Sentiment Analysis and Topic Modeling of Covid Vaccines in USA (1714 WORDS)

Ajinkya Pawale
Spring 2022
SOCIAL MEDIA MINING ILS Z-639

1. Introduction

To put a stop to the SARS-CoV-2 epidemic, a worldwide vaccination drive was underway; however, its success relied on people's desire to become vaccinated. Vaccine hesitancy was identified one of the top ten global health hazards by the World Health Organization (WHO) in 2019, citing complacency, inconvenient access, and a lack of confidence as the main factors. Shortly after the COVID-19 pandemic ravaged the globe, vaccine hesitancy garnered worldwide prominence, with many residents in countries like the United States refusing to take approved COVID-19 immunizations owing to concerns about safety, adverse effects, or an innate distrust of the government. [1] Uses NLP to preprocess the raw tweets and KNN Classification Algorithm to classify the processed data. The study suggests that Pfizer and Moderna vaccines have more positive sentiment among the general public, with ratings of 47.29 and 46.16, respectively, compared to AstraZeneca vaccines, which have a rating of 40.08. In [2] throughout the four months (December 2020 - March 2021), public opinion on Pfizer and Moderna vaccines appeared to be positive and consistent, with no major shifts between months. In contrast, public opinion toward the AstraZeneca/Oxford vaccination appears to be deteriorating over time, with a considerable drop between December and March. [3] performed topic modelling along with sentiment analysis. It found that topics revolving around positive tweets were about vaccination planning, getting vaccination and vaccination information. The topics revolving around negative tweet were about vaccine hesitancy, extreme side effects of the vaccines and vaccine supply/rollout.

2. Research Question

What was the attitude and sentiment towards SARS-CoV-2 vaccination in the USA? Which topics were prominent in these discussion threads?

3. Method

a) Data:

To address the above-mentioned research questions data was collected from Twitter. It was gathered using the snscrape package available in python. The search keywords used were Pfizer, Moderna, Johnson & Johnson and Vaccine.

Tweets were collected between the period from January 2021 to June 2021 since it was the time when vaccines were widely available to the people. Tweets were collected separately for the keywords Pfizer, Vaccine followed by Moderna, Vaccine and finally Johnson & Johnson, Vaccine. To make sure that tweets are extracted from each month, instead of giving an entire range from January to June 1000 tweets are extracted for each month.

Sentiment Analysis and Topic Modeling of Covid Vaccines in USA

For the keywords Pfizer, Vaccine a total of 4234 tweets were collected. For the keywords Moderna, Vaccine a total of 5322 tweets were collected. For the keywords Johnson & Johnson, Vaccine a total of 4432 tweets were collected. The reason why 6000 tweets were not extracted for each keyword pair is because snscrape did not find those many tweets for the specified time range. Tweets with language 'en' i.e., English are extracted. Location was set to near United States.

To perform seamless analysis, tweets are cleaned by converting to lower case, removing stop words, usernames, punctuations, URLs, and numbers. Since the data will be passed to a Machine Learning model, it is tokenized so what we get it in a numeric form. Finally, the tweets were lemmatized to get its root form.

b) Analysis:

First step involved in analysis is the sentiment analysis of the cleaned tweets. It is important to perform sentiment analysis on the cleaned data since unnecessary noise will give a wrong indication of the sentiment score. To avoid that, tweets are cleaned by removing username, hashtags, links, etc. as discussed above in the data section. The approach in this paper will calculate the sentiments of the tweets using a Machine Learning model. A prelabeled dataset named Sentiment140 is used which consists of 1.6 million tweets that are extracted using the twitter api. The tweets in this dataset are annotated as negative or positive. Thus, the Machine Learning model will be trained on the Senitment140 dataset and tested on the tweets collected for covid vaccines as mentioned above. Out of the 1.6 million tweets available in Sentiment140 100k tweets (50K each for positive and negative) are used for the training phase. The same data cleaning steps are applied to the training tweets as mentioned above for the extracted tweets. The training phase is further split into train and validation, the split is 80-20 i.e., 80% training and 20% validation. 5 Machine Learning algorithms are implemented on the 80% training data and are tested on the 20% validation data. The evaluation of the 5 models on the validation data is shown below in the table 1. From the table 1 it is clear that SVM is performing better as compared to the other models. Thus, SVM is used to predict the sentiments on the covid vaccine tweets.

Models	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.75	0.74	0.75	0.75
SVM	0.77	0.77	0.76	0.77
KNN	0.55	0.55	0.53	0.54
Ada Boost Classifier	0.69	0.69	0.68	0.69
Bagging Classifier	0.73	0.73	0.72	0.73

Table 1. Model Evaluation Metrics on the validation set

Second step involves topic modelling using BERT. In this work, embeddings are obtained using the sentence-transformers python package. Distilbert is used as the sentence transformer because it gives a good balance of speed and performance. Next before moving on to clustering similar topics, we first need to reduce the dimensionality of the embeddings as majority of the cluster algorithms don't handle high dimensions properly. Thus, UMAP package is used to perform dimension reduction with the final dimension size as 5 and the size of the local neighborhood (n-neighbors) is set as 15. Further, HDBSAN is used to cluster the documents as it works hand in hand with UMAP. In HDBSAN the cluster size is set to 15 and distance metric is selected as Euclidean. Figure 1 shows the topics visualized by reducing the sentence embeddings to 2-dimensional space. Now, once the clusters are formed, we need to figure out what makes a particular cluster different from another. To solve this class-based variant of TF-IDF is used. All the documents are treated as a single category and then TF-IDF is applied. As a result, each category would have a very long document, and the TF-IDF score would show the most essential terms in the topic. We take the top 20 words per topic based on their c-TF-IDF scores to construct a topic representation. Because the score is a proxy for information density, the higher the score, the more representative it should be of its topic.

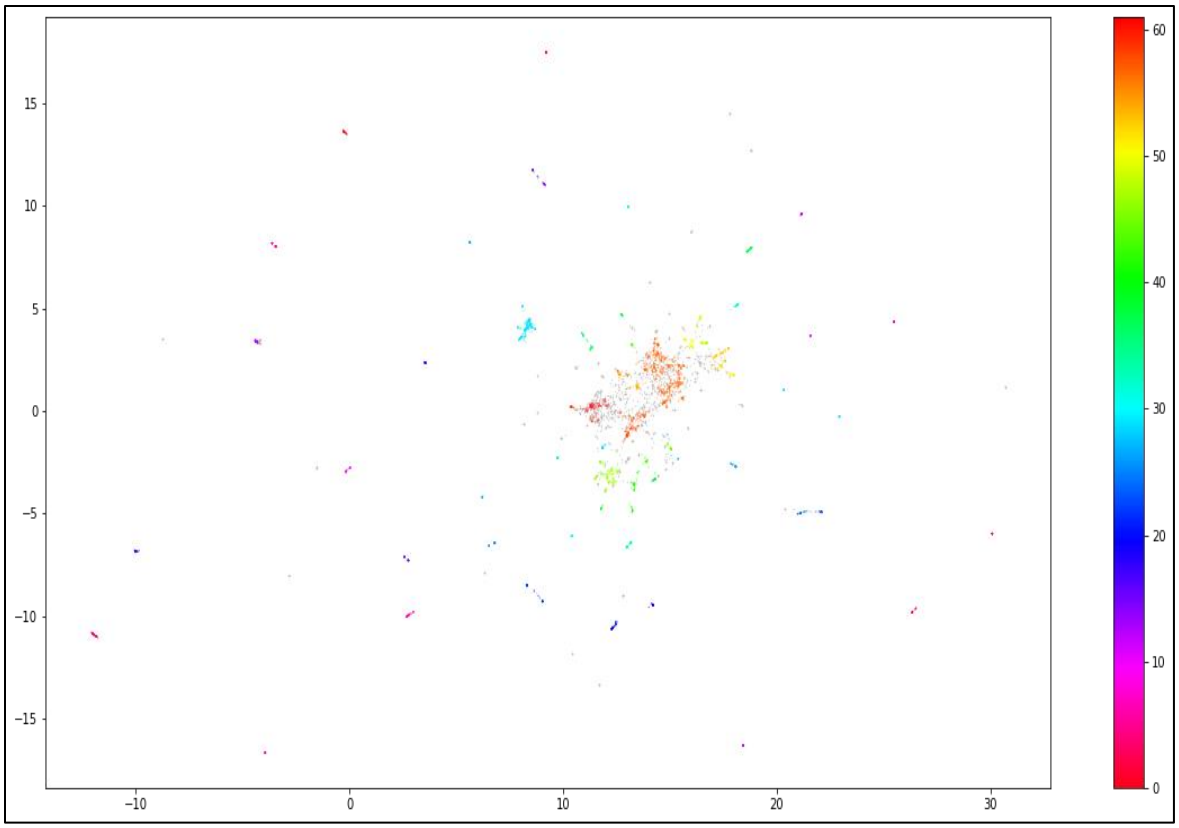

Fig 1. Topics visualized by reducing sentenced embeddings to 2-dimension space

4. Result

Coming to the results, let's focus first on the sentiment analysis scores. Out of the total 13988 tweets collected for Pfizer, Moderna and Johnson & Johnson 9410 tweets were positive, and 4578 tweets were labelled as negative. Overall, we can say that there was a positive attitude towards the vaccines. Table 2 gives the distribution of sentiments for individual vaccines. If we calculate the ratio of positive to negative sentiments for each vaccine in table 2, we find out that the ratio is maximum for Pfizer, followed by Moderna and Johnson & Johnson. Thus, we can say that Pfizer was involved in more positive discussions followed by Moderna and then Johnson & Johnson.

Vaccine	**Positive Sentiments**	**Negative Sentiments**	**Positive/Negative Sentiments Ratio**
Pfizer	3003	1231	2.43
Moderna	3647	1675	2.17
Johnson & Johnson	2760	1672	1.65

Table 2. Distribution of Sentiments across each vaccine

Analyzing the frequently occurring words in the tweets can give a good idea about what terms are talked about often as compared to others. In table 3 we can see the top 5 common words occurring across the vaccine categories except for the words that are by default common like the vaccine name, covid and coronavirus. From the table we can say that the tweets around Pfizer concentrate around their collaboration with the BioNTech company. Tweets from Moderna are more revolving around the testing and trial of the vaccine and the work around the mRNA technology. Similarly, tweets from Johnson & Johnson involve discussions around the human trial and the start of the vaccine administration.

Vaccine	**Top 5 Common Words in Tweets**
Pfizer	could, ready, biontech, fall, company
Moderna	mrna, trial, company, phase, testing
Johnson & Johnson	human, trial, september, start, begin

Table 3. Top 5 common words in tweets across each vaccine

Looking at the sentiments at a more granular level can give us more insights. Figure 2, 3 and 4 gives us insight into the positive to negative tweet ratio (p/n ratio) for each vaccine over the period of January-June 2021. For Pfizer the maximum p/n ratio was during May followed by June, it might be probably due to successful administration of both the doses of vaccine. Similarly, for Moderna the maximum p/n ratio was during February followed by June. The explanation for June might be similar to that of Pfizer during June and probably during February since the rollout was started people were more optimistic about the vaccine. Lastly, for J&J maximum p/n ratio was during June followed by March, the explanations might be similar to that of Moderna.

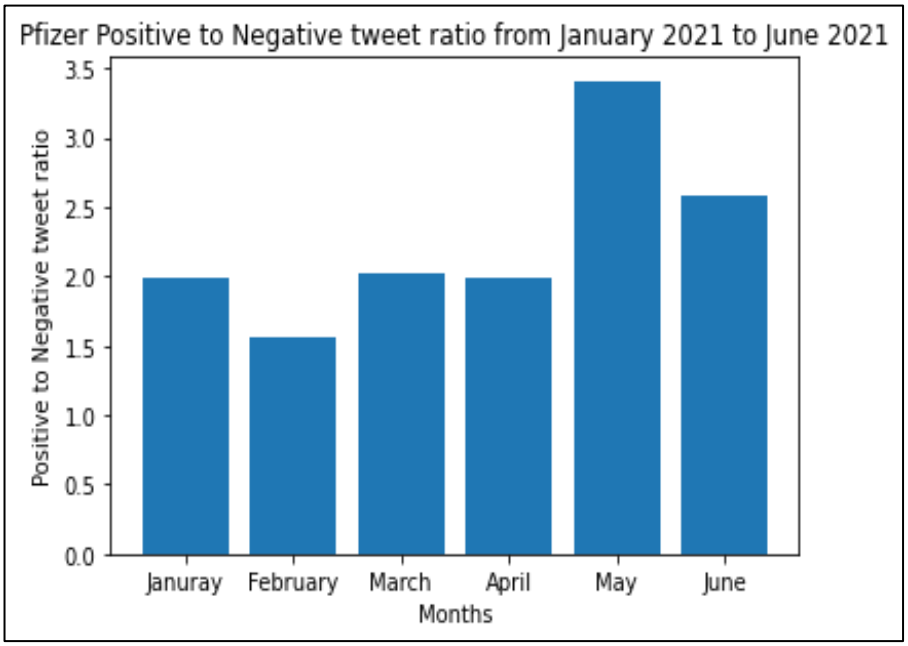

Fig 2. Pfizer Positive to Negative tweet ratio from January 2021 to June 2021

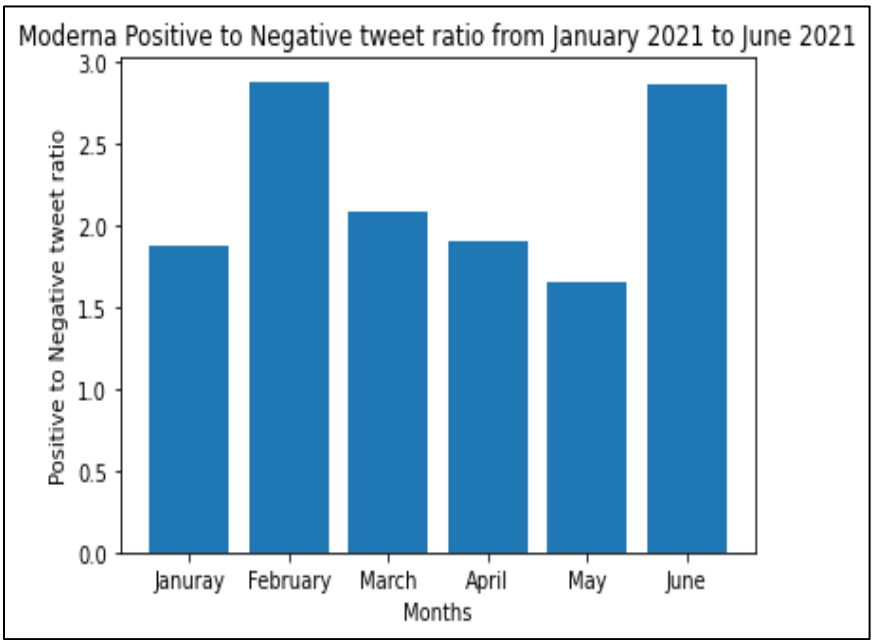

Fig 3. Moderna Positive to Negative tweet ratio from January 2021 to June 2021

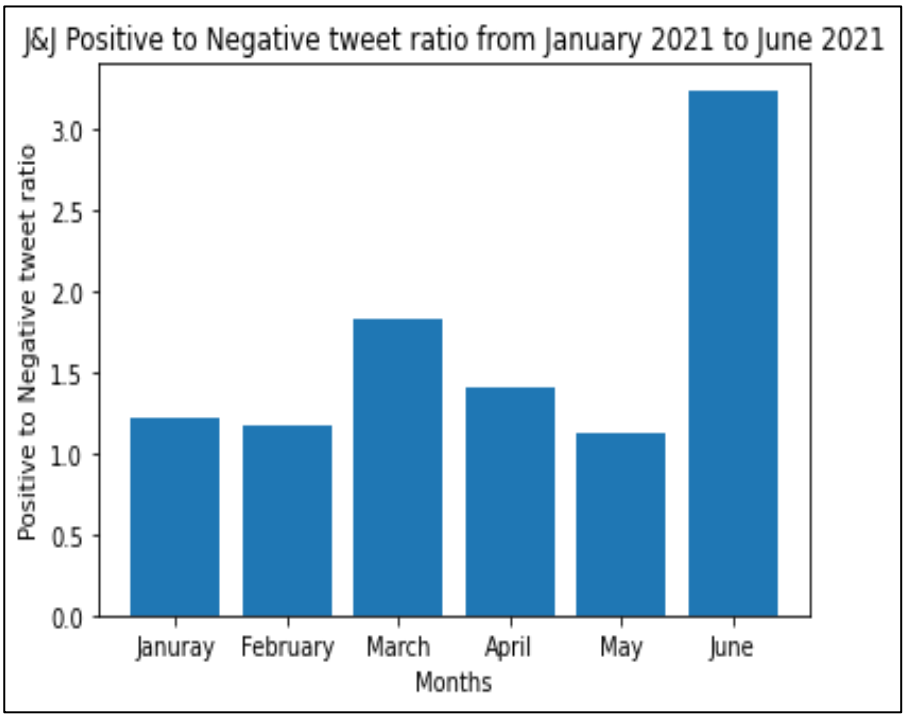

Fig 4. Johnson & Johnson Positive to Negative tweet ratio

Topic modeling gave some interesting results for all the three vaccines. For each vaccine, top 3 topics were selected with most occurrences in tweets. For those 3 topics, top 10 keywords within them were selected to get an idea of the topic being discussed. Table 4, 5, 6 shows the results of topic modeling for Pfizer, Moderna and Johnson and Johnson respectively. Topics in Pfizer revolved around the comparison of Pfizer with Moderna, merger with the German company BioNTech for worldwide supply, vaccine testing and trials. Topics in Moderna revolved around stock and availability of the vaccine, merger with the Trump administration, successful trials and availability to public, comparisons with Pfizer. Topics in Johnson & Johnson revolved around the second phase of trial, benefits vs risks of the J&J vaccine, positive response towards the vaccine.

Topic no.	Top 10 Keywords	Repeated in no. of tweets	Inferring the topic from keywords
1	vaccine, pfizer, covid, coronavirus, biontech, company, moderna, trial, development, mrna	407	Topic 1 is related to the comparison of pfizer and moderna. It also talks a little about the merger with BioNTech.
2	germany, vaccine, biontech, Pfizer, german, trial, covid, coronavirus, company, potential	213	Topic 2 tries to address the partnership of Pfizer and the German company BioNTech which helped to create a lot of vaccine supplies across the globe.
3	week, vaccine, testing, pfizer, coronavirus, begin, say, fall, trial, early	125	Topic 3 hints the testing of the vaccine and it's trials. It also suggests the start of the vaccine administration after successful trials.

Table 4. Top 3 topics for Pfizer according to the BERT algorithm

Topic no.	Top 10 Keywords	Repeated in no. of tweets	Inferring the topic from keywords
1	vaccine, moderna, covid, million, stock, trial, mrna, coronavirus, company, trump	690	Topic 1 talks about the stock availability and trials of the moderna vaccine. It also talks about the merger of Trump administration with Moderna to create more doses of the vaccine.
2	vaccine, moderna, covid, mrna, work, trial, work, good, development, phase, thank	275	Topic 2 tries to talk about the successful trials and the administration of the vaccines to the public.
3	vaccine, moderna, trial, month, week, phase, covid, start, bitontech, testing	255	Topic 3 hints about comparison of Moderna and Pfizer-BioNTech vaccine.

Table 5. Top 3 topics for Moderna according to the BERT algorithm

Topic no.	Top 10 Keywords	Repeated in no. of tweets	Inferring the topic from keywords
1	johnson, vaccine, trial, human, covid, september, coronavirus, begin, start, july	361	Topic 1 talks about J&J's initiation of the Phase 1/2a first-in-human clinical trial of its investigational covid vaccine. Initially scheduled to begin in September, the trial was commenced in the second half of July.
2	vaccine, don, johnson, want, work, going, need, know, people, make	205	Topic 2 tries to talk about the benefits of the vaccine versus its risk. People were skeptical as it was just a one-shot vaccine.
3	expected, september, johnson, start, trial, human, coronavirus, vaccine, what, yay	124	Topic 3 tries to convey the start of the Phase 1/2a human trial as mentioned in Topic 1. It also shows positive response around the vaccine.

Table 6. Top 3 topics for Johnson & Johnson according to the BERT algorithm

5. Conclusion and Limitation

From the above analysis and results there are 3 takeaways for the given experimental dataset. The positive to negative tweet ratio was the best for Pfizer, which suggests that people widely accepted it. The most widely discussed topic for Pfizer and Moderna is their merger with the firms mentioned above. Topics surrounding Johnson and Johnson are more regarding the trials of the vaccine.

Although this research gives us insight into the kind of characteristics that may give user feelings on social media for covid vaccines, there are a few limitations that suggest areas for further research. One of the most significant of these limits is the scale. This research examines about 14000 tweet which is just a sample. As a result, more study is needed in which the scope is broadened, more samples are tested and possibilities to study more tweets exists. The future work could extend on analyzing more tweets from different geo locations/countries and analyze the behavior in those regions. Apart from that a more wider time frame can be considered and compared against each other.

This research comes up with a customized scheme for determining the topics across tweets. Hyperparameters in the BERT topic model can be changed to obtain better results. More work can be done to understand the various tags/topic across the tweets and derive meaningful insights.

6. References

[1] Shamrat, F. M. J. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, O. M. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. Indones. J. Electr. Eng. Comput. Sci, 23(1).

[2] Marcec, R., & Likic, R. (2021). Using twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgraduate Medical Journal.

[3] Huangfu, L., Mo, Y., Zhang, P., Zeng, D. D., & He, S. (2022). COVID-19 Vaccine Tweets After Vaccine Rollout: Sentiment–Based Topic Modeling. Journal of medical Internet research, 24(2), e31726.