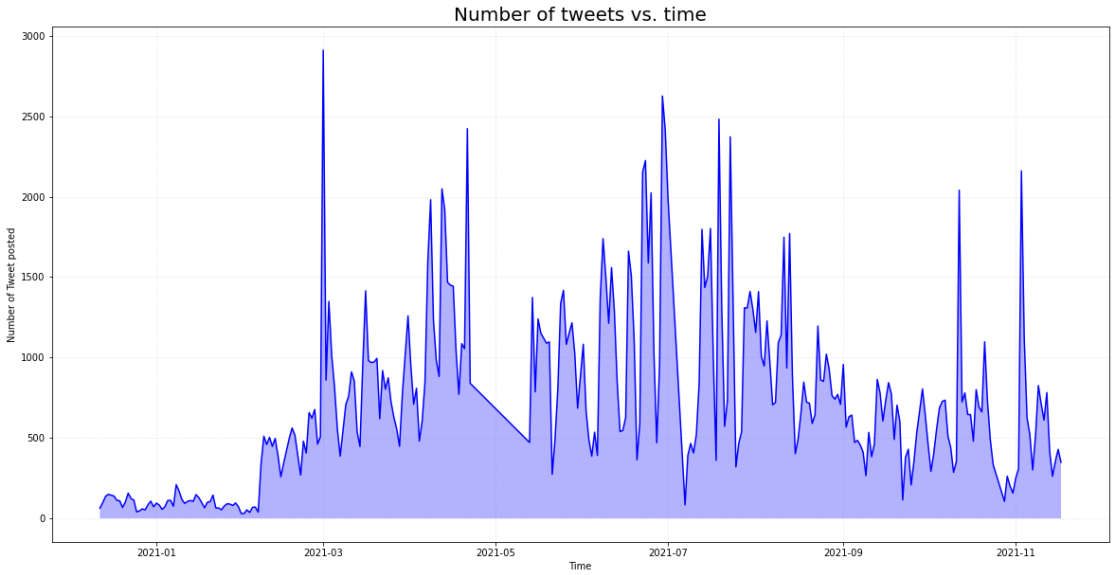
Data Description

The dataset used in this study is the COVID-19 All Vaccines Tweets dataset from Kaggle. The tweets data was gathered using tweepy Python package that extracts text data from the Twitter API. The dataset consists of 16 columns and over 220,000 rows each representing a twitter post that involves discussion on Covid-19 vaccines.

The dataset contains the following fields:

1. Id, a unique identifier number for each post
2. User_name, Twitter username
3. User_Location, the location where the tweet was posted
4. User_description, the personal description of the user
5. User_created, the time this Twitter user is created
6. User_followers, the number of followers
7. User_Friends, the number of users following
8. User_Favorites, number of favorite posts
9. User_verified, Verification Status
10. Date, Date the tweet is posted
11. Text, Text Body
12. Hashtags, a list of hashtags included in the post
13. Source, source of the Tweet
14. Retweets, Number of Retweets
15. Favorites, Number of favorites
16. Retweet Status, if the tweet has been retweeted

The main reason we are using this dataset for the study is that the dataset poster gathered the twitter posts based on the names of different types of COVID-19 vaccines, such as Pfizer, Sinopharm, etc. The data gathering procedure aligns with our research topic, which makes this dataset incredibly useful.



Graph 1: The frequency of tweets plotted against date

Graph 1 above shows the frequency of tweets plotted against the date. We can see the number of tweets on COVID-19 vaccines had a huge surge starting from March 2021. A possible reason for this increase would be the announcement made by J&J on March 1, 2021, saying that the first batch of vaccines would be sent to states and pharmacies and distributed to the public. After that, the number of tweets fluctuated significantly.

The frequency graph only shows one preliminary aspect of the tweet trends on COVID-19 vaccines. In order to make generalizations and conclusions on public attitudes, we have to make considerable modifications to the text body.

Data Preprocessing

After reading in the data comes the crucial step of cleaning and preprocessing. The raw data is messy and unstructured. There are emojis, hashtags, and non-English texts in the columns we are interested in. In order to make the dataset workable, we follow the procedures below:

1. Removing unnecessary columns, fixing column types
2. Noise removal and letter casing
3. Removing stopwords
4. Stemming

Removing unnecessary columns

The 16 columns contained in the original data are id, user name, user location, user description, time when user account was created, number of followers, number of friends, user's number of favorites, if whether or not the user is verified (True or False), date of tweet, tweet

content (text), hashtags, source, number of retweets, tweet's number of favorites, and if whether or not the tweet is a retweet. Most of this information is irrelevant to our research questions. So the columns we decide to keep are id, user location, date, text, hashtag, and retweet. Next, we change the column types to string for user location and hashtag. The id column is a categorical variable, and specific time is excluded from the date column. We are only interested in what day of the month of the year the tweet is tweeted, but not the specific time. The majority of the analysis later will focus on the column 'text', because that is the column with the tweet content.

Noise removal and letter casing

Removing non-English text, punctuations, and emojis is essential for the analysis later. Here we evoke the function 'clean_txt' from the util.my_functions module. This function eliminates anything but alphabets from A to Z, and numbers from 0 to 9. We apply this function to the text, location, and hashtag columns.

Removing stopwords

The nltk library defines a list of stopwords. These stopwords do not contribute to our understanding of the sentiment. We evoke the function "rem_sw" from util.my_functions to eliminate the stopwords in the text column.

Stemming

Stemming is the process of removing affixes in a word to obtain the word stem. For example, skating and skate carry the same meaning so we would like to treat them as one. To achieve this, we use the function "stem_fun" in the util.my_functions module. This function uses the Porter Stemmer technique. It chops off the end of any word.

Methodology

Sentiment Analysis

The main sentiment analysis methods utilized in this study are Valence Aware Dictionary and sEntimentReasoner (VADER) and Textblob. Both methods use natural language processing and text analysis to systematically assign a polarity score that indicates the attitude of the text body towards a certain topic. [1]

VADER is a model used for sentiment analysis based on lexicons of sentiment-related words to determine whether a text body is generally positive, negative or neutral in terms of attitude. VADER generally creates three columns each shows the positive score, negative score and neutral score. After creating these metrics, a compound score will be calculated based on

[1] Peru, B, "K-Means vs Vader Vs TextBlob Sentiment Analysis", https://www.kaggle.com/accountstatus/k-means-vs-vader-vs-textblob-sentiment-analysis, accessed on Dec 17, 2021

the sum of all the lexicon ratings normalized to a -1 to 1 scale, where -1 indicates extremely negative and 1 indicates extremely positive. We are categorizing the sentiment of each of the tweets into positive tweets with positive compound scores, neutral tweets with 0 compound scores, and negative tweets with negative compound scores.

The other method is Textblob. Textblob is a python library built upon the NLTK architecture that performs multiple NLP tasks such as sentiment analysis, noun phrase extraction, etc. The Textblob library automatically calculates the polarity score and subjectivity score based on the emotions expressed in the text. The polarity score is similar to the compound score from VADER, where -1 indicates extremely negative sentiment and 1 indicates extremely positive sentiment. Subjectivity indicates the expression of opinions or emotions from the text.

We utilized both methods to create sentiment metrics for sentiment analysis. The result shows that there are 1,189 tweets that have different polarity scores. However, considering there are 224,249 total tweets included in the dataset, only 0.5% of the tweets are having different polarity scores.

Keywords extraction and word cloud generation

Extracting keywords from the text body is useful in analyzing the general features and trends of the vaccine tweets. The main method utilized to extract keywords is to calculate the frequency of words and visualize the prominence of each keyword using WordCloud library.

The WordCloud library creates a visualization of clustered keywords. The words in the word cloud generated have varied font sizes and color. A bigger word font generally means higher prominence level compared to other words included in the plot. The prominence, in this specific study, is determined by the frequency of each word. [2]

Time Series Analysis

Time series analysis can be a useful method to examine the change in frequency and the general trend of the topics of the Tweets. The main methodology of analysis is to visualize the time series data and to identify potential trends and changes in the target variables.

A quantitative approach to analyze the time series data is to incorporate the Autoregressive Moving Average model to describe the weakly stationary stochastic time series. However, the central topic of this study is the sentiment analysis on the tweet text body, instead of the time series analysis on panel data. Therefore, we didn't include the time series analysis in the study.

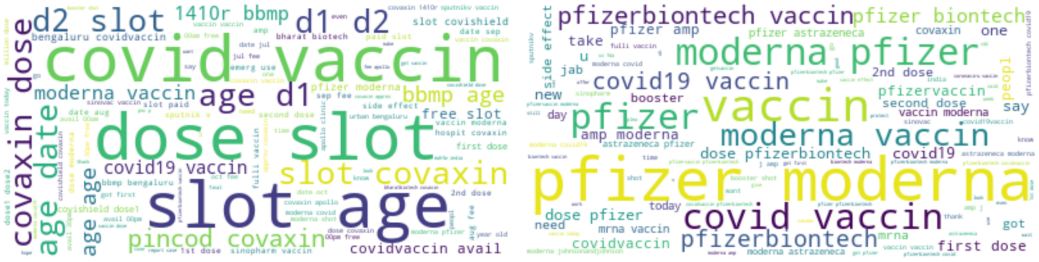
[2] Shah, R, "How to Build Word Cloud in Python?", Analytics Vidhya, May 20, 2021, Accessed Dec 17, 2020

Findings

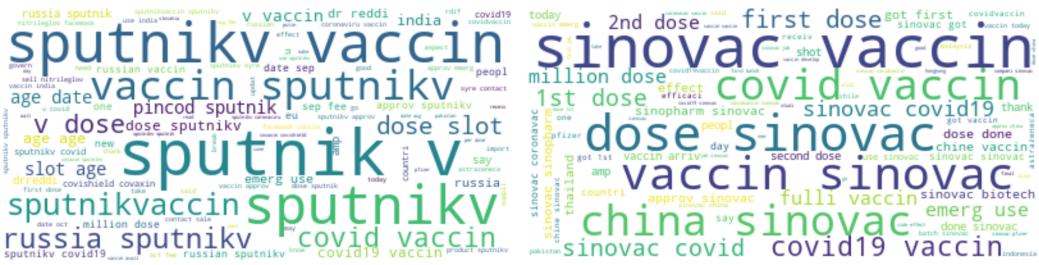
How do discussions about the vaccines differ by different kinds of COVID-19 vaccines?

After the outbreak of Covid-19 pandemic, the experts and the majority of people view the vaccines as the key to overcoming the virus. Meanwhile, the high cost of developing a qualified type of vaccines means that only the developed countries and great power can afford the expense. Moreover, there is also controversy about these vaccines' effectiveness and allocation.

Thus, in the first part, the opinion on twitter about five major types of vaccines(three western vaccines, one Chinese vaccine incorporating two different brands, one Russian vaccine) are categorized into different groups. The meaningful findings are then picked out to analyze. Generally speaking, what and where they are talking about are the attractive points.



Graph 2: The frequency of tweets about all kinds of vaccines(left side)
Graph 3: The frequency of tweets about Pfizer vaccine(right side)



Graph 4: The frequency of tweets about SputnikV vaccine(left side)
Graph 5: The frequency of tweets about Chinese vaccine(right side)

The first group of graphs are word cloud pictures that show the most common keywords used in tweets on the selected Pfizer and SputnikV vaccines. And there is also a word cloud of all the tweets of vaccines as a comparison.

The first interesting thing is that people focus on "age" in the general word cloud, which shows the concern about side effects on elderly people. However, the topics about specific vaccines don't show that strong concern, which perhaps means that this concern is a common issue and not biased to some specific types of vaccines.

The second finding is that whether the word cloud of Pfizer or Moderna, AZ has a big discussion volume about each other. People tend to compare these vaccines about their effectiveness or side effects. However, the Russian vaccine's word cloud doesn't duplicate this kind of relationship. So are the Chinese vaccines. One can assume that the different vaccinated positions and the difference between mRNA and inactivated vaccines are the main reasons. But this gap also shows a divided perspective between Western and non-Western society.

To extend the findings above mentioned, the TextBlob is utilized to gain a sentiment analysis on tweets about different types of vaccines.

Total	Amount	Percentage
neutral	12,314	56.28
positive	6,284	28.72
negative	3,280	14.99

Table 1: Proportion of sentiment of tweets mentioning Pfizer

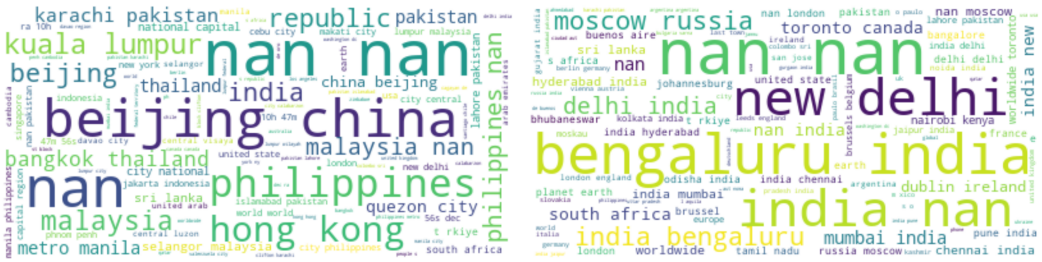
Total	Amount	Percentage
neutral	11,938	66.86
positive	4,051	22.69
negative	1,866	10.45

Table 2: Proportion of sentiment of tweets mentioning SputnikV

From table 1 and 2, it is obvious that people are more positive and also more likely to express their opinion on the Pfizer vaccine. The results of AZ and Moderna show similar structure. And the neutral sentiment is more dominant in the discussion of Russian vaccines (the Chinese vaccines' data has very close results), which implies a kind of ignorance or uncare about things unrelated to Western society.

Graph 6: The user_location frequency of tweets about AZ vaccine(left side)

Graph 7: The user_location frequency of tweets about pfizer vaccine(right side)



Graph 8: The user_location frequency of tweets about Chinese vaccine(left side)
Graph 9: The user_location frequency of tweets about Russian vaccine(right side)

The third finding is about the location where people talk about vaccines. The initial understanding is that the people from the countries who developed and manufactured vaccines are most likely to talk about this type of vaccines, such as Britain and America as the graph 6 and 7 has shown. China and Russia aren't English speaking countries, but China's huge population still leads Beijing and Hong Kong to the top of user_location. But India and cities like New Delhi and Bengaluru are the top of the locations where people talk about SputnikV vaccine, which is confusing. The possible reason is that in May of 2021, India authorized the use of Russian vaccines, then in July, India was authorized to manufacture the SputnikV. The high volume of discussion in India is probably a result of the public health cooperation between India and Russia, which shows the "vaccine policy" of Russia.

Actually, the location at least reveals the effectiveness of international vaccine policy of different countries to different parts of the world. Chinese vaccines also has a high volume of discussion in the Philippines, Malaysia, Thailand and Pakistan, which shows China tends to donate vaccines to Southeast Asia. AZ vaccines have an influence in Germany, Canada, Australia and Sri lanka. While Pfizer is discussed mainly in Canada, Australia, India, Malaysia and the United Arab Emirates.

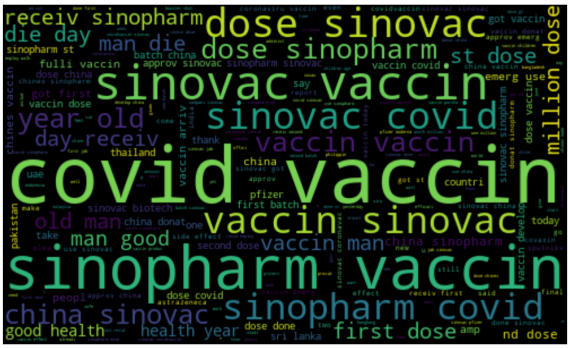
The user_location's amount demonstrates the different donation sphere, which is surely a sign of international relationship and geographic distance. However, the absence of Africa and South America is worthy of concern. Part of the reason is that they are not English speaking areas, but the main reason might be the "vaccine nationalism" criticized by the World Health Organization. The vaccine nationalism is an important source of unceasing variant viruses such as the Omicron recently discovered in South Africa.

What are the differences in sentiment between Chinese vaccines and non-Chinese vaccines?

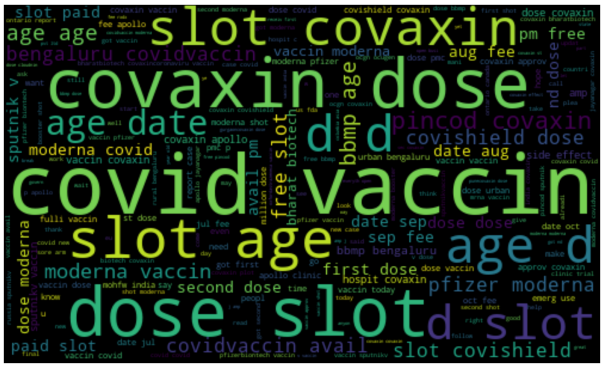
COVID-19 is a global pandemic, and therefore COVID-19 is a huge challenge to the whole world and to the entire human race. However, the difference in political ideologies among different countries makes the pandemic a key component in political conflicts and discussions. Vaccine, the most reliable and most effective method to control the spread of the pandemic, also suffers from the same political controversies. Which country produces the best vaccine? Are the vaccines produced by Chinese pharmaceutical companies reliable? Questions like these are constantly brought up by mainstream media and news companies.

The main research goal for this section is to see if the public's opinions are biased against Chinese vaccines specifically. The research disregards the real effectiveness of Chinese vaccines.

The word cloud for Sinopharm



The word cloud for All other Vaccines



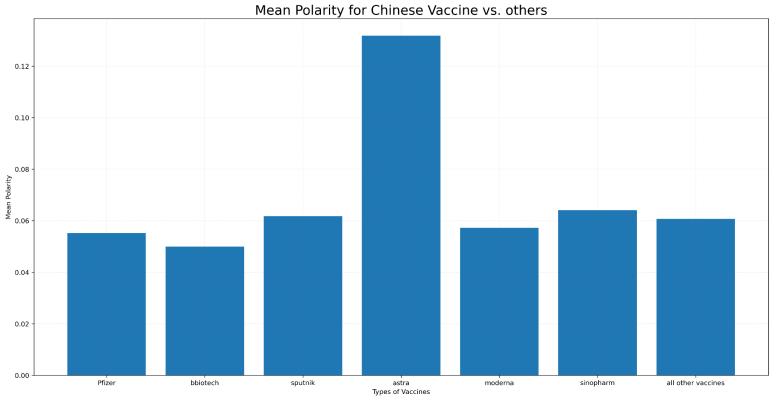
Graph 10: Word clouds comparing keywords within Sinopharm tweets and non-Sinopharm tweets

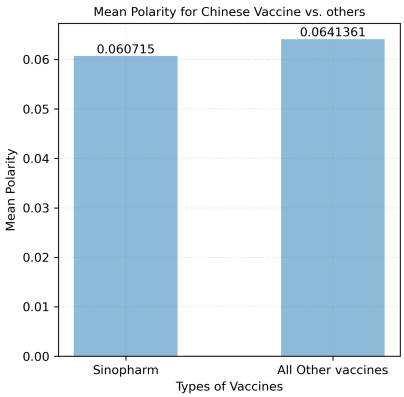
The graph 10 is a word cloud that shows the most common keywords used in tweets on the Chinese vaccine, Sinopharm. From graph 10, we noticed some concerning phrases such as "die day", "man die". These words are considered to be a huge indication of negative attitudes towards Chinese vaccines. From the word cloud, we can make a preliminary generalization that the public are having concerns about the reliability of Chinese vaccines.

In order to approach this issue from a more quantitative perspective, we utilized TextBlob to create the sentiment value based on the content of each data. TextBlob is a Python library built upon nltk, that can be used to generate polarity scores for sentiment analysis. The polarity score generated by Textblob can be used as a good inference for a generalization of attitude. The higher the polarity score, the more positive the text body shows.

In order to compare the polarity for tweets on Chinese vaccines and those on non-Chinese vaccines, the dataset is divided into two subsets. Two reference variables are defined to filter the dataset. The reference variables are used as the standard to filter the data. Generally, all tweets that include the keywords "sinovac", "sinopharm", "chinese", "china", or "sino" are considered to be tweets on Chinese vaccines. Tweets that include keywords like "pfizer", "covax", "modern" are categorized as tweets on other vaccines. (A complete list of reference word will be included in the Appendix I)

Then the mean polarity scores are calculated for the two groups, tweets on sinopharm, and tweets on all other vaccines. The results are shown in the graph 11 shown below.





Graph 11: The Mean polarity across all types of vaccines (Left)
Graph 12: The Mean polarity for Sinopharm vs. all other vaccines (Right)

Graph 11 shows the mean polarity across all types of covid vaccines. We can see that astra has the highest mean polarity compared to all other vaccines, and for the other vaccines, including sinopharm, have generally the same mean polarity in terms of tweet sentiments.

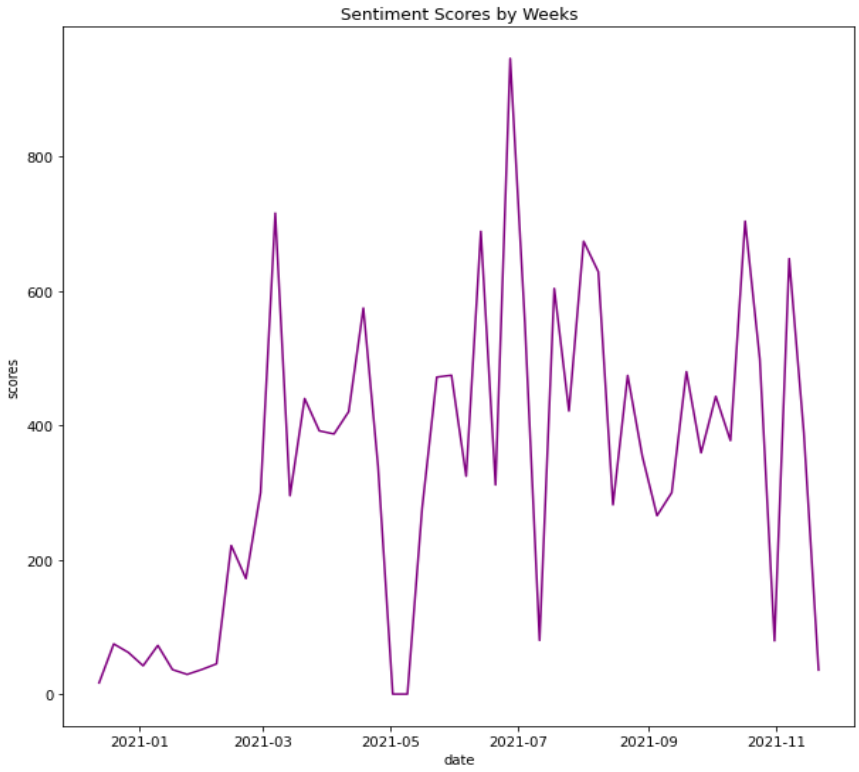
The comparison between Chinese vaccines and all other vaccines as a whole is more clearly visualized in Graph 12. We can see that the mean polarity scores for tweets on both types of vaccines are similar quantitatively, with the mean polarity score for tweets on sinopharm slightly lower. The result suggests that the general opinion on Chinese vaccines is slightly more negative compared to all other vaccines. This is a somewhat concerning conclusion, as there is evidence that the public's opinion is biased based on the type of vaccine. However, we cannot make any arbitrary conclusions suggesting that the public opinion is shifted by the influence of mainstream media, since the objective evaluation on the reliability of each vaccine is not in the concerns of this specific study. In order to evaluate the reasoning behind the trend, we have to take the reliability of each vaccine into consideration.

How do discussions over vaccines vary over time?

We are highly interested in seeing how sentiment towards COVID-19 vaccines change in time. We know some people are suspicious towards the notion of a vaccine. Consequently, conspiracy theories started to develop. However, there are also many who believe in science. Frankly, we are not certain what the sentiment trend will look like. Previously, a column of polarity scores was added to the data. The polarity scores are provided by SentimentIntensityAnalyzer via VADER. We also labelled each score; positive if the score is greater than 0, neutral if the score is equal to 0 and negative if the score is less than 0. The labels are stored in the column called 'labels'.

The first step is to create a date index. Date index will be used for the time series analysis. Then, we can sum the polarity score by day, week, month, or year. A large positive sum indicates a more positive sentiment overall, whereas a small positive sum or a large negative sum indicates a less positive sentiment in general. We find grouping polarity scores by

weeks the best because it shows a clear trend of sentiment changes in time, and it is not too messy to examine. The figure below is generated by matplotlib.pyplot.



Graph 13: Sentiment scores over time

From the figure, we see an interesting trend. The sentiment is low at the beginning of 2021. This was when the vaccines were first developed. People were suspicious towards COVID-19 vaccines and they worried about the side effects and efficacy. As the vaccination rate increases, the sentiment towards the vaccines becomes more positive. The peak of the positive sentiment happens in July. However, as new variants emerge, people start to question the efficacy of vaccines once again. In November, governments started to recommend the booster shot in order to fight the Omicron variant. The sentiment towards vaccines turns sharply negative.

Does discussion on mRNA vaccines differ from discussion on non-mRNA vaccines?

Discourse over mRNA vaccines is of particular interest because of how novel the technology is, providing ample opportunities for comparison against discussion on vaccines employing more traditional technologies. mRNA vaccines work differently from traditional vaccines. In layman's terms, mRNA vaccines carry over a harmless piece of mRNA from a virus, usually a protein found on the outer layer of the virus, in this case the spike protein from the COVID-19 virus. Encountering these unrecognized spike proteins, cells are then stimulated to produce antibodies without actually encountering the actual virus. This is a significant departure from traditional vaccination in which the inactivated or weakened virus is introduced to

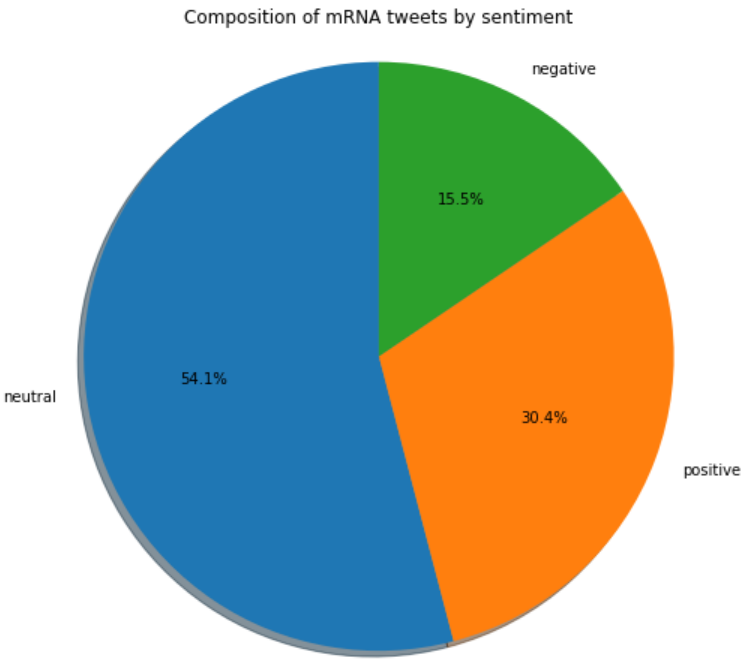
the body's cells.[3] The technology was very promising; the efficacy rate was especially high when results were initially released. The Pfizer/BioNTech vaccine was found to be 95% effective against COVID-19 in the phase 3 trials in 2020.[4] Similarly, the Moderna vaccine was found to be 94% effective in the company's phase 3 trials.[5] And both of these vaccines were developed at an exceptional speed of less than one year.[6]

In turn, the novelty and the efficacy of the mRNA technology have provoked strong reactions on the Pfizer/BioNTech and Moderna vaccines. On one hand, there was excitement and marvel at how fast the vaccines were developed, particularly in late 2020 and early 2021 when COVID-19 cases were at their highest levels.[7] On the other hand, the speed garnered suspicion and skepticism within some circles.[8] In this section, we test two competing hypotheses. First, we hypothesize that mRNA vaccines are more likely to cause skepticism and hesitancy, which induces a lower sentiment score for tweets mentioning the two mRNA vaccines. Second, we hypothesize that mRNA vaccines are more likely to inspire excitement and provide hope, which leads to higher sentiment scores for mRNA vaccines. The null hypothesis would be that there is no substantive and significant difference in the discussions between the two categories.

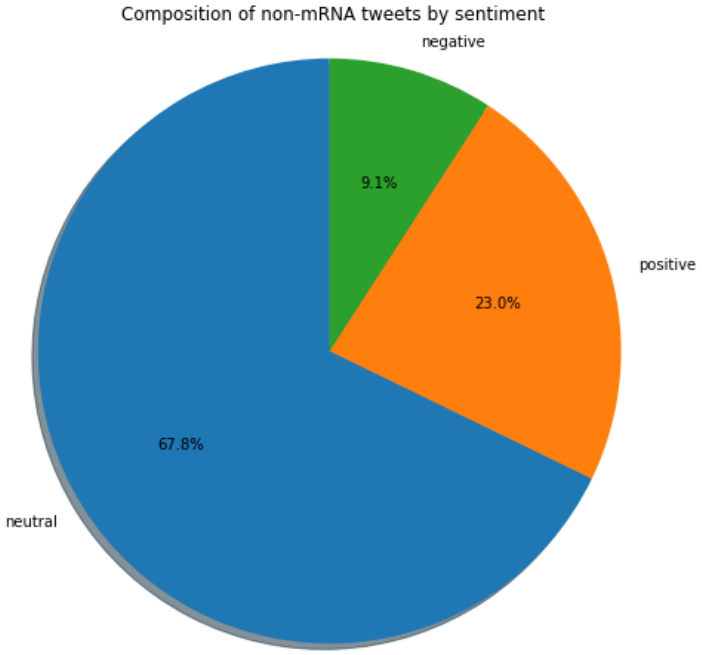
Using the sentiment analysis algorithm VADER, we find that there is indeed a substantial difference between the sentiment of tweets mentioning mRNA vaccines and the sentiment of those mentioning non-mRNA vaccines. The sentiment around mRNA vaccines is significantly more positive than the sentiment around non-mRNA vaccines, which are more neutral. From graph 13 and graph 14, 30.4% of tweets mentioning mRNA vaccines are of positive sentiment, compared to 23.0% in the tweets mentioning non-mRNA vaccines. The non-mRNA tweets, in contrast, have more neutral sentiment compared to mRNA tweets at 67.8% and 54.1%, respectively.

---

[3] "What Are MRNA Vaccines and How Do They Work?: MedlinePlus Genetics," accessed December 17, 2021, https://medlineplus.gov/genetics/understanding/therapy/mrnavaccines/.
[4] "Pfizer and BioNTech Conclude Phase 3 Study of COVID-19 Vaccine Candidate, Meeting All Primary Efficacy Endpoints | Pfizer," accessed December 17, 2021, https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-conclude-phase-3-study-covid-19-vaccine.
[5] Lindsey R. Baden et al., "Efficacy and Safety of the MRNA-1273 SARS-CoV-2 Vaccine," *New England Journal of Medicine* 384, no. 5 (February 4, 2021): 403–16, https://doi.org/10.1056/NEJMoa2035389.
[6] "COVID-19 Vaccine: How Was It Developed so Fast?," November 13, 2021, https://www.medicalnewstoday.com/articles/how-did-we-develop-a-covid-19-vaccine-so-quickly.
[7] Berkeley Lovelace Jr, "Pfizer Says Final Data Analysis Shows Covid Vaccine Is 95% Effective, Plans to Submit to FDA in Days," CNBC, November 18, 2020, https://www.cnbc.com/2020/11/18/coronavirus-pfizer-vaccine-is-95percent-effective-plans-to-submit-to-fda-in-days.html; Denise Grady et al., "F.D.A. Panel Endorses Moderna's Covid-19 Vaccine," *The New York Times*, December 17, 2020, sec. Health, https://www.nytimes.com/2020/12/17/health/covid-vaccine-fda-moderna.html.
[8] Elie Dolgin, "The Tangled History of MRNA Vaccines," *Nature* 597, no. 7876 (September 14, 2021): 318–24, https://doi.org/10.1038/d41586-021-02483-w.



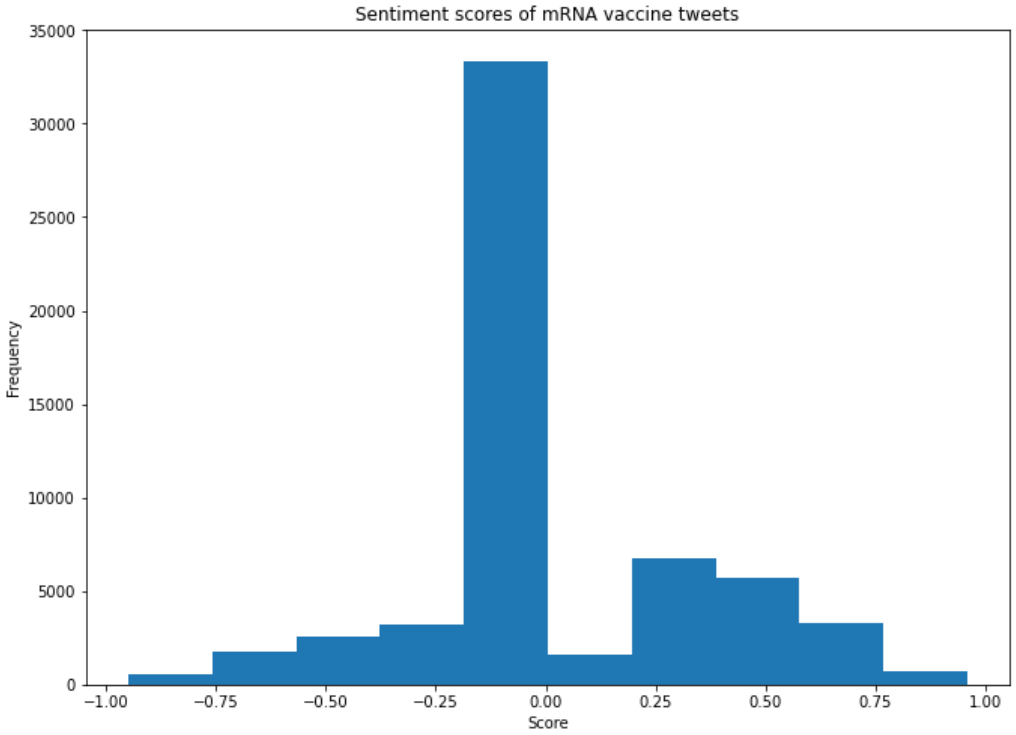
Graph 14: Pie chart of proportion of mRNA tweets by sentiment



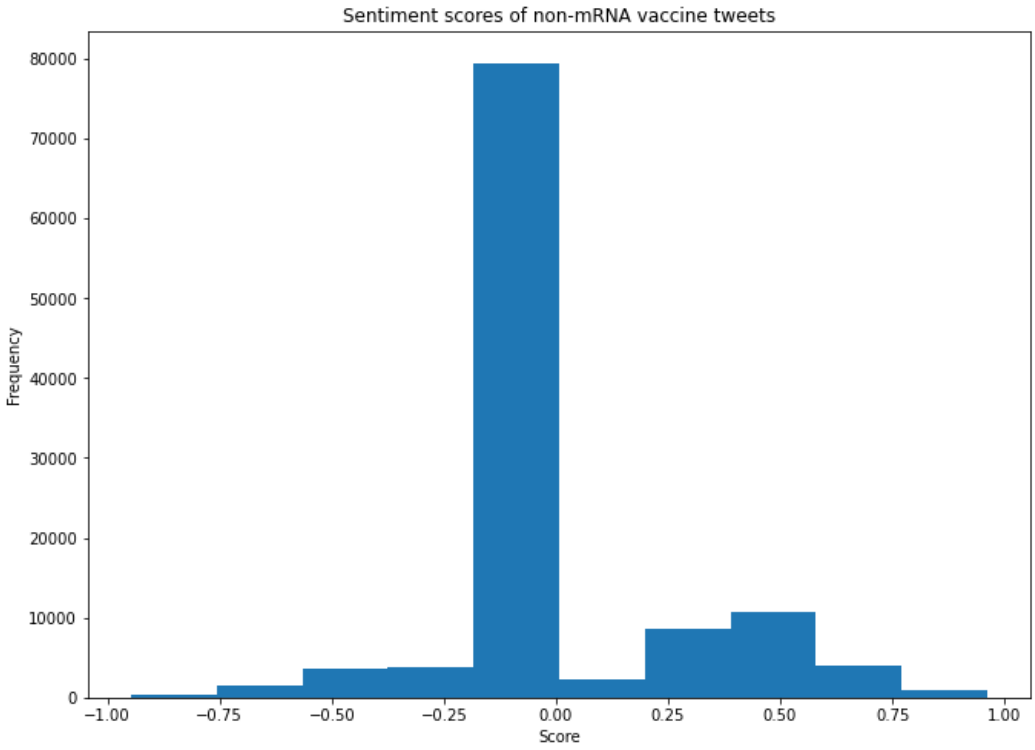
Graph 15: Pie chart of proportion of mRNA tweets by sentiment

The distribution of compound scores for each tweet category also supports the finding above. Compound scores are the sum of positive, neutral and negative sentiment scores normalized between -1 (most negative) and 1 (most positive). Graph 16 shows the distribution of compound scores for tweets mentioning the mRNA vaccines. While most tweets are neutral

based on the compound score (~35000 tweets), the distribution skews to the right, indicating more tweets with positive sentiment. Graph 16 shows the distribution for tweets mentioning non-mRNA vaccines. The distribution is similar to that of mRNA tweets, with most tweets having neutral sentiment (~78000 tweets) and the distribution slightly skewed to the right. However, we can see that the skew is higher in graph 16 than in graph 17, considering the number of tweets in each category. We can deduce that tweets about mRNA vaccines tend to have more positive sentiment compared to tweets about non-mRNA vaccines based on sentiment score.



Graph 16: Distribution of compound sentiment scores of mRNA tweets



Graph 17: Distribution of compound sentiment scores of non-mRNA tweets

The results above lead us to reject the null hypothesis that there is no meaningful difference between online discourse on mRNA vaccines and online discourse on non-mRNA vaccines. Tweets mentioning the mRNA vaccines tend to be significantly more positive than tweets mentioning non-mRNA vaccines. This finding supports the second hypothesis that the mRNA vaccines tend to excite people and lead to more positive conversations online against the first hypothesis that the mRNA vaccines are more likely to cause skepticism and negative sentiment on Twitter. While we did not employ any statistical tests, the evidence is a great starting point for more thorough analysis of online discourse around mRNA vaccines.

Conclusion

This study has shown that online conversations around vaccines do differ by kind, origin and technology, and over time. Western-made vaccines like Pfizer/BioNTech and AstraZeneca tend to induce tweets with higher positive sentiment than non-Western ones. The Chinese vaccines have a slightly lower sentiment score on Twitter than non-Chinese vaccines. Tweets about mRNA vaccines tend to inspire more positive sentiment than tweets about non-mRNA vaccines. We can thus reject the hypothesis there is no difference in online conversations between different types and categories of vaccines. However, our findings are limited to that extent only; there are many more interesting questions and more precise analyses that can be undertaken. To build upon our study, we suggest multiple paths forward. First, we call for statistical tests to be employed to test for significance in the differences. Second, there could be a way to subset the sample to obtain more isolated and precise estimates in differences between groups. Third, more noise can be filtered out by removing spam and tweets by bots. Fourth, one can use specific dictionaries such as medical terms and emotions to measure sentiment more accurately. Lastly, one can analyze additional effects such as networked effects and spatial effects for geolocated tweets. For now, our findings strongly suggest that different vaccines do provoke different reactions online.

References

Baden, Lindsey R., Hana M. El Sahly, Brandon Essink, Karen Kotloff, Sharon Frey, Rick Novak, David Diemert, et al. “Efficacy and Safety of the MRNA-1273 SARS-CoV-2 Vaccine.” *New England Journal of Medicine* 384, no. 5 (February 4, 2021): 403–16. https://doi.org/10.1056/NEJMoa2035389.

“COVID-19 Vaccine: How Was It Developed so Fast?,” November 13, 2021. https://www.medicalnewstoday.com/articles/how-did-we-develop-a-covid-19-vaccine-so-quickly.

Dolgin, Elie. “The Tangled History of MRNA Vaccines.” *Nature* 597, no. 7876 (September 14, 2021): 318–24. https://doi.org/10.1038/d41586-021-02483-w.

Grady, Denise, Abby Goodnough, Carl Zimmer, and Katherine J. Wu. “F.D.A. Panel Endorses Moderna’s Covid-19 Vaccine.” *The New York Times*, December 17, 2020, sec. Health. https://www.nytimes.com/2020/12/17/health/covid-vaccine-fda-moderna.html.

Jr, Berkeley Lovelace. “Pfizer Says Final Data Analysis Shows Covid Vaccine Is 95% Effective, Plans to Submit to FDA in Days.” CNBC, November 18, 2020. https://www.cnbc.com/2020/11/18/coronavirus-pfizer-vaccine-is-95percent-effective-plans-to-submit-to-fda-in-days.html.

“Pfizer and BioNTech Conclude Phase 3 Study of COVID-19 Vaccine Candidate, Meeting All Primary Efficacy Endpoints | Pfizer.” Accessed December 17, 2021. https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-conclude-phase-3-study-covid-19-vaccine.

Peru, B, “K-Means vs Vader Vs TextBlob Sentiment Analysis”, Accessed Dec 17, 2021 https://www.kaggle.com/accountstatus/k-means-vs-vader-vs-textblob-sentiment-analysis

Shah, R, “How to Build Word Cloud in Python?”, Analytics Vidhya, May 20, 2021, Accessed Dec 17, 2020 https://www.analyticsvidhya.com/blog/2021/05/how-to-build-word-cloud-in-python/

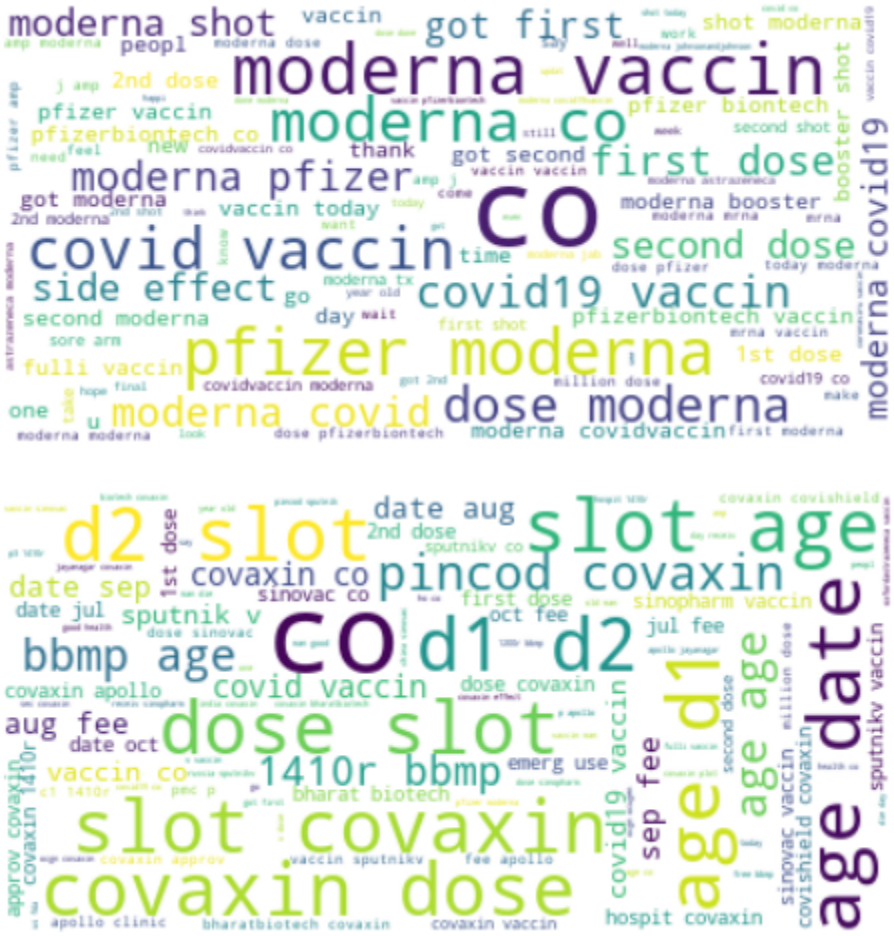
“What Are MRNA Vaccines and How Do They Work?: MedlinePlus Genetics.” Accessed December 17, 2021. https://medlineplus.gov/genetics/understanding/therapy/mrnavaccines/.

Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A., and Sharif, S. (2021). “An analysis of COVID-19 vaccine sentiments and opinions on Twitter.” *Int. J. Infect. Dis*. 108, 256–262. doi: 10.1016/j.ijid.2021.05.059 https://reader.elsevier.com/reader/sd/pii/S1201971221004628?token=6D8B25723AB60D05B994D18FC705630AB6720731EFE33ADAB5F53210DB5EB877BB301A2B38CB3B018A9357CFBDF1BBE0&originRegion=us-east-1&originCreation=20211217200239

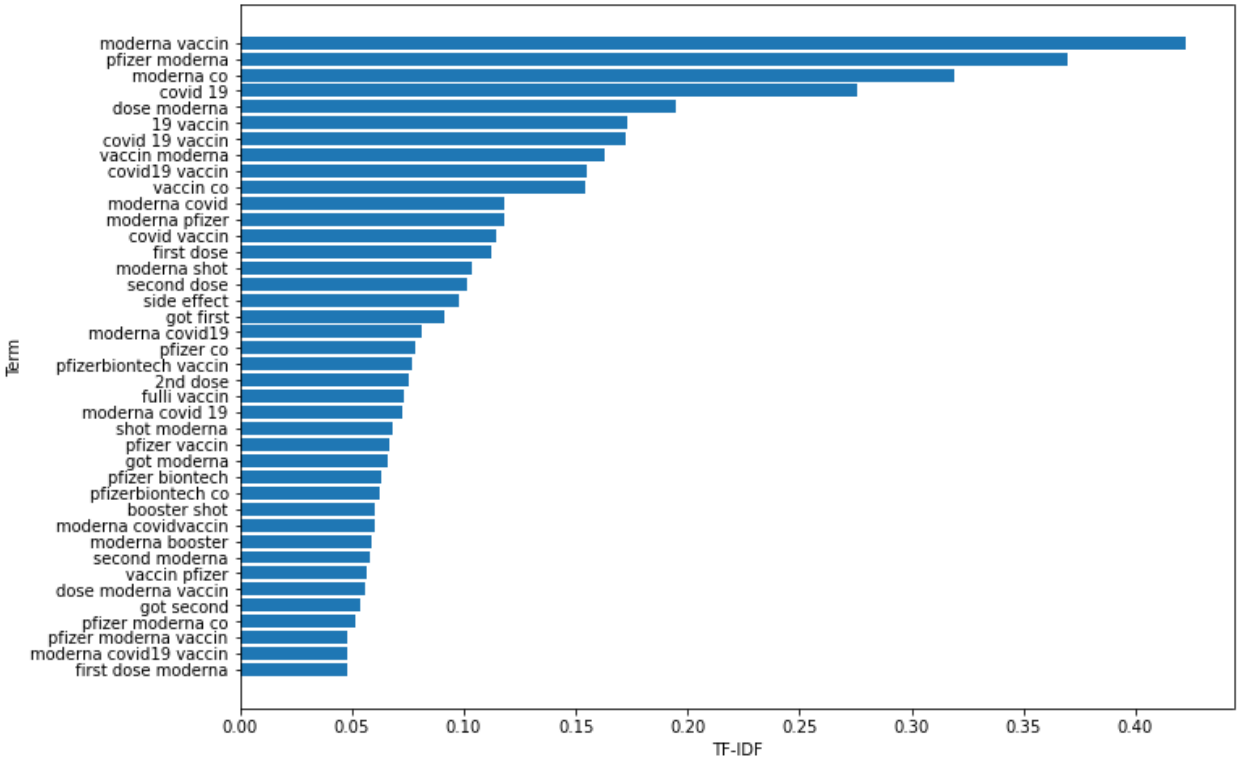
Appendix I: Reference keywords used to filter types of tweets

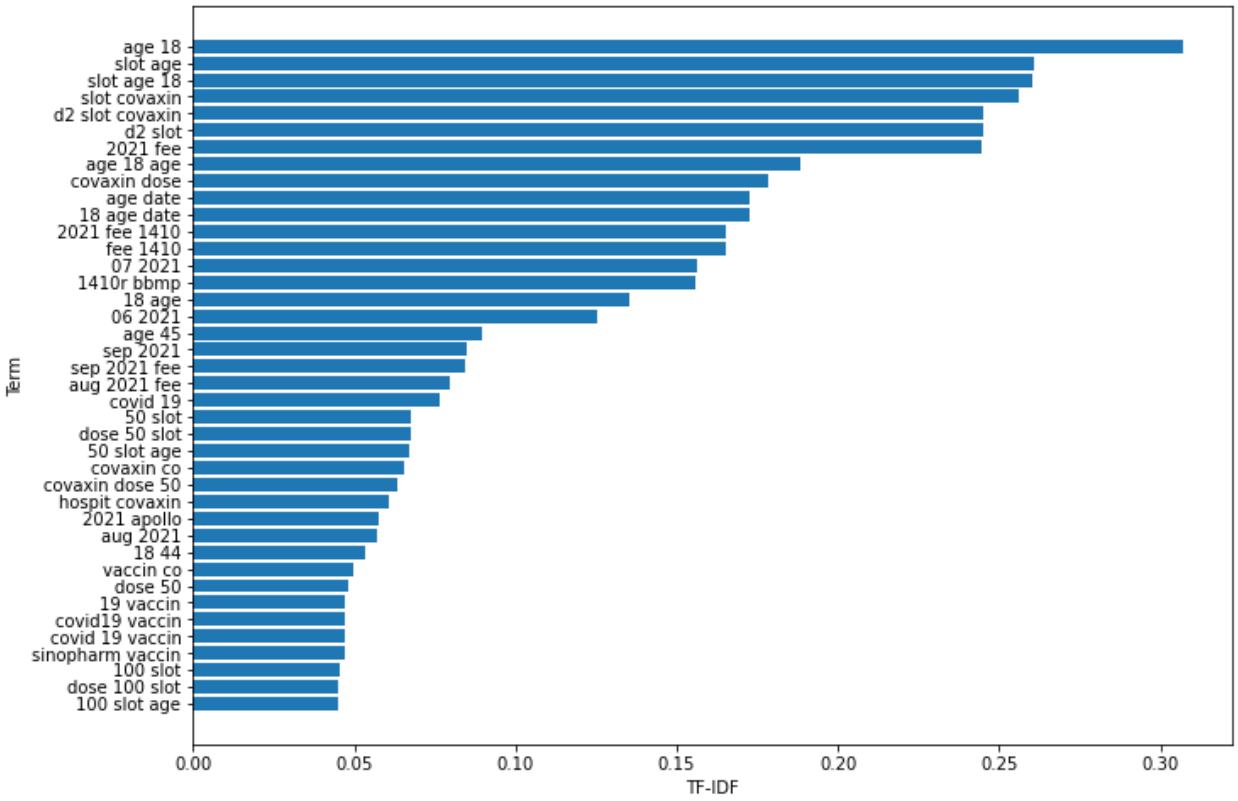
```
pfizer_refs = ["pfizer","pfizer-bioNtech","biontech"]
bbiotech_refs = ["covax","covaxin","bharat biotech","bharatbiotech"]
sputnik_refs = ["russia","sputnik","v"]
astra_refs = ['sii','adar poonawalla','covishield','astra','zenca','astrazenca','oxford-astrazenca','serum institiute']
moderna_refs = ['moderna','moderna','mrna-1273','spikevax']
sinopharm_refs = ['sinovac','sinopharm', 'chinese', 'china', 'sino']
```

Appendix II: Word Clouds for mRNA and non-mRNA tweets



Appendix III: Top words by TF-IDF among mRNA and non-mRNA tweets





Appendix IV: Topic models for mRNA and non-mRNA tweets