

Text Data Analysis on COVID-19 Vaccination Tweets

All code related to this project is hosted on this GitHub repository: <https://github.com/jameslee98331/mls-uda-final-project>

Contents

1	Introduction	1
2	Dataset Selection	2
2.1	Data Source	2
2.2	Choice of Dataset	2
3	Problem Statement	2
4	Exploratory Data Analysis	2
4.1	Basic Description of Dataset	2
4.2	Data Quality Issues	2
4.2.1	Missing Data	2
4.2.2	Duplicate Data	3
4.3	Tweets from Bot Accounts	4
4.4	Lack of Labels	4
5	Data Pre-processing	4
5.1	Data Cleaning	4
5.2	Labelling	4
5.3	Resampling	5
5.4	Tweet Pre-processing	5
5.4.1	Feature Selection	5
5.4.2	Tokenization and Lemmatization	5
6	Tweet Sentiment Analysis and Modelling	6
6.1	Tweet Sentiment by Vaccine Manufacturers	6
6.2	Tweet Sentiment over Time	7
6.3	Logistic Regression Models	7
6.3.1	Word Embedding	7
6.3.2	Model Training	8
6.3.3	Sentiment Predictions vs. Vaccine Manufacturers and Time	9
6.3.4	Comparison with Others' Works	9
7	Summary	10
7.1	Findings	10
7.2	Limitations and Further Work	10

1 Introduction

As we, hopefully, approach the end of the COVID-19 pandemic and begin to look back at how this global event has shaped our world, we cannot overlook the influential role social media has played in shaping public opinions, particularly around vaccinations [2]. This project aims to dissect into the public sentiment towards COVID-19 vaccination through analysing tweets.

2 Dataset Selection

2.1 Data Source

The COVID-19 All Vaccines Tweets dataset published by Gabriel Preda on Kaggle [7] is chosen for this project. Tweets, in English, related to COVID-19 vaccines (Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin and Sputnik V) within this dataset are collected through searching for relevant terms via Twitter's API, using code that can be found here: <https://github.com/gabrielpreda/covid-19-tweets>. The dataset is expected to be updated daily (last updated 2021-11-23). This dataset is released under the "CC0: Public Domain" License. The dataset is downloaded for this project at 2022-12-15 23:13.

2.2 Choice of Dataset

This dataset is chosen as it presents an interesting challenge for text data analysis. Tweets require significant attention as they contain more complex components on top of words, such as @user tags, hashtags, emojis, hyperlinks... etc. Tweets also tend to be in more irregular "internet-style" grammatical structures; it would be interesting to see how the methods discussed in the Unstructured Data Analysis module work on them.

3 Problem Statement

The aim of this project is to

1. explore methods of cleaning and pre-processing tweet text data,
2. compare Count vectorization against TF-IDF vectorization,
3. build and analyse sentiment classification models, and
4. analyse the predicted sentiments against time and vaccine manufacturers.

4 Exploratory Data Analysis

Before analysing the textual components, it is important to first understand some basic properties of the dataset.

4.1 Basic Description of Dataset

This dataset is presented in CSV format and contains 228,207 rows and 16 columns. Each row represents a tweet, only tweets in English were extracted. The columns can be broadly split into 2 categories, 8 columns each:

1. Information about the user who posted the tweet, e.g. `user_name`, `user_location`, `user_description`, `user_created`, `user_followers`, `user_friends`, `user_favourites`, and `user_verified`, and
2. Information about the tweet, e.g. `id`, `date`, `text`, `hashtags`, `source`, `retweets`, `favourites`, and `is_retweet`.

The earliest tweet recorded was from 2020-12-12 11:55:28 and the latest from 2021-11-23 20:58:08.

4.2 Data Quality Issues

4.2.1 Missing Data

Some of the columns contain missing (NaN-like) values. 5 out of 16 columns contain missing data: `source`, `hashtags`, `user_description`, `user_location`, and `user_name`. Since the analysis will mainly focus on the text column, the large numbers of missing `user_description` and `user_location` can be accepted. The `hashtags` column also contain a large number of missing data due to the fact that not all tweets contain hashtags, so this missingness is also expected. Very few number of tweets have missing `source` and `user_name`. Figure 1 shows a visualisation of the counts of missing data in the 5 columns.

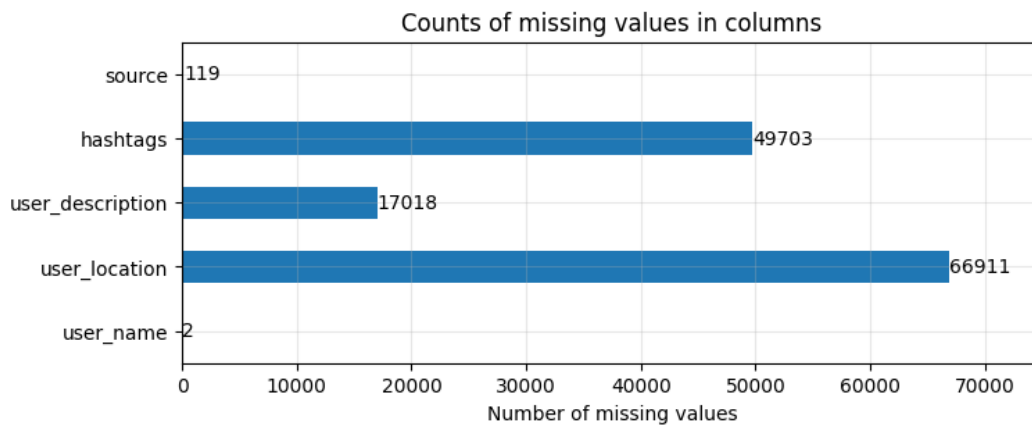


Figure 1: Count of missing data in columns

There are also no tweets recorded for some of the dates, e.g. around May and July 2021 as seen from Figure 2. It is not certain whether this missing data is caused by the scraping code or that there were genuinely no tweets satisfying the requirements around those dates. While this is not necessarily a problem for analysing the existing text, it must be dealt with carefully when trying to analyse properties of the tweets, such as sentiment over time.

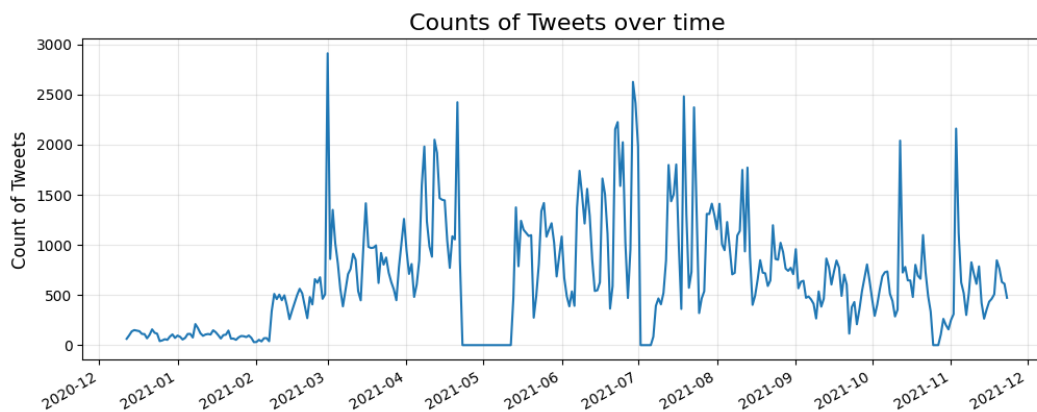


Figure 2: Count of Tweets over time from 2020-12-12 to 2021-11-23

Aside from the days with missing tweets, some clear trends can also be observed from the count of tweets plot. There is a clear peak around early March 2021, with increasing volumes until around July/August 2021. There are also 2 other clear peaks around October/November 2021.

Some tweets are also truncated. Reading the discussion around this dataset on <https://www.kaggle.com/datasets/gpreda/all-covid19-vaccines-tweets/discussion>, we can see that this is a problem with how the author has collected the data. While this reduces the usefulness of the data, the sentiment of the available part of the tweet can still be analysed.

4.2.2 Duplicate Data

	user_name	user_location	date	text
218116	Grace Schmitz	Calumet, MI	2021-11-05 14:33:14	Boosted! #moderna
225230	Natasja S	United Kingdom	2021-11-15 17:41:46	Boosted! #moderna
224386	Gary Wien	New Jersey, USA	2021-11-19 17:40:11	Boosted! #moderna

Figure 3: Identical tweet by different users at different times

There are no duplicate tweet ids but it was found that there are identical tweet texts. Below in Figure 3 shows an example of an identical tweet “Boosted! #moderna” by three different users at different times. In this case, these tweets should be treated as distinct. On the other hand, there are also rows of identical tweets from the same user generated within a very short period of time, e.g. in Figure 4, these should be treated as duplicates and only one should be kept. Section 5.1 will revisit the handling of these duplicate rows in the formation of the processed text dataset.

	user_name	date	text
227202	Craig Gordon	2021-11-18 23:23:50	@RepThomasMassie The vaccine you are referring to is #Covaxin #Ocugen #BharatBiotech
227203	Craig Gordon	2021-11-18 23:23:28	@RepThomasMassie The vaccine you are referring to is #Covaxin #Ocugen #BharatBiotech
227204	Craig Gordon	2021-11-18 23:23:15	@RepThomasMassie The vaccine you are referring to is #Covaxin #Ocugen #BharatBiotech

Figure 4: Identical tweet by the same user within a few seconds

4.3 Tweets from Bot Accounts

By checking the user names with the most tweets, it is also discovered that this dataset contains a few bot accounts for vaccination slow alerts. For example, accounts “CowinBangalore” and “CoWIN Blore 18-44” alone contributed over 22,000 tweets, which is almost 10% of tweets in this dataset. Upon further investigation, these accounts appear to be part of the Indian government’s vaccination effort [6]. The tweets by these accounts are all related to vaccination slots with times and locations, rather than expressing sentiments on vaccination. Hence, these accounts will be disregarded. Section 5.1 will revisit the removal of these bot accounts from analysis.

4.4 Lack of Labels

It is important to note that this dataset does not contain any labels on the tweets. This will be discussed in more detail in section 5.2.

5 Data Pre-processing

5.1 Data Cleaning

As discussed in sections 4.2.2 and 4.3, there are duplicate rows and tweets from bot accounts that should be removed from the dataset. Duplicate rows from the same user with the same text posted on the same date are removed (the first one is kept). Tweets from bot accounts are also removed by filtering on the **source** column, as they provide a large quantity of factual information about vaccine slots, which adds a lot of noise and little value to text sentiment analysis. Only columns **user_name**, **user_location**, **fulldate** (removed the time component from original **date** column), **text**, and **source** are kept from the original dataset as they’re more likely to be relevant to this study. After down-selecting, the reduced dataset has 189,147 rows and 5 columns.

5.2 Labelling

The reduced dataset does not contain any sentiment labels, or any other data that could be interpreted as a sentiment label. So, to allow for model building in subsequent sections, the sentiments of the original texts are classified using the **cardiffnlp/twitter-roberta-base-sentiment-latest** [5] model through the “🤗 Transformers” library.

This model is based on **roBERTa** [4] and is trained on around 124 million tweets, which is then fine-tuned for sentiment analysis against **TweetEval** [1], which is a “unified benchmark for tweet classification”. Fine-tuning is the process of taking a pre-trained large language model (**RoBERTa**) and updating the weights with additional training data, for a specific task; sentiment classification in this case. This model is chosen as it was fine-tuned with tweets from early 2018 to late 2021 for sentiment analysis. This covers tweets relevant to the COVID-19 pandemic, which aligns with the objective of this project. As [5] suggests, the Twitter corpus used to build the **twitter-roberta-base** model is limited to English, which is also suitable for our dataset.

The `twitter-roberta-base-sentiment-latest` model is used to generate scores for all three sentiments (negative, neutral and positive) for each tweet. The sentiment with the highest score is extracted as the sentiment label. These sentiment classifications are used as proxies to the ground truth labels. Any subsequent model building are evaluated against these labels.

5.3 Resampling

The generated labels are not balanced: there are many more neutral tweets than negative and positive tweets. For fairer modelling in later sections, the reduced dataset is balanced by resampling with replacement to 35,000 observations per sentiment label. Figure 5 shows the counts of the labels before and after resampling. After resampling, the reduced dataset has 105,000 rows and 5 columns.

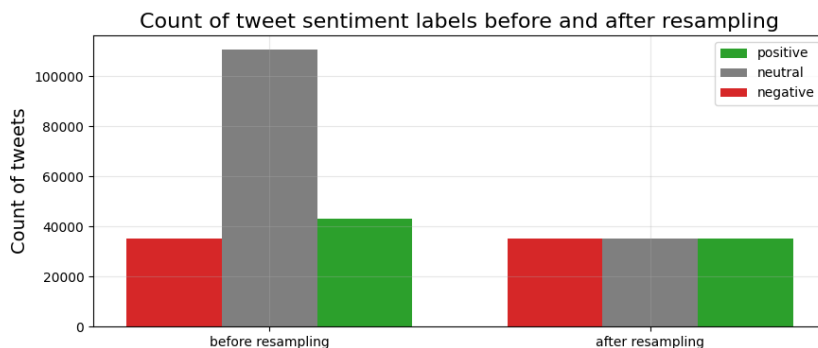


Figure 5: Count of Tweet sentiment labels before and after resampling

5.4 Tweet Pre-processing

The pre-processing step removes features that are unimportant for analysis, to reduce the complexity of the text before modelling in section 6.

5.4.1 Feature Selection

For a basic level of text data cleaning, all words are converted to lower case, extra white-spaces, tabs, and newline characters are stripped out. From section 4, we see that some tweets contain hyperlinks to the original tweet, these are also removed using regular expressions matching. We also see that the tweets contain more complex elements than just words, such as emojis, hash tags (#) and user tags (@). Hash symbols (#) are removed from hash tags and the word is retained. For example, “#covid” is reduced to “covid”. @user tags are kept as it is expected that many tweets may include mentions of national leaders or organisations. It would be interesting to see if these user tags have any bearing on sentiment.

Emojis are also kept as they can easily change the sentiment label. Sentences with identical words but different emojis can express very different sentiments. Taking the sentences in Figure 6 as an example: both lines were supplied to the `twitter-roberta-base-sentiment-latest` model [5], adding the grinning face emoji 😊 switches the sentiment from neutral to positive, while adding the worried face emoji 😟 gives a negative sentiment. Each emoji is represented as a unicode character, for example the grinning face emoji 😊 is represented as “\U0001F600”, so they can be treated as a single token in the next section.

5.4.2 Tokenization and Lemmatization

The column of pre-processed text is the corpus, and each pre-processed tweet is a document. Each documents is split into tokens, each token represents a word, a symbol, an emoji, or a user tag. Tokens that are classified by the SpaCy `en_core_web_sm` pipeline as punctuation or stop words (e.g. “a”, “the”, ...etc.) are removed. The SpaCy pipeline is

text	text_sentiment_scores	max_score_label
I received a dose of #pfizerbiontech covid-19 vaccine	{'negative': 0.04958874, 'neutral': 0.83540446, 'positive': 0.11500678}	neutral
I received a dose of #pfizerbiontech covid-19 vaccine 😊	{'negative': 0.0078060045, 'neutral': 0.26736185, 'positive': 0.7248322}	positive
I received a dose of #pfizerbiontech covid-19 vaccine 😞	{'negative': 0.81133777, 'neutral': 0.17565854, 'positive': 0.013003681}	negative

Figure 6: Emojis significantly changing the sentiment of a sentence with identical words

modified to keep negation words, such as “no”, “not” and “never”, as tweets with negative sentiments are expected to include them.

Each token is then lemmatized, which is a process of grouping words that came from the same root together. For example, “got”, “gets”, and “getting” would all be reduced to “get”. The pre-processed texts are then saved as **preprocessed_text** in the reduced and balanced dataset. Another column named **preprocessed_text_no_noun** is also added to the data, as set of pre-processed text without nouns and proper nouns like “covid” and “vaccines”. Figure 7 provides a visualisation of the most common words in the tweets after pre-processing, with nouns and proper nouns removed.

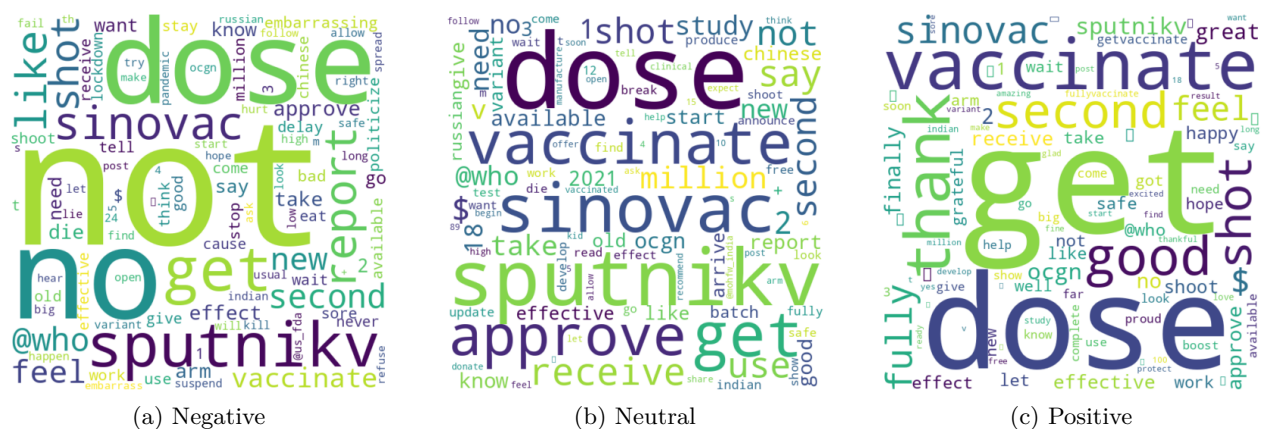



Figure 7: Word Clouds extracted from pre-processed text, split by sentiment label

From the negative word cloud, the negation words “no” and “not” are very prominent as expected. We can also see words that are potentially related to vaccine side effects, such as “effect”, “arm”, “hurt”, and “die”. There is also an interesting mix of words surrounding the theme of COVID-19 policies, such as “politicize”, “lockdown”, and “embarrassing”.

The neutral word cloud shows words that we may expect more from news reports, such as “study”, “approve”, “million” and “receive”. The positive cloud appears to show how people “feel” after vaccination, with words like “good”, “thank” and “effective” being the most common.

It is interesting to observe that the mention of World Health Organization’s Twitter account “@who” is more frequent among negative and neutral tweets than in positive tweets. It is also interesting to see that the syringe emoji  appeared as one of the top 10 positive tokens (before filtering out proper nouns and nouns), with 3,850 occurrences. It is surprising to see “sputnikv” and “sinovac” on the word clouds despite the filtering with SpaCy.

6 Tweet Sentiment Analysis and Modelling

6.1 Tweet Sentiment by Vaccine Manufacturers

The pre-processed tweets are grouped into the 3 sentiment labels generated in 5.2, and then by vaccine manufacturers. Figure 8 shows both the counts and the percentages of sentiment labels for each vaccine. It can be observed from Figure 8(a) that Covaxin and Moderna had the most Tweets. While Moderna and Covaxin has the highest proportions

of positive tweets, Sputnik and Sinopharm appears to have the lowest percentage of positive tweets. The Oxford-AstraZeneca vaccine has the highest proportion of negative tweets, although a much smaller number of tweets relative to Moderna and Covaxin.

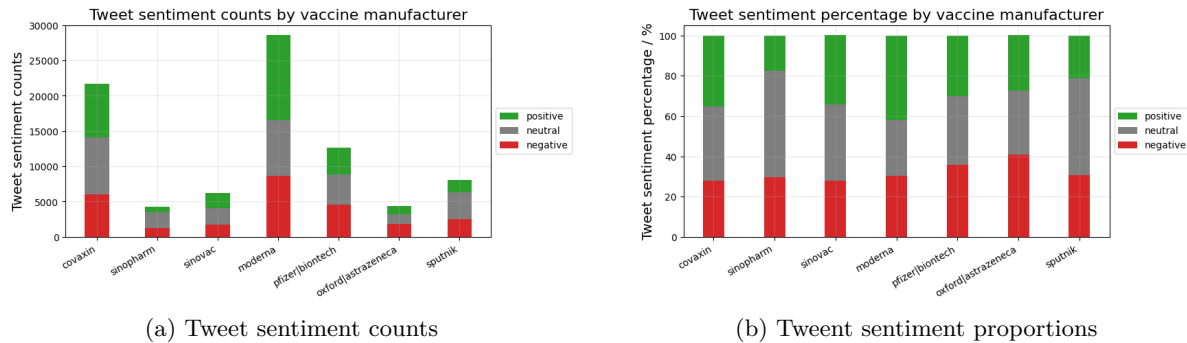


Figure 8: Counts and percentages of tweet sentiment labels grouped by vaccine manufacturers

6.2 Tweet Sentiment over Time

The proportions of tweet sentiment is also measured over time. The gap in the line graphs show the dates with no tweets. Although the proportions between negative, neutral and positive is not very distinct, we can clearly see points where negative and positive sentiments switch around. For example, from Figure 9, there is a strong peak of negative sentiments in early July 2021, the negative sentiment is significantly higher. This is around the same time as the FDA releasing results on 12th July 2021 on the potential side effects of the vaccines [8]. Inspecting the data for this date, tweets like “This is only the beginning of consequences of these vaccines. No one knows what the long term effects are. #JohnsonAndJohnson #Pfizer #moderna” can be found. From a broader view, we can also see that the proportion of negative tweets is trending up while the proportion of positive tweet is trending down, albeit very slowly.

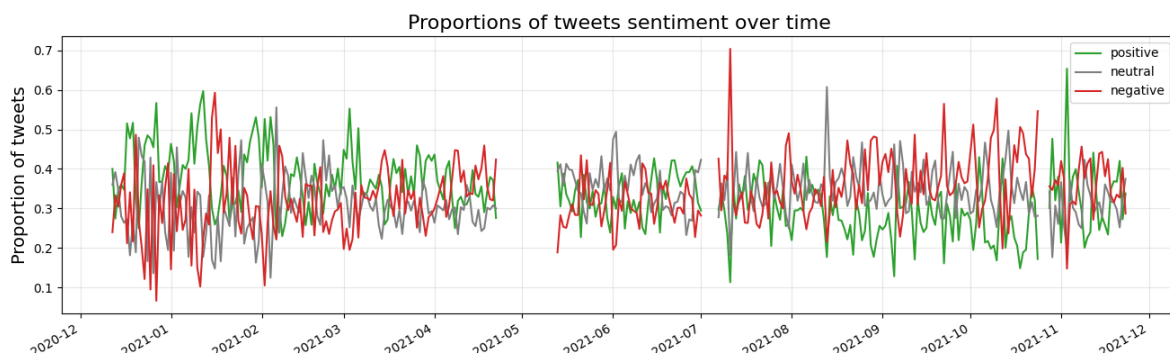


Figure 9: Variation of tweet sentiment proportions over time

6.3 Logistic Regression Models

6.3.1 Word Embedding

The pre-processed tweets, both with and without nouns, are vectorized using both Count Vectorization and Term Frequency-Inverse Document Frequency (TF-IDF) Vectorization.

The count vectorizer converts the corpus of pre-processed tweets into a matrix of token counts, with each row representing a tweet and each column representing a token. The values at each entry represents the frequency of the token within a tweet. The TF-IDF vectorizer not only takes into account the frequency of a token within a tweet, but also measures the importance of the token using its frequency across the corpus (Inverse Document Frequency).

The more frequent a token appears in other documents, the less important it is. The TF-IDF measure for each token t in document d is given by

$$\text{TF-IDF}(t, d) = \frac{\text{count of } t \text{ in } d}{\text{total number of tokens in } d} \cdot \log \left(\frac{\text{total number of documents in corpus}}{\text{number of documents containing } t} \right)$$

For both vectorizations, the maximum number of feature is set to 2,500 to keep run time low, while keeping enough features to build an accurate model in the next sections. The maximum feature parameter keeps the most frequent 2,500 tokens.

These two vectorizers are chosen as it is observed from the word clouds that different sentiments can be characterised by high frequencies of different sets of words. For example, “thank” can be expected to have high frequency in positive tweets while “no” would be frequent in negative tweets. Hence, obtaining measures on the tokens in terms of word frequency would be useful for modelling.

6.3.2 Model Training

Four multinomial logistic regression models are built. The pre-processed text data, with and without nouns, are each vectorized by both the Count and the TF-IDF Vectorizers. For training, the dataset is divided into 80-20 train-test splits. The logistic regression models are fitted against the sentiment labels generated in section 5.2. Table 1 summarises the model performances:

Pre-processed Text	Vectorization	Training Accuracy	Test Accuracy
no noun	Count	73.45%	71.17%
no noun	TF-IDF	73.25%	71.16%
with noun	Count	77.02%	74.60%
with noun	TF-IDF	76.72%	74.58%

Table 1: Summary of model accuracy

Figure 10 presents the confusion matrices for the trained models, predicting on the test set.

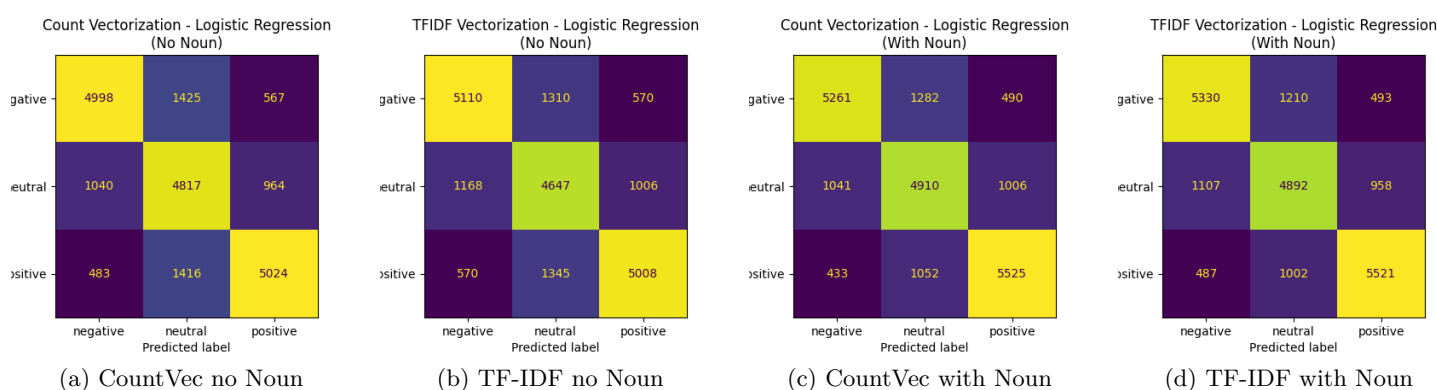


Figure 10: Confusion matrices of fitted logistic regression models

The models trained on data with nouns perform better than models trained on text without nouns. For the same pre-processing strategy, the logistic regression models achieved a similar level of accuracy between Count and TF-IDF Vectorization. This is surprising at first, as it was expected that the “inverse document frequency” component of TF-IDF should help to provide a better vectorization by penalising words that appear too often across tweets. However, the TF-IDF vectorizer might be limited in this case as many of the tweets, regardless of sentiment labels, contain common words such as “dose”, “get”, etc. The tweets are also fairly short due to the truncation issue discovered in

section 4, which reduces the benefit of the TF-IDF Vectorizer. The average number of tokens per document is 10.24 with nouns and 4.23 without nouns.

6.3.3 Sentiment Predictions vs. Vaccine Manufacturers and Time

The best model with the highest test accuracy, trained on count vectorized text with nouns, was used to predict the sentiment labels for the full dataset. This prediction is compared against the sentiment labels generated by [5] through doing the same sentiment analysis as in sections 6.1 and 6.2, as shown in Figures 11 and 12. The conclusions appear to be the same, that Moderna and Covaxin has the highest proportions of positive tweets while Oxford-AstraZeneca has the highest proportions of negative tweets.

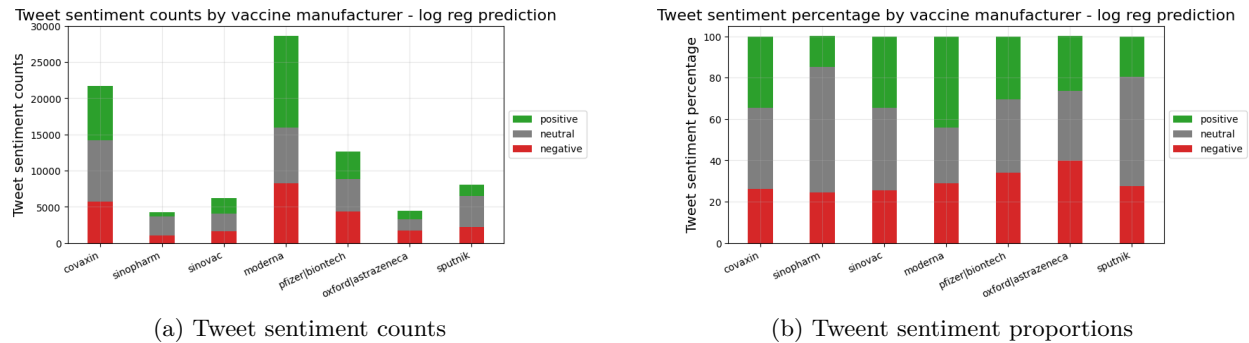


Figure 11: Count and percentage of tweets with sentiment labels predicted by logistic regression model, grouped by vaccine manufacturers

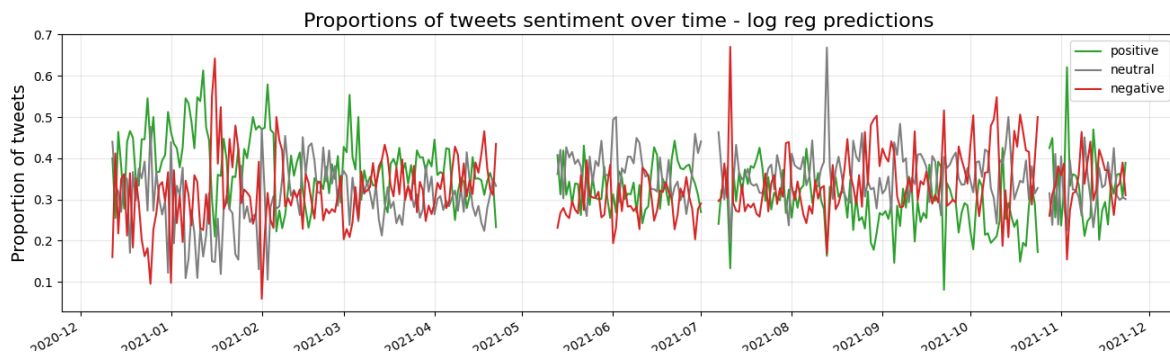


Figure 12: Variation of tweet sentiment proportions over time - predicted by logistic regression model

6.3.4 Comparison with Others' Works

While the accuracy of this model cannot be compared directly with other's work as the mechanism to generate the labels in section 5.2 is different to what others has done, it is still useful to compare the overall methodology. Looking at what other people have done on this dataset on Kaggle, two examples are particularly worth mentioning:

1. Instead of using a model trained on tweets to generate the sentiment labels, the author of MK IV Supra TFIDF used the `SentimentIntensityAnalyzer` provided by the `nlTK` library to create the labels. This sentiment analyzer uses a rule-based system by giving sentences scores using a pre-defined set of word-score mapping. This author only considered positive and negative sentiment labels, and used TF-IDF to vectorize the text, without limiting the maximum number of features. This author built a random forest classifier that achieved 86.1% test prediction accuracy.
2. The author of Sentiment Analysis of Covid19 Vaccination tweets creates the labels by first finding a separate dataset of tweets with sentiment labels. The author fine-tuned a BERT [3] model and then made predictions on

the COVID-19 All Vaccines Tweets dataset [7]. From their analysis on “Variance with Time”, it also appears that the sentiment is trending towards the negative side, agreeing with findings in section 6.2.

7 Summary

7.1 Findings

1. Tweets have much more complex structures than text data in books, speeches and articles. It contains less structured grammar, and more categories of symbols such as emojis, hash tags and user tags...etc., which could vastly change the sentiment of the tweet.
2. Moderna and Covaxin appears to have many more tweets than other manufacturer, and they have the highest proportion of positive tweets. Sputnik and Sinopharm have the lowest proportion of positive tweets and Oxford-AstraZeneca has the highest proportion of negative tweets.
3. Although the proportions of positive, negative and neutral tweets are fairly similar, the proportion of negative tweet sentiments is trending up while positive tweet sentiment is trending down slowly. We can also observe points in time when negative sentiments vastly outweighs positive sentiments, which could be linked with negative news such as vaccine safety concerns.
4. Logistic regression models with features embedded using TF-IDF and Count Vectorization performs similarly. The models using pre-processed text with nouns kept performs better than with nouns filtered out. This could mean that some of the nouns have certain bearing on the sentiment of the tweet, which makes sense considering the differences in positive sentiment proportions for different vaccine manufacturers.
5. TF-IDF vectorization might be limited by the very short documents in this dataset.

7.2 Limitations and Further Work

To avoid some of the data quality issues mentioned in 4, such as truncated tweets, missing dates...etc., and to have greater control over the method of extraction, for example which hashtags to be searching for, further work could be done with datasets created directly using the Twitter API. More advanced models could also be explored, for example Recurrent Neural Networks, specifically the Long Short-Term Memory architecture, as well as Transformers.

References

- [1] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification*, 2020. DOI: 10.48550/ARXIV.2010.12421. [Online]. Available: <https://arxiv.org/abs/2010.12421>.
- [2] F. Cascini, A. Pantovic, Y. A. Al-Ajlouni, *et al.*, “Social media and attitudes towards a covid-19 vaccination: A systematic review of the literature,” *eClinicalMedicine*, vol. 48, Jun. 2022, ISSN: 2589-5370. DOI: 10.1016/j.eclinm.2022.101454. [Online]. Available: <https://doi.org/10.1016/j.eclinm.2022.101454>.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018. DOI: 10.48550/ARXIV.1810.04805. [Online]. Available: <https://arxiv.org/abs/1810.04805>.
- [4] Y. Liu, M. Ott, N. Goyal, *et al.*, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019. DOI: 10.48550/ARXIV.1907.11692. [Online]. Available: <https://arxiv.org/abs/1907.11692>.
- [5] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, and J. Camacho-collados, “TimeLMs: Diachronic Language Models from Twitter,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 251–260. DOI: 10.18653/v1/2022.acl-demo.25. [Online]. Available: <https://aclanthology.org/2022.acl-demo.25>.
- [6] Ministry of Health and Family Welfare, Government of India, 2021. [Online]. Available: <https://www.cowin.gov.in/> (visited on 12/31/2022).
- [7] G. Preda, *COVID-19 All Vaccines Tweets*, 2021. DOI: 10.34740/KAGGLE/DSV/2845240. [Online]. Available: <https://www.kaggle.com/dsv/2845240>.
- [8] U.S. Food & Drug Administration, *Initial Results of Near Real-Time Safety Monitoring of COVID-19 Vaccines in Persons Aged 65 Years and Older*, Jul. 2021. [Online]. Available: <https://www.fda.gov/vaccines-blood-biologics/safety-availability-biologics/initial-results-near-real-time-safety-monitoring-covid-19-vaccines-persons-aged-65-years-and-older>.

This work is completed independently.