COVID-19 Vaccination Tweet Clustering

Derek Xu, Molly Shillabeer, Ekim Karabey

Problem Definition

Twitter operates as an online town square, a place for everyone to talk about anything. This means that tweets are very useful in providing insight into the different opinions of various groups of people. COVID-19 vaccination was a very controversial topic during the pandemic that was discussed extensively online. Understanding the various ways vaccination is discussed online is valuable in informing future vaccine efforts and could be used to provide more specific recommendations to users.

The goal of our project was to create a text clustering model to understand the different ways people talk about vaccination. We do this by using a dataset of tweets collected during the pandemic that contained key words related to COVID-19 vaccines to build an unsupervised model for natural language processing. The results were then analyzed to understand the important features of each cluster.

Data Description

The dataset used was found on Kaggle and contains 228,207 samples of tweets collected between December 2020 and November 2021 that were discussing vaccines. There are 16 features: tweet id, username, user location, user description, user created date, user’s number of followers, user’s number of friends, user’s number of favorites, verified status, date of tweet, tweet text, hashtags, source, number of retweets, number of favorites, and retweet status. The only feature used in creating the model was the tweet text, but other features were used to analyze the data and help characterize the resulting clusters. Due to a limit with Twitter’s API, the text data was originally all cut off at 140 characters and had to be recovered using tweet hydration. The text had also not been processed in any way and contained hashtags, mentions, and emojis.

When analyzing the user creation dates in Figure 1, it is visually clear that there was a huge spike in the number of accounts created during the spring of 2021, coinciding with vaccination campaigns worldwide. This is suspicious behaviour because so many accounts were created and immediately started tweeting about COVID-19 vaccinations. It is possible that these are fake accounts or bots, that could be used to tweet information about where to find vaccines or possibly sway public opinion on vaccination. This finding helped us analyze our clustering model, as these spam accounts tweet similar content and could be clustered together.

For the analysis of numeric features, Table 1 shows that there is a heavy skew towards tweets not having any retweets or favorites. Even at the 75th percentile, the number of retweets is 0 and the number of favorites is only 2. A similar trend is seen in the number of friends, followers, and favorites each user has, where the majority have few and a small number of users between the 75th percentile and maximum have a lot. This aligns with our understanding of Twitter usage, the vast majority of accounts are small and don't get a lot of attention, while a small number of accounts are very popular and get a lot of attention, shown by the large maximum values. This trend of numeric features being right-skewed is also shown visually in our univariate analysis, Figure 4. Figure 3 shows that despite all the numeric features having a similar distribution, most are not heavily correlated with each other. The only 2 features that have high correlation are number of likes and retweets. There is a medium correlation between number of user favorites and friends, and verified status and number of followers. All of these correlations are expected because it makes sense that popular tweets get a lot of likes and retweets, people with a lot of friends get more likes, and being verified means you are more popular and will have more followers.

From our analysis of the tweet text, we noticed a few features in Figure 2 that could indicate an interesting distribution of content in the dataset. "Covaxin" (the Indian-manufactured COVID vaccine) was the fourth most common word after "vaccine", "dose", and "slots" which indicates that Indian users are overrepresented. This is supported by the fact that "India" is the only country name in the top 50 most common words. There are also specific relevant numbers on the list, such as 1st, 2nd, 2, 18, and 2021, which could indicate that vaccine rollouts are a common point of discussion, specifically discussion around dose numbers, age limit for vaccination, and the year of the vaccine campaign. The words "get", "got", and "getting" also indicate that getting the vaccine is a common point of discussion in all three tenses (past, present, future).

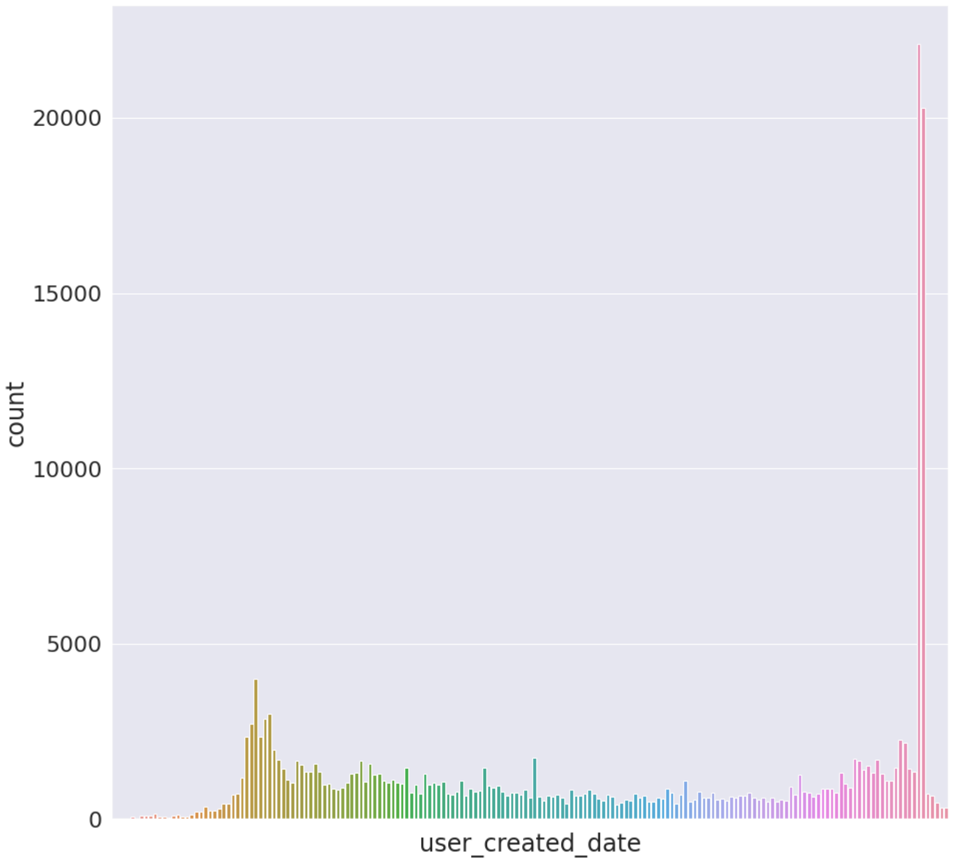

Figure 1: Number of users created by month, ranging from 2006 to 2021.

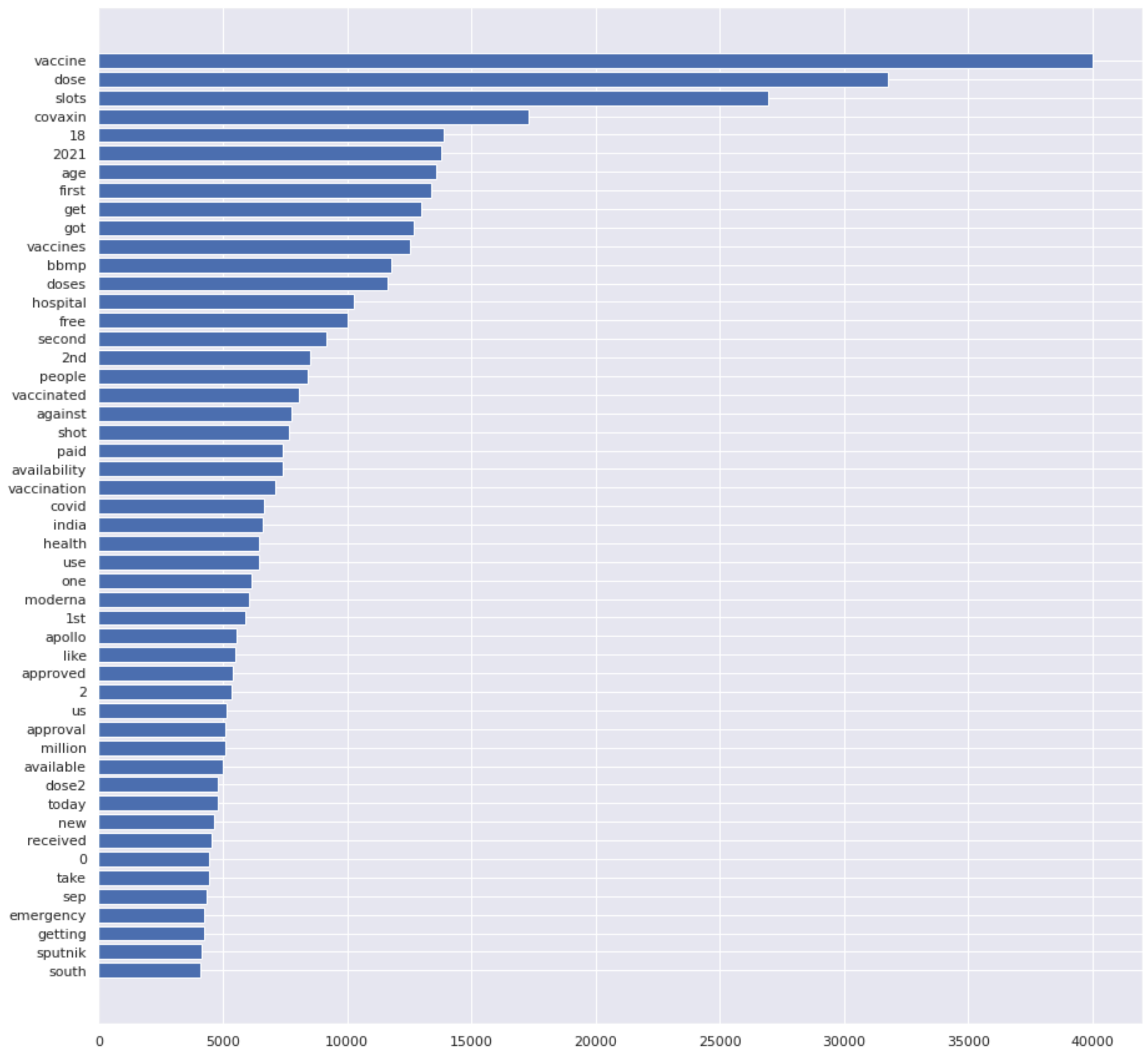

Figure 2: 50 most common words with stop words removed.

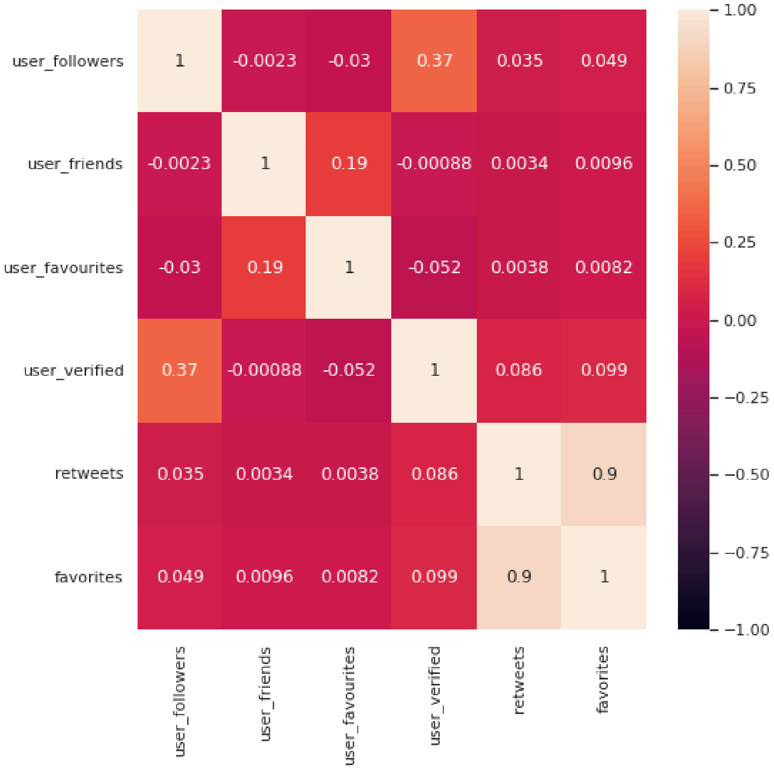

Figure 3: Correlation heatmap for numeric features.

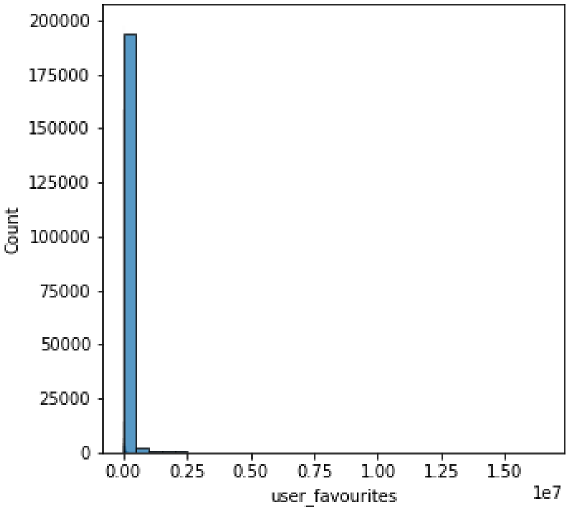

Figure 4: Univariate analysis for the number of favorites an account has. The graph is the same for the other numeric attributes, all are heavily right-skewed.

	user_followers	user_friends	user_favourites	retweets	favorites
count	199266.00	199266.00	199266.00	199266.00	199266.00
mean	109517.40	991.75	11998.47	2.54	11.12
std	895816.17	5560.53	39718.01	53.05	202.87
min	0.00	0.00	0.00	0.00	0.00
25%	57.00	17.00	57.00	0.00	0.00
50%	342.00	222.00	810.00	0.00	0.00
75%	1749.00	786.00	6861.00	0.00	2.00
max	16353048.00	582461.00	1299600.00	12294.00	54017.0

Table 1: Data distribution for numeric features.

Challenges

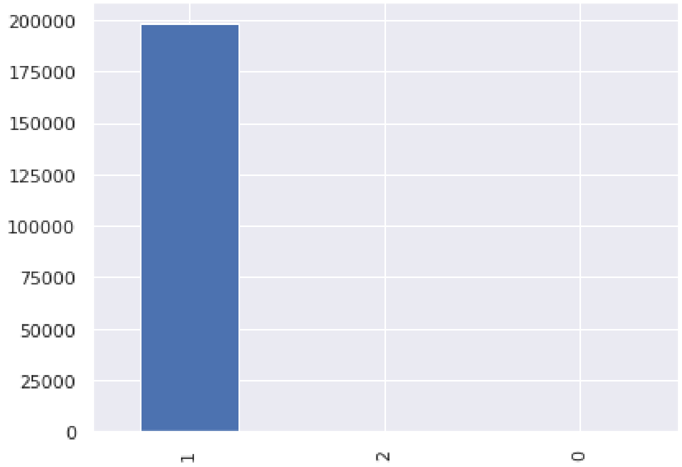

Figure 5: Sentiment from the pretrained model, showing that the resulting scores are overwhelmingly neutral (class 1).

Our original goal was to perform sentiment analysis, but this became our biggest challenge to overcome because the pre-trained model's results were overwhelmingly stacked in the neutral sentiment class. There are a few possible reasons for these results.

Firstly, the pre-trained model was trained on general Twitter datasets and not on a COVID-specific dataset. This could mean that the model is unable to recognize certain important features that are specific to COVID-19 vaccination.

Secondly, controversiality is not necessarily correlated with a negative sentiment. For example, the statement “vaccines cause autism” has a neutral sentiment (because it doesn’t express happiness, excitement, sadness, anger, etc.) but it is undoubtedly controversial. Similarly, compliance with vaccination efforts is not necessarily correlated with a positive sentiment. The statements “I got the vaccine today” or “I’m getting the vaccine tomorrow” also express a neutral sentiment.

This meant that identifying how vaccination was being discussed would require a more general text clustering approach. So, we decided to change our approach. We still decided to use clustering, but instead of trying to reproduce the sentiment results from the pre-trained model, we wanted to find a new feature such as controversiality or presence of bots by interpreting the results of a clustering model.

Methodology

Preprocessing Steps

The dataset was downloaded from Kaggle [1] and the tweet id column was uploaded to Hydrator [2] to perform tweet hydration and extract the full tweets. The 2 datasets were then merged, deleting any rows that did not have a full tweet available, which was about 13% of the dataset. These tweets may have been deleted or the accounts went private or were deleted, meaning that there is no access to the full tweet anymore.

The retweet status feature was removed because none of the samples are retweets. The tweet ID column was also removed because it was different for every tweet, providing no useful information. Time and date were separated for the user created date and tweet date. Hashtags and mentions were separated from the body of the text and put into their own features. At this point, data exploration was done, in which we created the graphs present in this report.

The text was cleaned by first fixing any encoding issues(\n or &amp;), lowercasing the text, then removing emojis, mentions, URLs, and hashtags. At this point, the data was clean enough for use in the RoBERTa pretrained model, while more processing was needed for our final model. Stopwords were removed and the text was lemmatized to reduce the total vocabulary. Question marks and exclamation points were encoded as hasQuestion and hasExclamation because they contain important information regarding the tone of a tweet. After they were encoded, all remaining punctuation was removed. Tweets that contained no words after this process were removed.

The data was then transformed into a vector space using N-grams where n=4 to preserve spatial information in the text, which was the input for TF-IDF. A vocabulary size of 700 was used for the vectorization, meaning the 700 most common words were used to create the N-grams and

TF-IDF representation of the data. This was the final, fully-processed dataset used as input for training.

Pretrained Model

RoBERTa model from huggingface was the pre-trained model we selected for sentiment analysis. The huggingface transformers library was installed using:

```
pip install transformers
```

Then, we downloaded the pre-trained model, tokenizer, and label mapping using the settings:

```
task='sentiment'
MODEL='cardiffnlp/twitter-roberta-base-sentiment'
```

The transformer model's encoding and decoding steps were run as shown on the documentation page: [3] (see relevant code block for full code). Lastly, we saved the sentiment results to a CSV file since the transformer model takes about 4 hours to run.

Unsupervised Model

K-means clustering was the final model we selected. Hyper-parameter tuning was done manually, resulting in 3 clusters, 75 initializations, and 1000 max iterations. The number of initializations is fairly large because K-means clustering depends on where the centroids start. So for a large dataset, having more initializations makes it more likely that the best clustering will be found. The cluster label for each sample was extracted after training and used to evaluate the model. K-means was selected because of its efficiency and easily-interpreted results. We attempted to use other clustering algorithms (hierarchical and spectral), but they were too inefficient and could not be implemented successfully.

The model was evaluated using silhouette score, Davies-Bouldin index, and Calinski-Harabasz index. The silhouette score looks at the distance between each sample and others in the same cluster and all others in the next nearest cluster, 1 being the best possible value and -1 the worst. The Davies-Bouldin index evaluates the similarity of clusters, so 0 is the best possible score and indicates better partitioning. The Calinski-Harabasz index (variance ratio criterion) measures the ratio between the distances of samples in different clusters and distances of samples within the same cluster, and higher values align with dense and well-separated clusters.

The model was analyzed by calculating the number of samples in each cluster, finding the 10 most common words in each, and printing out a selection of tweets from each cluster to identify trends. The graphs in Figure 6 were created by plotting the number of friends, favorites, followers, retweets, and word count for each cluster and also aided in analyzing the final result.

Evaluation

Cluster Evaluation Metrics

Silhouette Score:	0.51
Davies-Bouldin Index:	0.6
Calinski-Harabasz Index:	472 559

Tweet Distribution

Cluster	Number of Samples	Percentage
0	42,432	22.13%
1	74,138	38.67%
2	75,138	39.19%

Conclusions

The final chunk of our codebase is dedicated to displaying the resulting clusters, and analyzing the main characteristics of each.

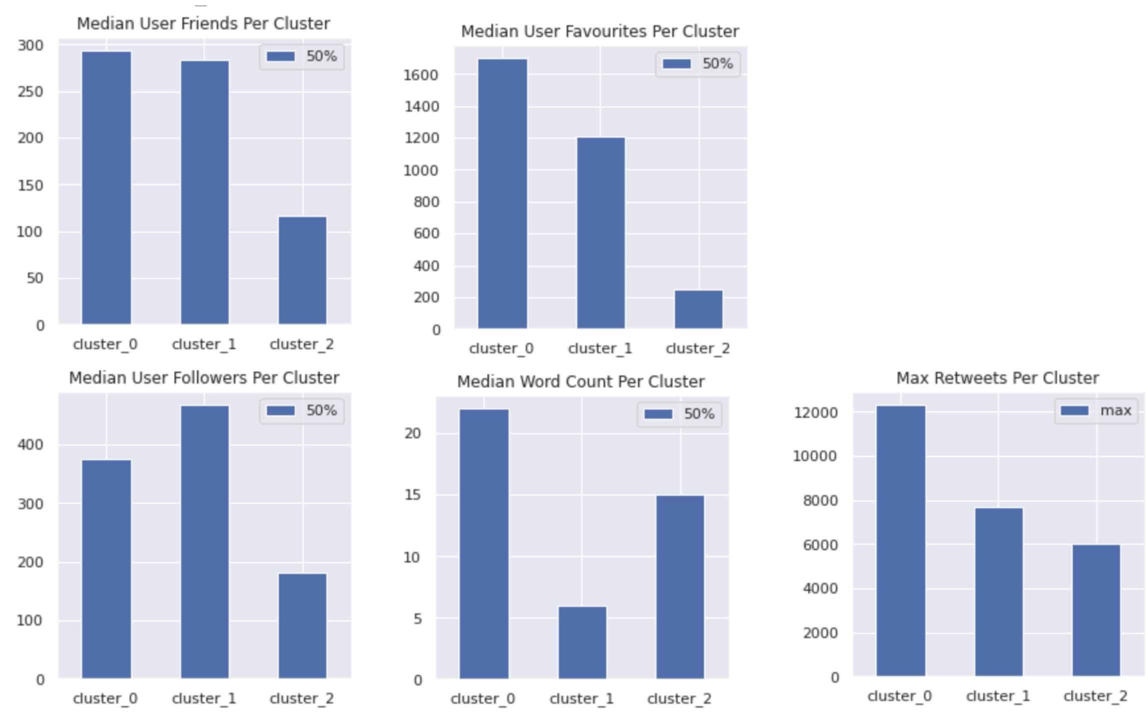

Figure 6: Median number of friends, followers, favorites, retweets, and word count for each cluster.

Before looking at the text itself, we first look at the characteristics of the tweets. From Figure 6, we can see some takeaways: cluster 0 is the most wordy, cluster 1 is the least wordy and cluster 2 falls behind on all metrics except word count.

Next, we took a look at the most commonly occurring words in each cluster. They are as follows:

- **Cluster 0:** 'hasexclamation', 'hasquestion', 'vaccine', 'get', 'dos', 'first', 'people', 'got', '2nd', 'shot'
- **Cluster 1:** 'hasexclamation', 'age', 'hasquestion', 'got', 'first', 'get', 'shot', 'slot', 'vaccinated', 'second'
- **Cluster 2:** 'slot', 'covaxin', '18', '2021', 'hasexclamation', 'free', 'hasquestion', 'hospital', '1410', 'paid'

From these insights we can try to construct a representation of what each cluster holds.

Cluster 0: Covid Related Rants - 22.13% of data
Cluster 0 is the most likely to ask questions, and has the longest median word count. It importantly has a lot of words relating to the first and second doses, as well as “people”, . Get is more used than got, suggesting an accusatory tense of grammar. E.g. “Why did the government think it was a good idea to… “

Cluster 1: Covid Related Queries - 38.67% of data
Cluster 1 is likely to ask questions, and has the shortest median word count. It contains words in past tense, which suggests the user asking about other people’s experiences. Since the tweet body tends to be brief, and the engagement is the highest, this type of tweet is likely someone looking for advice. E.g. “How old were you when you got your first shot?”

Cluster 2: Templated Bot Tweets - 39.19% of data
Cluster 2 is the least likely to ask questions, and the least likely to generate engagement from users. Interestingly, slot is the most commonly used word. Words like free, paid and 1410 (max price for a vaccine appointment set by the Indian government) appear a lot, hinting at the monetary nature of the tweets. Indeed, a manual scan of the tweets shows that the text is auto generated, and the accounts are often controlled by bots. E.g. “Vaccine slot available 14 2021 paid 1410 rs at nearby hospital.”

Future Work

For future work on this project, we would like to implement time series analysis to see how the content of tweets about COVID-19 vaccination changed over the course of the pandemic. We would also like to analyze our current model with location data and account popularity/verified status to see how they are distributed in the clusters. We could also create a new clustering model including emojis because they could be important in indicating the tone of a tweet. Lastly, we could attempt to use a large number of clusters to identify more specific trends in the data.

Team Contribution

Team Member	Tasks Completed	Percent of Project
Derek Xu	Pretrained model and sentiment analysis, text encoder (roBERTa model), word distribution visualizations	33%
Molly Shillabeer	Univariate & bivariate analysis, tweet hydration, data cleaning, clustering, model evaluation	33%
Ekim Karabey	Cluster interpretation, text encoding, hashtag and tag extraction/exploration	33%

References

[1] https://www.kaggle.com/datasets/gpreda/all-covid19-vaccines-tweets

[2] https://github.com/DocNow/hydrator

[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment