Sentiment Analysis of Covid-19 Vaccine Related Tweets and Their Relationship With Vaccination Rates

Group 8

1 Introduction

This project aims to predict Twitter users' sentiment towards Covid-19 vaccines among English-speaking countries using Natural Language Processing (NLP) machine learning techniques and neural networks. In other terms, the main objective of this report is to assess whether statements on Twitter can help us understand the general public's opinion towards Covid-19 vaccines. By drawing a comparison to data on vaccination rates, these results may also partially help to explain patterns and fluctuations in vaccination rates within the regions studied.

This project may be useful in guiding policymakers or health organisations to make recommendations to deal with the public eye's hesitancy or target negative feelings and low trust. It can also guide them into knowing which demographic to focus their strategy on, based on the rates of negative sentiments.

We will first review the literature which relates to our topic. After training our selected models and analysing their performance on our validation sets, we will then fine-tune our best model and fit it to our test set to see how well it can predict tweeters' sentiments towards the vaccine. Finally, we will discuss how our results correlate with the vaccination rates in certain English-speaking countries.

2 Literature Review

Since the development and distribution of Covid-19 vaccines is a very recent event, relevant background material dates from the last 12 months. Despite this time constraint, we have found multiple research papers that relate closely to our topic and can be used as a backbone for our project and analysis.

The first paper collected tweets from a large-scale COVID-19 Twitter chatter data set (Lyu et al., 2021) which contained one or more keywords related to 'vaccine' (e.g., vaccination, vaccinated, etc.). The methodology used in this paper would differ to the approach we have chosen but the results may be of interest to us — the paper found that the sentiment was increasingly positive in general.

Another paper by Ansari and Khan (2021) collected data via the Twitter API and used the Textblob library for sentiment analysis which we opted to use for our data analysis. In this paper, however, the researchers found that tweets regarding Covid-19 vaccines were mainly negative in tone and highlighted a significant difference in sentiments between male and female tweeters.

Finally, an article by Dua (Towards Data Science, 2021) used Textblob API and word cloud visualisations to investigate tweets related to Covid-19 vaccines. Using the same dataset and a similar methodology to what we chose for our analysis, Dua found most tweets to be neutral. However, she also found that tweets were more positive before the major vaccine rollout between February and March 2021.

This report will investigate these claims and determine if the results align with what previous papers have synthesised, adding to the discourse surrounding global vaccination.

3 Data

3.1 Datasets Overview

The first dataset was collected from www.kaggle.com/gpreda/all-covid19-vaccines-tweets. The data was gathered via the Twitter

API with Tweepy Python package from December 12th 2020 – November 23rd 2021. It consists of 228,207 tweets regarding the Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/Astrazeneca, Covaxin and Sputnik V vaccines. The locations captured included 25,000 cities around the world.

The second dataset was collected from www.kaggle.com/gpreda/covid-world-vaccination-progress and consists of daily vaccination rates globally. The data was collected from December 4th 2020 – January 25th 2022.

3.2 Data Cleaning

Firstly, the location data was cleaned using python Regular Expression library, to extract the country names from long-name locations and emoji symbols. Afterwards, English-speaking courtiers were chosen for the analysis, and the cities data for these countries was collected, to identify country location from locations that only have city names. Lastly, the cities and countries data were merged to create a

$$'processed_location'$$

variable with just chosen country names in it.

Text data from social media platforms like Twitter is often considered 'dirty' as it can contain spelling or punctuation errors, hashtags, at signs, and emojis. To maintain consistency and facilitate the pre-processing stage, the data had been lower-cased. Using the 're' module, the at signs have been removed along with the word attached to it, which is the twitter user's username, as it generally brings no additional value to the textual data.

However, with hashtags, only the symbol was removed since hashtags often contain relevant information as they are used to indicate the tweet's pertinence to a specific subject. Indeed, they help to better communicate the topic addressed or the user's feelings. In our dataset the hashtags often contain keywords directly related to coronavirus or vaccines. Links and emojis have been entirely removed as they are not relevant for our analysis.

Figure 1 below is a simple visualisation representing average Tweet retweet by country. Retweets serve as a good measure of engagement which also has implications for sentiment; e.g. people who feel strongly about a statement are more likely to retweet it.

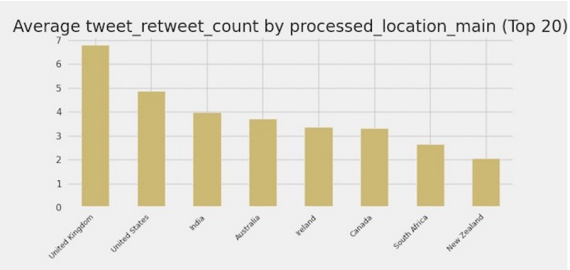


Figure 1: Retweet Count per country

3.3 Data Pre-Processing

Pre-processing the clean textual data using techniques known as tokenization, stopword removal, and lemmatization will make the text more consistent and will lead to better performance downstream.

Tokenization is common when using text data, as it splits a document into smaller and more meaningful subunits, or tokens. This can be implemented using the spaCy library, which can also be used for stopword removal. This means more focus can be applied to the meaningful parts of the text and thus potentially increase classification accuracy of the project. Stopword removal is generally avoided when doing certain NLP related tasks, but in this particular case it is a useful technique to apply. The processed tweet content is stored into a csv file to cut down on processing time.

We then lemmatize our data. Such text normalisation is useful because it makes the text more understandable for the model as well as reducing the number of unique words in the data, bringing down training time. This technique, applied in this project via the TextBlob library, is particularly useful for weak labelling. Snorkel assigns weak, provisional labels to the tweets based on whether they contain certain keywords that are matched to negative, positive, and neutral sentiments. By reducing the tweets to root words, the list of keywords can be much more concise and effective while still catching all of the instances where they appear in a particular tweet.

4 Labelling

4.1 Manual Labelling

Our dataset did not contain any label, so we decided to manually label a portion of our dataset to use as part of our training, validation and test sets. We randomly selected 1050 rows from our clean dataset to divide and annotate as 'positive' which would take the value 1, 'neutral' as 0, and 'negative' as -1. By manually labelling part of our data, we hoped to train our models to make correct predictions.

This task proved to be quite difficult. Many tweets had a mainly negative tone and used negatively connoted words, even though the sentiment expressed concerning the vaccine was positive. Since we wanted to correctly identify the sentiment expressed towards the Covid-19 vaccine and not the overall sentiment expressed in the tweet, we had to take into account the underlying assumptions of each tweet and the context – was it a negative sentiment towards the vaccine or was it a negatively connoted tweet defending the vaccine? If it were the former, it would be negative, if it were the latter it would be labelled as positive for our analysis. As such, there may be some human error where the interpretation of the context was complicated.

Many tweets did not communicate any sentiment and were simply facts about the Covid-19 vaccines or general questions about the vaccine – these were labelled as neutral. There were also tweets which only contained one word such as 'ok' or no word at all, simply a response tweet containing previous urls or hashtags. These were also labelled as neutral, as they provided no insight into sentiment.

4.2 Weak Labelling using Snorkel

Snorkel was used to label the rest of the training dataset (306,566 tweets). Snorkel is a programmatic building technique which allows labelling of large datasets. We applied Snorkel's Labelling Function programmatic operation, which uses supervision techniques that require heuristic rules for labelling data (weak supervision) (Snorkel.org, 2022). Figure 2 visualizes how the labelling function works.

Figure 2: Snorkel Labelling Function Visualized (retrieved from Snorkel.org)

When defining the labelling function, we used keyword matching techniques. We created lists of keywords from our manually labelled data to make a bag of keywords that define positive and negative sentences. Below are the list of keywords that were used in the labelling function rules.

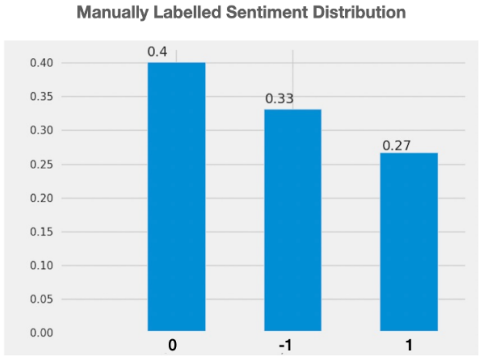
- Labelling function 1 : label all sentences with positive keywords as positive (+1) else return neutral (0). Positive keywords = ['safe', 'disinformation', 'misinformation', 'denier', 'denial', 'expert', 'studies', 'anti-vax', 'anti-covid', 'protect', 'protection', 'severe', 'spread', 'severity', 'vulnerable', 'antibodies', 'Resistance', 'science', 'proven', 'hospitalisation', 'remain', 'help', 'evidence', 'understand', 'effective', 'fight', 'mitigate', 'freed', 'save', 'support'].

- Labelling function 2 : label all sentences with negative keywords as negative (-1) else return neutral (0). Negative keywords = ['weak', 'issue', 'side', 'effect', 'ridiculous', 'efficacy', 'ineffective', 'unvaccinate', 'choice', 'sceptic', 'tyranny', 'tyrant', 'censorship', 'censor', 'conceal', 'debunk', 'exempt', 'freedom', 'narrative', 'propaganda', 'media', 'brainwash', 'conspiracy', 'force', 'mandatory', 'fail', 'reaction', 'opposition', 'coerce', 'control', 'danger', 'enslave', 'failure', 'opposition', 'harmful', 'fail'].

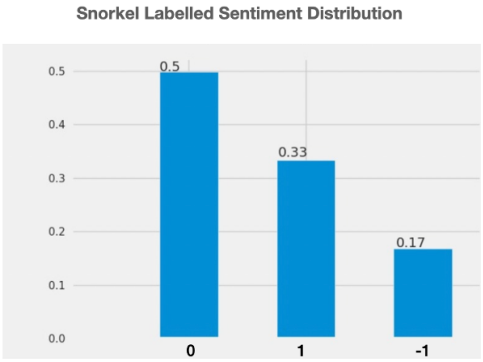
- Labelling function 3 : label all sentences with question mark (?) as neutral (0).

We can see the distribution of sentiments across our Snorkel labelled data in the following graphs.

Figures 3(a) and 3(b) show a side-by-side comparison of the distribution of sentiment for each labelling method. For both manual and Snorkel labels, the sentiment is neutral the majority of the time. However, the manual labelling revealed more negative classifications whereas Snorkel finds more positive, indicating a fundamental



(a) Distribution of sentiment of manual labeled tweets



(b) Distribution of sentiments of Snorkel labeled tweets

Figure 3: Infographics about our manual and Snorkel labelled data

difference in what defines sentiment for the two techniques. This has a significant impact on the performance of the various classification models.

We now run our model on Wordcloud to observe the most prominent patterns in our Snorkel labels.

(a) Most recurrent words in positive tweets

(b) Most recurrent words in negative tweets

Figure 4: Wordcloud analysis on our Snorkel labelled data

We can observe from the word clouds that the word 'mandate' is highly present in both negative and positive tweets, along with 'think', 'getting' and 'virus'. Those words describe general neutral actions whereas terms referring to sentiment are, as expected, distinctly pessimistic for negative tweets ('dying', 'issue', etc.) and optimistic for positive tweets ('protect', 'effective', etc.).

4.3 Training Methods

The modelling process was based around three experimental settings to incorporate the weak labels from Snorkel. Firstly, the models were trained on the manual labelled data merged with the weak labelled data (Snorkel data) and shuffled before evaluating it on only the manual labelled validation set. For the second experiment, the models were trained and evaluated only on manually la-

belled data, and thirdly, the models were trained and evaluated only on the data labelled by Snorkel.

5 Models Results

5.1 Supervised Learning

5.1.1 RNN

We decided to apply a Recurrent Neural Network (RNN) to our three subsets of training data as a supervised learning method. An RNN takes a sequential input of words and recursively loops back on itself using backpropagation, which allows information to be stored, or remembered, over time within the hidden states of the network and shared weight matrices.

The tweets need to be converted into numerical data rather than words in order to act as an input to the RNN. This is done by creating a dictionary of the data's training vocabulary, mapping each unique word in the training set to an index. Also included in the dictionary is a padding token [PAD] and an out-of-vocabulary [OOV] token. The padding token is a placeholder to ensure that the sequences put into the model are all padded to the same length (matched to the longest sequence, which in this case is 66), since individual tweets will be different. The OOV token catches any unknown words that appear in the validation and testing sets. Next, each word in the vocabulary is transformed into its 100-dimensional GloVe embedding representation along with its corresponding index. Once the values are scaled between -1 and 1, the RNN can then process this as an input, be fit to the data, and return a prediction of sentiment.

Despite using 4 LSTM layers to incorporate a strong memory component into the network, the accuracy of the model when evaluated on the manual labelled validation data was very low. After 8 epochs, the validation accuracy had already peaked at 44% with a high level of loss. The second model, trained and evaluated on the manual labels, only reached 38% accuracy. The final model though, using only Snorkel labels, achieved 88% on the validation set.

5.1.2 GloVe Embeddings for Multinomial Naive Bayes and Logistic Regression

We have applied the Multinomial Naive Bayes model as it performs well in text classification, including sentiment analysis, in addition to being fast, reliable and accurate. Similarly to RNN, the input feature is a Glove embedding representation of each word. Fundamentally, Naive Bayes's technique is based on calculating the count of each word, independent of the sentence, and selecting the predicted class with the highest probability. When applying it during two experiments to the weak and manually labelled training data and evaluating it on the manually labelled validation set, the accuracy score stagnates around 40%. However, when applying Multinomial Naive Bayes to the Snorkel labelled training set and validation set, the accuracy improves and reaches 45%. Performing grid search on each of the three experiments did not improve the accuracy score.

We wanted to run a discriminative model to compare results obtained with the Naive Bayes generative model. Accordingly, we have used Logistic Regression to run all three experiments. The model performs best on the training and validation Snorkel labelled sets, with an accuracy score of 44.5%. Its worst performance occurs when fitting it on all manually labelled sets with 33% accuracy. Finally, when fitted to the weakly labelled training and manually labelled validation set, the accuracy score averages out to 40%. According to the results, the model is highly biased towards neutral sentiment and wrongly classifies a majority of tweets as neutral. This can be explained by class imbalance. In three out of two experiments, the model fails to successfully classify negative tweets; which is a trend that can also be observed with Naive Bayes. Although Logistic Regression's performance scores do not drastically differ from Naive Bayes', the latter delivers higher and more consistent results.

5.2 Unsupervised Learning

Flair, Textblob and VADER algorithms were used for further sentiment analysis of the dataset.

	On Snorkel Labelled Training Set			
	Snorkel Label	VADER	TextBlob	Flair
Positive (1)	**102,546**	232	121,591	64,891
Neutral (0)	**152,847**	306,591	120,896	65,458
Negative (-1)	**51,802**	372	64,708	176,846
	On Manually Labelled Training Set			
	Sentiment	VADER	TextBlob	Flair
Positive (1)	**172**	1	254	126
Neutral (0)	**250**	628	238	136
Negative (-1)	**207**	0	137	367

Figure 5: Predicted labels of VADER, TextBlob and Flair models

5.2.1 VADER

We start by feeding our snorkel labelled training set into the VADER algorithm and calculate the polarity scores of each word in our cleaned tweets. We then get the overall compound score and sentiment for each tweet. According to VADER, the dataset contains a substantial amount of neutral tweets — with 306,591 tweets labelled as neutral and only 372 and 232 tweets labelled as negative and positive respectively.

This is a surprising result. Our Snorkel labelled data also contains more neutral tweets than positive or negative tweets, but it is not as drastic a difference. We want to check if this polarising result is due to a mis-labelling on Snorkel's part or if VADER is not able to pick up on the subtleties of our dataset.

We then feed our manually labelled training set into VADER — it again labels the majority of the tweets as neutral (628 out of 629), only labels one tweet as positive and finds no negative tweet. This contradicts our manually labelled sentiments — we annotated 250 as neutral tweets, 207 as negative tweets and 172 as positive tweets. VADER is not performing as well as hoped with the requirements of our dataset. Since the results are not satisfactory, another sentiment analysis algorithm is applied to our training sets and evaluated.

5.2.2 TextBlob

A function is defined to capture each tweet's polarity: if Textblob calculates the polarity to be over 0.1, it would classify the tweet as positive, if the polarity falls between 0.1 and -0.05 it would be classified as neutral, and it would classify any other tweet whose polarity is below -0.05 as negative.

We then feed our Snorkel labelled data to Textblob. Textblob finds a majority of positive tweets in our dataset, closely followed by a large number of neutral tweets and labels a smaller number of tweets as negative. These results seem more in line with the results of the weak-labelling algorithm. Even though Snorkel labelled more neutral than positive tweets, the sentiments' range of values are similar. This is a great improvement from the performance of the VADER algorithm.

We now want to feed our manually labelled data into TextBlob to see if the performance varies. The model predicted more positive tweets than neutral or negative tweets. In our case, we classified more tweets as neutral, then quite a few negative and less positive tweets. However, the range of values was similar and it appears that Textblob performs better than VADER when distinguishing between neutral/positive/negative — it does not simply label a majority as neutral.

5.2.3 Flair

We use TextClassifier to predict the sentiment of each tweet in our snorkel labelled training set. We first define a function to return 1 if the sentiment detected is positive and if the model is over 70% confident that it is positive; -1 if the sentiment detected is negative and the model is over 70% confident that it is negative; and 0 (neutral) otherwise.

We then feed our training set to the model and extract the Flair predicted sentiments. The Flair algorithm labelled a majority of the tweets as negative and labelled a similar number of tweets as positive or neutral. These results contradict with Snorkel's labels — Snorkel found a majority of neutral and positive tweets, and labelled almost three times as less tweets as negative than Flair. We want to investigate these results by feeding our manually labelled training set into Flair to see if this contradiction is due to Snorkel mis-labelling our data or if Flair does not perform well for the requirements of our analysis.

When running our model with manually labelled data, the difference between Flair and Snorkel's labelling stems from the distribution between negative and neutral tweets. Indeed, both models labelled a similar amount of tweets as positive. However, Snorkel labelled more tweets as neutral while Flair predicted more tweets as negative.

6 Models Evaluation

6.1 Supervised Learning

Our evaluation method is a classification report (Figure 6). We compared all model's accuracy, precision, recall, and F1 score across the three experiments.

Model/Label	Label	Precision	Recall	F1-Score	Support
Multinomial Naive Bayes with Glove Embedding	-1	0.35	0.11	0.16	65
	0	0.44	0.62	0.52	82
	1	0.28	0.34	0.31	53
	Accuracy	0.38			
Logistic Regression with Glove Embedding	-1	0.00	0.00	0.00	65
	0	0.41	0.98	0.58	82
	1	0.20	0.02	0.03	53
	Accuracy	0.41			
RNN with Glove Embedding	-1	0.20	0.43	0.27	30
	0	0.62	0.45	0.52	113
	1	0.45	0.42	0.44	57
	Accuracy	0.44			

(a) Accuracy, F1 score, recall and precision table for experiment 1

Model/Label	Label	Precision	Recall	F1-Score	Support
Multinomial Naive Bayes with Glove Embedding	-1	0.31	0.34	0.32	65
	0	0.48	0.44	0.46	82
	1	0.41	0.42	0.41	53
	Accuracy	0.40			
Logistic Regression with Glove Embedding	-1	0.21	0.17	0.19	65
	0	0.39	0.52	0.45	82
	1	0.32	0.23	0.26	53
	Accuracy	0.33			
RNN with Glove Embedding	-1	0.58	0.37	0.45	104
	0	0.30	0.40	0.35	62
	1	0.25	0.38	0.30	34
	Accuracy	0.38			

(b) Accuracy, F1 score, recall and precision table for experiment 2

Model/Label	Label	Precision	Recall	F1-Score	Support
Multinomial Naive Bayes with Glove Embedding	-1	0.40	0.21	0.28	38
	0	0.50	0.64	0.56	91
	1	0.37	0.34	0.35	71
	Accuracy	0.45			
Logistic Regression with Glove Embedding	-1	0.00	0.00	0.00	38
	0	0.45	0.97	0.62	91
	1	0.17	0.01	0.03	71
	Accuracy	0.45			
RNN with Glove Embedding	-1	0.71	0.90	0.79	30
	0	1.00	0.81	0.89	113
	1	0.80	1.00	0.89	57
	Accuracy	0.88			

(c) Accuracy, F1 score, recall and precision table for experiment 3

Figure 6: Evaluation and performance of our final model

We found that experiment 1 (Figure 6.a) and experiment 2 (Figure 6.b) have mediocre accuracy scores ranging between 33-44%. Experiment 3 (Figure 6.c) presents the best result across all experiments. RNN with GloVe Embedding model is the best model predicting the sentiment analysis of Covid vaccine with 88% accuracy score. This is the final model chosen to fit on the unseen test set.

6.2 Unsupervised Learning

Arguably, VADER performs the worst out of all three, but since it's unclear how well Snorkel fits our data or how accurately it labelled our tweets for the purpose of our analysis, comparing Flair and TextBlob's performance against our pre-trained labels may not be particularly insightful. Indeed, they might be performing better than Snorkel, but this cannot be verified under the current project scope.

VADER and TextBlob predicted a majority of tweets to be neutral, while Flair predicted a majority of tweets to be negative. By looking at the vaccination rates in a selected number of countries later on, we may be able to detect patterns or trends in the data that would be correlated with what we have found in our analysis.

7 Final Model Results and Evaluation

The following is the summary of the final model used for sentiment prediction on the test set to evaluate overall performance on the unseen data. We used an RNN model based fully on Snorkel labelled data. We used code from Bartolo (2022) related to advanced RNNS.

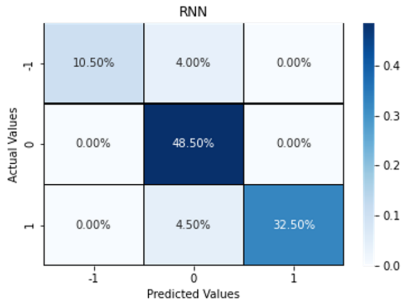
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 66, 100)           5659700
_________________________________________________________________
lstm (LSTM)                  (None, 66, 64)            42240
_________________________________________________________________
lstm_1 (LSTM)                (None, 66, 32)            12416
_________________________________________________________________
lstm_2 (LSTM)                (None, 16)                3136
_________________________________________________________________
dense (Dense)                (None, 1)                 17
=================================================================
Total params: 5,717,509
Trainable params: 57,809
Non-trainable params: 5,659,700
_________________________________________________________________
```

Figure 7: Summary of RNN model

The results of the final model demonstrate that the model has an accuracy of 93% on the test data, which is relatively high and can be considered as good performance. Furthermore, as can be seen, the model is most accurate at predicting the neutral and positive tweets while it struggles slightly to classify negative tweets. This could be due to less negative tweet data available within the dataset: just under 0.2 in proportion.

	precision	recall	f1-score	support
-1	0.86	0.93	0.89	27
0	0.99	0.89	0.94	108
1	0.88	1.00	0.94	65
accuracy			0.93	200
macro avg	0.91	0.94	0.92	200
weighted avg	0.94	0.93	0.93	200

(a) Accuracy, F1 score, recall and precision table



(b) Confusion matrix

Figure 8: Evaluation and performance of our final model

According to the confusion Matrix, 0% of data was labelled positive instead of negative and vice versa, which is a good indication that the model is able to distinguish quite strongly between these two categories. However, the model did struggle to identify the difference between positive and neutral by 4.50%, and negative and neutral by 4%. This may be due to difficulties in differentiating the categories for longer and more complex tweets.

8 Vaccination Rates Analysis

8.1 Overview of Tweet Sentiments and Vaccination Rates

The number of tweets concerning Covid-19 generally dropped between the end of 2021 and the beginning of 2022 in all countries, whereas there were peaks of Twitter activity around November 2021, which corresponds to the first cases of Omicron variant known.

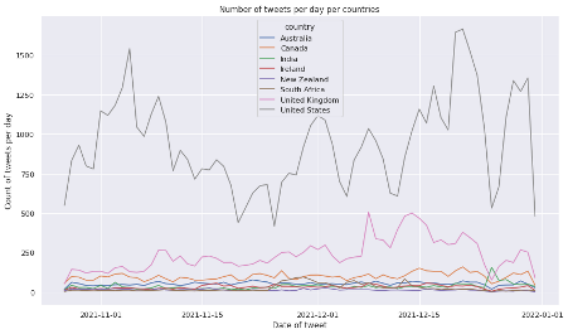
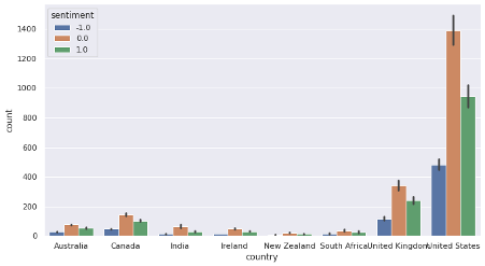
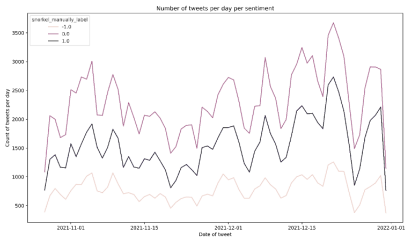


Figure 9: Evolution of daily tweets per country

As shown in Figures 10(a) and 10(b), the general trend is that most of the tweets are neutral, with positive tweets coming second and negative tweets being the least frequent. Nonetheless, the gap between negative and positive is more consequential in some countries than others, such as in the United States. The US has more negative tweets, contrary to Australia where the difference is less noticeable.



(a) Distribution of sentiment per country



(b) Evolution in the sentiment of tweets over time

Figure 10: Average sentiment of tweets

The vaccination rate in all six countries has remained quite high and stable over our analysis period, with no significant drop in the rates. However, Figure 11 displays the decrease of people getting vaccines overtime. Also, a trend that coincides is the United States' decrease of vaccination rates. Indeed, we notice that some countries such as the United States have a high amount of

negative tweets and similarly, there is a significant drop in the daily vaccination rates, even though only 65% of the population has been vaccinated.

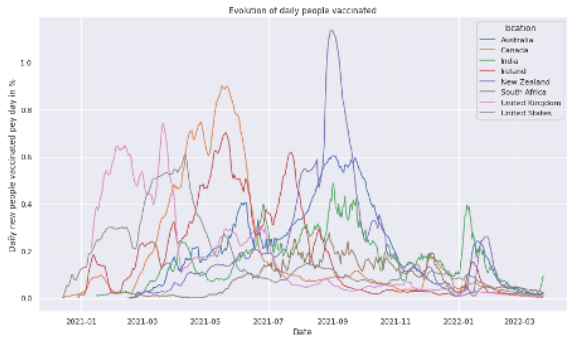


Figure 11: Evolution of daily people vaccinated per country

8.1.1 Key Takeaways

Overall, we cannot confidently draw a direct correlation between Covid-19 tweets' sentiment and vaccination rate in the six countries of interest. The graphs above show that although the share of 'anti-vaxxers' is high, it is still less prevalent than positive and neutral tweets. In addition, vaccine mandates are still obligatory as per government requirement. Nonetheless, there still seems to be some relationship between them, since in some countries a decrease of vaccination rate occurs whilst there is also a high amount of negative tweets.

9 Conclusion and Discussion

By running multiple models and experimenting with three different methods of training and evaluation, we were able to draw several conclusions about the project results.

Performance changes significantly based on which subsets of data were used, implying that the models (particularly RNN) do very well at classifying sentiment based purely on the weak labelling done by Snorkel. However, the dramatic decrease in accuracy when the manual labels are added demonstrates a discrepancy between what task the models are completing. This can be attributed to the complexity of the question posed: rather than asking for a simple sentiment analysis of each tweet, we wanted to classify the sentiment about the vaccines specifically, which requires logical reasoning. The distributions of sentiment in the manual labelled data versus the Snorkel labels shows how the models may be misinterpreting the task.

The results of the project do have interesting implications for country-level perceptions towards vaccines and Covid-19 itself. Graphs comparing Twitter activity and vaccination rates per country show similar patterns of activity, allowing us to extrapolate the classification results to the same trends and hypothesise how changes in sentiment may follow changes in vaccinations rates. High amounts of negative tweets may coincide with drops in vaccination rates, particularly for large countries where the division between perceptions towards the vaccines can be significant.

However, the public discourse on social media platform may not be representative of the overall population's sentiment towards the vaccination campaigns that have been carried out in most countries. A more well-rounded analysis could include data taken from a wider variety of sources.

10 Limitations

The main limitation of the analysis is that the research focuses on the sentiment of the tweet in the relation to vaccine (measuring how positively or negatively a person has spoken about vaccination), which can be hard to determine since algorithms mainly pick up on individual words to calculate the sentiment rather than the full context in relation to a specific subject.

Another limitation would be that drawing a direct correlation between vaccination rates and sentiment analysis is potentially inaccurate since the two datasets have slightly different timestamps (November 2021-January 2022 for tweets, and January 2021-March 2022 for vaccination rates).

The third limitation is the relatively small sample of manually labelled data, 600 for train, 200 for dev, and 200 for test, in comparison to the full dataset which has more than 300,000 entries. Furthermore, the manually labelled data had some complex and lengthy tweets which made it harder to fully grasp the sentiment. The data was also labelled by 6 people whose judgements on sentiment labelling can differ, adding further bias

into the labels.

With more time and resources, we would have collected more tweets to enlarge the analysed timeframe and get a more accurate representation of vaccine sentiment. It would have been interesting to gather tweets posted during the first few months of the vaccination campaigns, to capture initial sentiments and analyse how they have evolved over time. Moreover, it would have been beneficial to take advantage of crowdsourcing platforms and pay third parties on websites such as Amazon Mechanical Turk to increase the quantity of labels and maximise models' performance.

11 Contribution

We divided our work based on below allocation : Alizee : hand labelling, data inspection, data cleaning, data visualisation, model training Logistic Regression and Naive Bayes with Glove Embeddings, report writing, presentation. Carissa : hand labelling, weak labelling with Snorkel, model training Logistic Regression and Naive Bayes with Glove Embeddigs, report writing, presentation. Daria : hand labelling, data inspection, data cleaning, data visualisation, model training RNN Classification, report writing, presentation. Leila : hand labelling, data pre-processing, model training RNN Classification, report writing, presentation. Noemie : hand labelling, model training with Vader, Textblob, and Flair, data visualisation, report writing, report editing, presentation. Sina : hand labelling, data visualisation, researching the link between vaccination rates and sentiment analysis, Flair, graphs, report writing, presentation.

References

Ali, W.W., Malik, A., Basray, R., Haq, W., Malik, M.W., Ranjha, M.A., Chaudhry, A., Ashraf, N., Ali, H., Khan, M.A., Ansari, J.A. and Ikram, A., 2021. Assessment of COVID-19 linked fear perception in the community of Pakistan, 1 June to 31 July 2020. Global Biosecurity, 3(1), p.None. DOI: http://doi.org/10.31646/gbio.119

Analytics Vidhya (2021) Creating Customized Word Cloud in python. Available at: https://www.analyticsvidhya.com/blog/2021/08/creating-customized-word-cloud-in-python/. Accessed: 19 March 2022

Andrade, F. (2021, March 13). Ways to Tokenize Text in Python. Retrieved from Towards Data Science: https://towardsdatascience.com/5-simple-ways-to-tokenize-text-in-python-92c6804edfc4

Bartolo, Max (2022) Lecture 12: Advanced RNNs.

Cambridge University Press. (2008). Retrieved from Standford.edu: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Dua, S. (2021, March 22). Towards Data Science. Retrieved from https://towardsdatascience.com/sentiment-analysis-of-covid-19-vaccine-tweets-dc6f41a5e1af

Jain, S. (2018, February 11). Retrieved from https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/

Kohli, S. (2019, November 18). Medium. Retrieved from Understanding a Classification Report For Your Machine Learning Model: https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

Medium (2021) Sentiment Analysis with NLTK, TextBlob and Flair. Available at: https://medium.com/analytics-vidhya/sentiment-analysis-with-nltk-textblob-and-flair-a321d1460867. Accessed: 20 March 2022.

Md Tarique Jamal Ansari, N. A. (2021). Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content. Electronic Journal of General Medicine, 2.

Robert Marcec, R. L. (2021). Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. BMJ, 1-5.

Shah, P. (2020, June 27). Sentiment Analysis using TextBlob. Retrieved from Towards Data Science: https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524

Snorkel. (2022). Get Started with Snorkel. Retrieved from Snorkel.org: https://www.snorkel.org/get-started/

Yan, Yuxin, Yoongxin Pang, Zhuoyi Lyu, Ruiqi Wang, Xinyun Wu, Chong You, Haitao Zhao, Sivakumar Manickam, Edward Lester, Tao Wu, and Cheng H. Pang. 2021. "The COVID-19 Vaccines: Recent Development, Challenges and Prospects" Vaccines 9, no. 4: 349. https://doi.org/10.3390/vaccines9040349