COVID-19 Tweets Natural Language Processing (MLP) & Topic Modelling

Table of Contents

Introduction

According to Chaffey (2023), approximately 4.8 billion people use social media for an average of 2 hours 24 mins daily. The high utility makes it a powerful tool for expressing opinions, engaging with others, and sharing information. Through Natural Language Processing (NLP), we can interpret and dissect the subtleties of textual information. According to Liddy (2001), NLP is a theoretically grounded class of computational methods for evaluating and representing naturally occurring texts at one or more levels of linguistic analysis to obtain human-like language processing for various tasks or applications. In this report, we will be examining tweets relating to COVID-19 and the general populace sentiment online with the application of BERTopic.

Data Pre-Processing

The COVID-19 All Vaccines Tweets data set was extracted from the Kaggle platform. This dataset contains 16 variables with over 200,000 rows of tweets. However, before performing data analysis or machine learning, data pre-processing is required to ensure an appropriate analysis or model training format.

Null value check was performed and noticed that null values do exist, which are removed subsequently. Following this, a duplicate data check was done to ensure data integrity.

```
In [8]: lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))

def process_tweet(text):
    # Remove HTML Encoding
    text = re.sub('<[^>]*>', '', text)
    # Remove URL Formatting
    text = re.sub(r'\S*https?:\S*', '', text)
    # Remove @Username
    text = re.sub('@[\w]+', '', text)
    # Identify emoticons
    emoticons = re.findall('(?::|;|=)(?:-)?(?:\)|\(|D|P)', text)
    # Remove special characters, casefold, and move emoticons to the end
    text = (re.sub('[\W]+', ' ', text.lower()) +
            ' '.join(emoticons).replace('-', ''))
    return text
```

Figure 1: Data Pre-Processing

After the above steps, Lemmatization was first performed to reduce words to the root form for grouping to ensure they are treated as a single term. Similarly, stop_words was included to remove words of non-significance and reduce dimensionality. The remaining text was then checked and stripped of special characters and formatting to ensure consistency.

Data Exploration

Bag of Words (BOW) Word Cloud

Figure 2: Positive BOW Word Cloud

In the positive BOW word cloud, the popular words are ban, lifted, stopped, congress, and vaccine. This could indicate that the ban is lifted and people are able to perform certain activities in Ahmedabad with the approval of a certain bill being stamp over, explaining the positive sentiment.

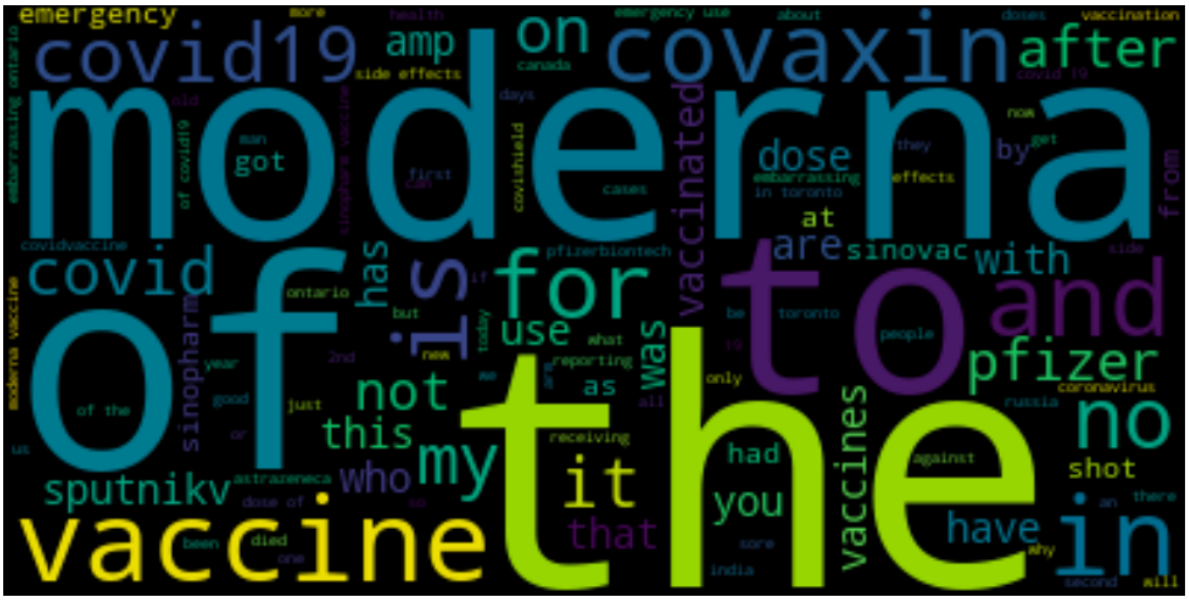

Figure 3: Negative BOW Word Cloud

In the negative BOW word cloud, the most prominent words are moderna, vaccine and other similar words if we were to exclude words of insignificance (the, to, of). Apart from concerns about vaccines, no other context appears relevant apart from emergency.

TF -IDF Word Cloud

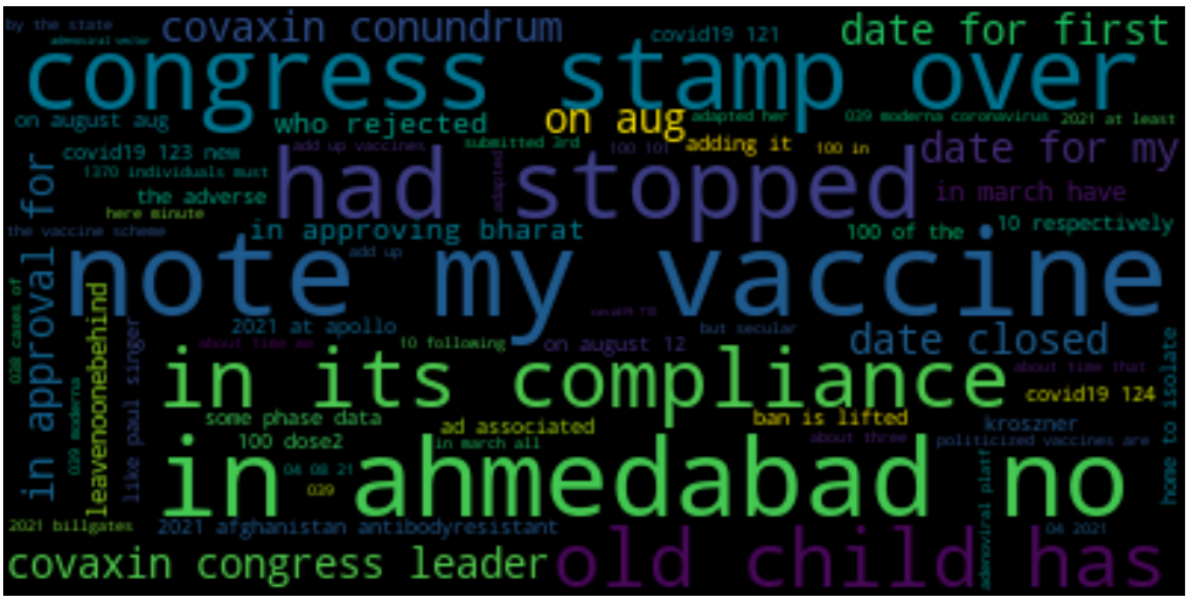

Figure 4: Positive TF-IDF Word Cloud

In the positive TF-IDF word cloud, the most common words appear to be similar to those of the positive BOW word cloud.

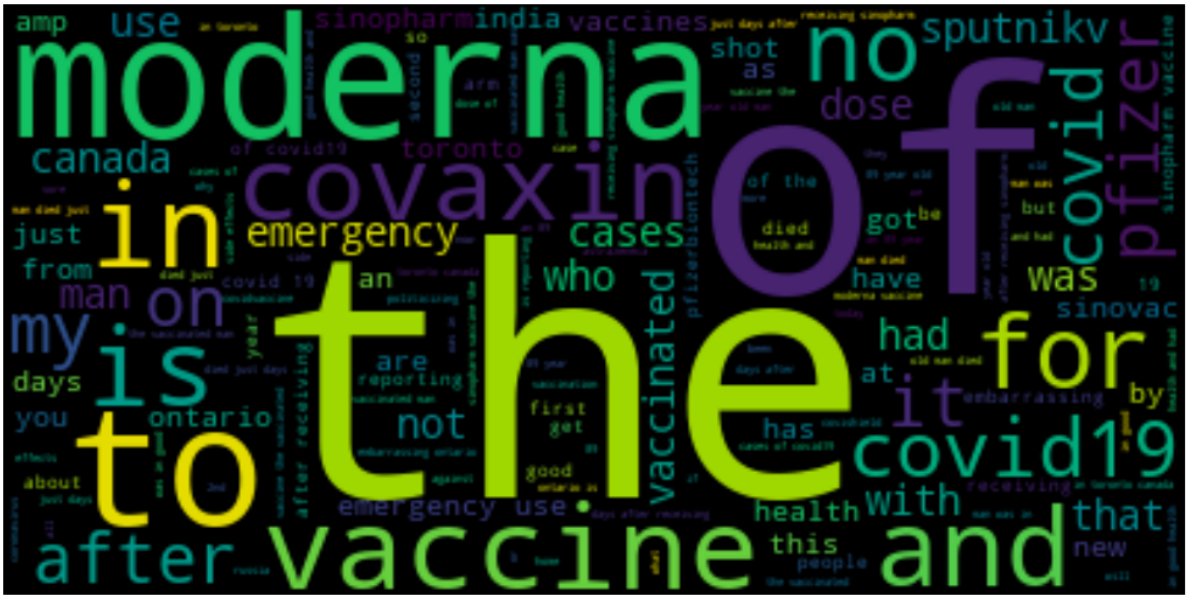

Figure 5: Negative TF-IDF Word Cloud

In this negative word cloud, we can observe various changes as canada, and vaccinated appears to be a new entry. This could indicate that most of the topics are in relation to Canada vaccination, which probably ran into obstacles, gaining negative sentiments.

Sentiment Analysis

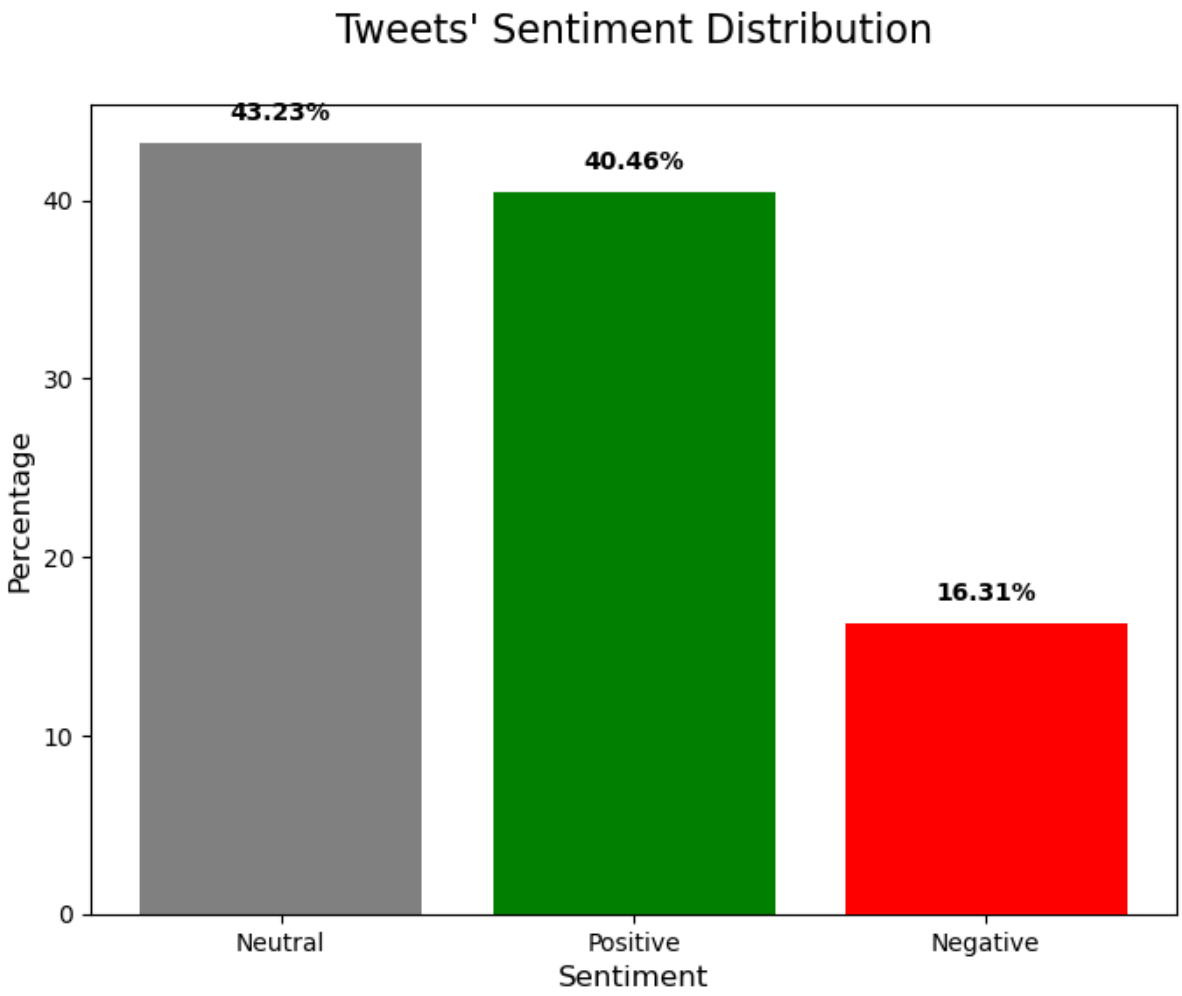

Figure 6: Overall Sentiment Distribution

From an overview, we can conclude that the general populace has a positive sentiment pertaining to the COVID-19 vaccines, as observed in Figure 6. The positive sentiment is more than twice the negative sentiment, excluding approximately half with neutral sentiments.

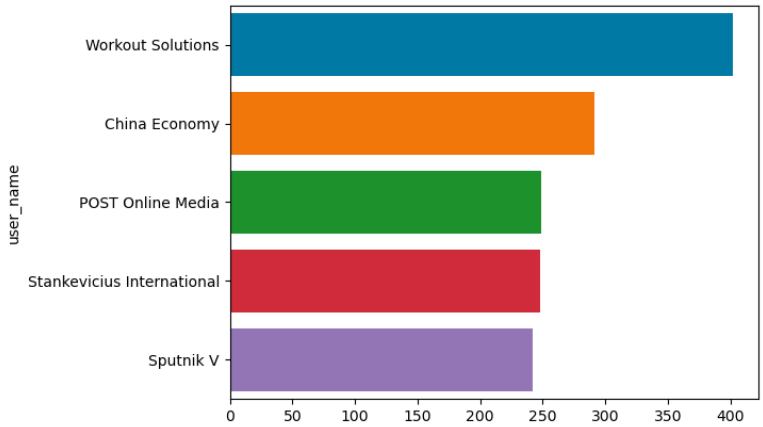

Figure 7: Top 5 Neutral Users

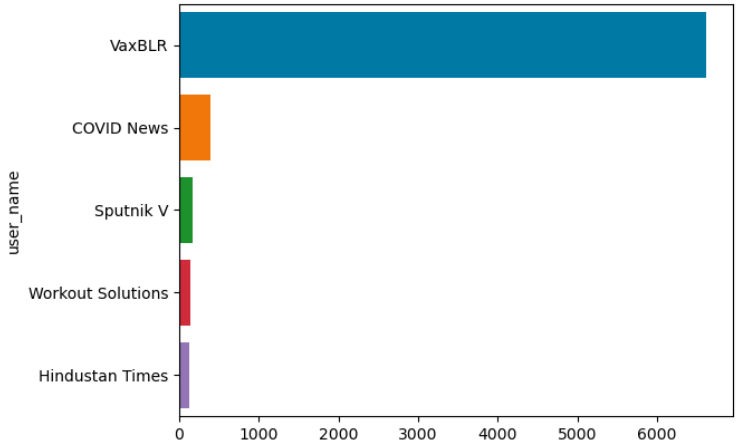

Figure 8: Top 5 Positive Users

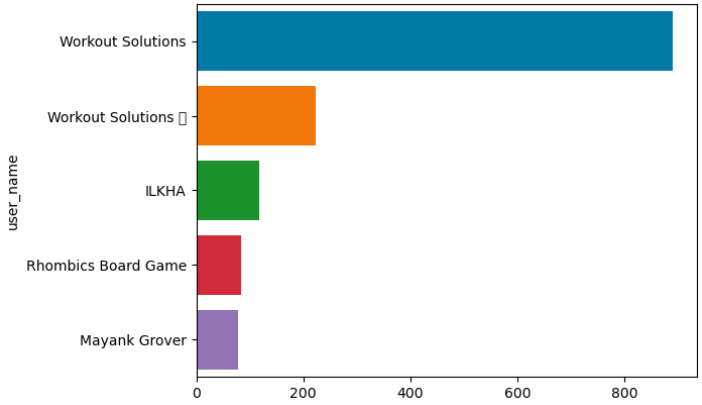

Figure 9: Top 5 Negative Users

From the above charts 7 and 9, we can observe that workout solutions is the top user for neutral and negative sentiment. This could be due to the negative news being disseminated. Therefore, an increased amount of negative and neutral tweets are made by Workout Solutions.

On the contrary, VaxBLR has a huge amount of positive tweets in Figure 8. This would require investigation as it leaps the next four combined by a huge margin. An interesting find is another user named Workout Solutions which has contributed significantly to negative tweets. This would require investigation.

Topic Modelling

Popular text analytical techniques for assessing text data include topic modelling. There are different topic modelling techniques that assume various types of linkages and constraints within datasets into account (Vayansky & Kumar, 2020). In this report, we will be applying BERTopic as our topic modelling of choice.

The BERT language model serves as the foundation for the topic modelling library known as BERTopic. Because BERTopic captures semantic meaning as opposed to conventional topic modelling approaches, which rely on counting word occurrences, it can easily handle synonyms, context, and polysemy. Its key advantage is its capacity to automatically identify subjects from massive amounts of unstructured text input without the requirement for pre-existing knowledge or specially designed characteristics. As a result, it is extremely useful for analysing huge text corpora.

However, before diving into the BERTopic model, the following libraries are used to support the mode:

1. os - file & folder management
2. pandas - general library for data manipulation
3. numpy - general library for numerical computation
4. matplotlib - supports data visualisation charts
5. seaborn - supports high-level data visualisation charts
6. re - supports pattern matching and text manipulation
7. nltk - library for processing NLP
8. wordcloud - to enable word cloud functionality
9. sklearn - general library for various data science utility
10. bertopic - BERTopic Modelling
11. umap - reduce dimensionality
12. gensim - supports topic extraction & similarity calculation

```
In [21]: # Set TOKENIZERS_PARALLELISM to false
         os.environ["TOKENIZERS_PARALLELISM"] = "false"

         vectorizer_model = CountVectorizer(ngram_range=(1, 5))

         topic_model = BERTopic(nr_topics=5, language='english', vectorizer_model=vectorizer_model,
                                calculate_probabilities=True, verbose=True)

         topics, probabilities = topic_model.fit_transform(tweet_positive.tolist())
```

Figure 10: BERT Model Parameters

In this BERT model, the ngram range is set to 5, which indicates that it will consider up to 5 consecutive word range. nr_topics=5 is set to dictate the model to identify 5 underlying topics.

Positive

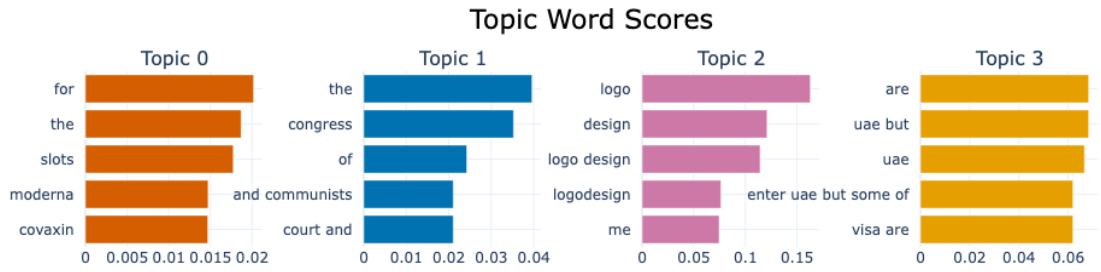

Figure 11: Positive Topic Scores

From figure 11, we can determine that various vaccine slots form the first group. This could indicate that the vaccine program is going well. Following this, congress was mentioned in a positive light which indicates that the public possibly views them in a favourable light.

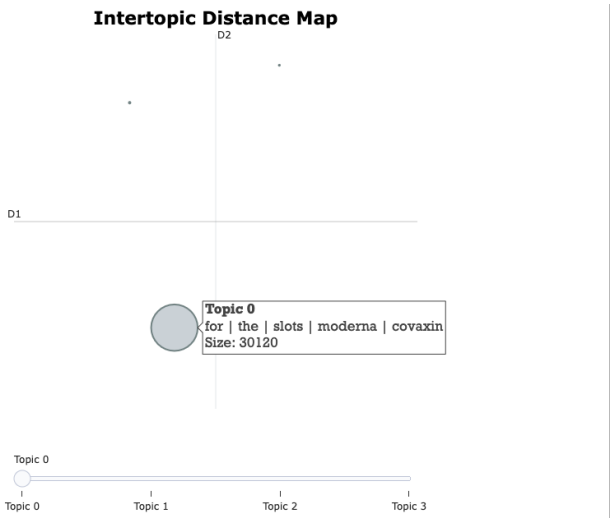

Figure 12: Positive Intertopic Map

From figure 12, we can observe the mentions of topic 0 pertaining to vaccine slots greatly outweighing the rest. This indicates that it was the talk of the town as it was newly made available to the public.

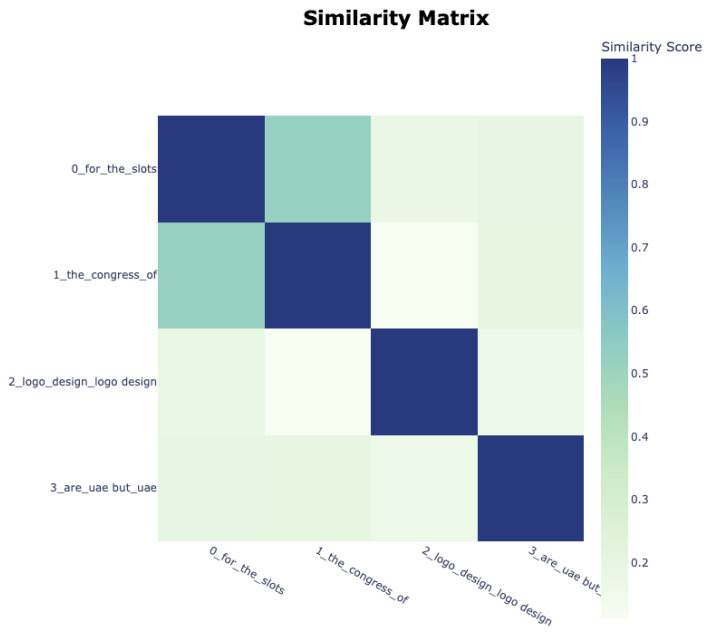

Figure 13: Positive Similarity Matrix

From figure 13, we can observe that BERTopic has clearly separated the topics. Each topic greatly differs from each other with a score of below 0.5.

Negative

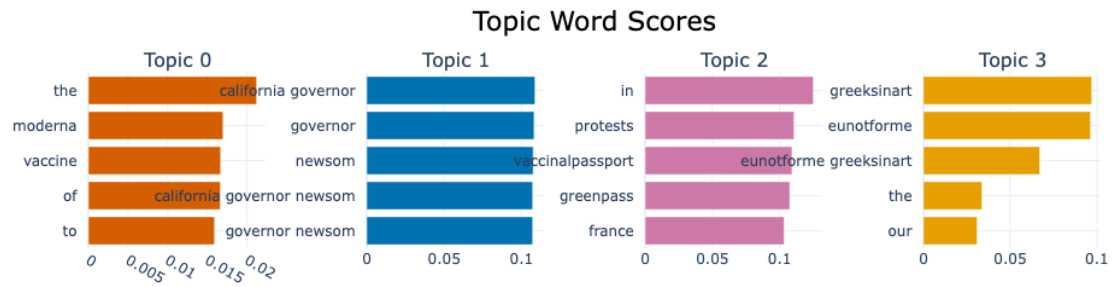

Figure 14: Negative Topic Scores

From the above, the moderna vaccine in particular was not well received online from the Tweets. Similarly, this could be said for the California officials that made up Topic 1.

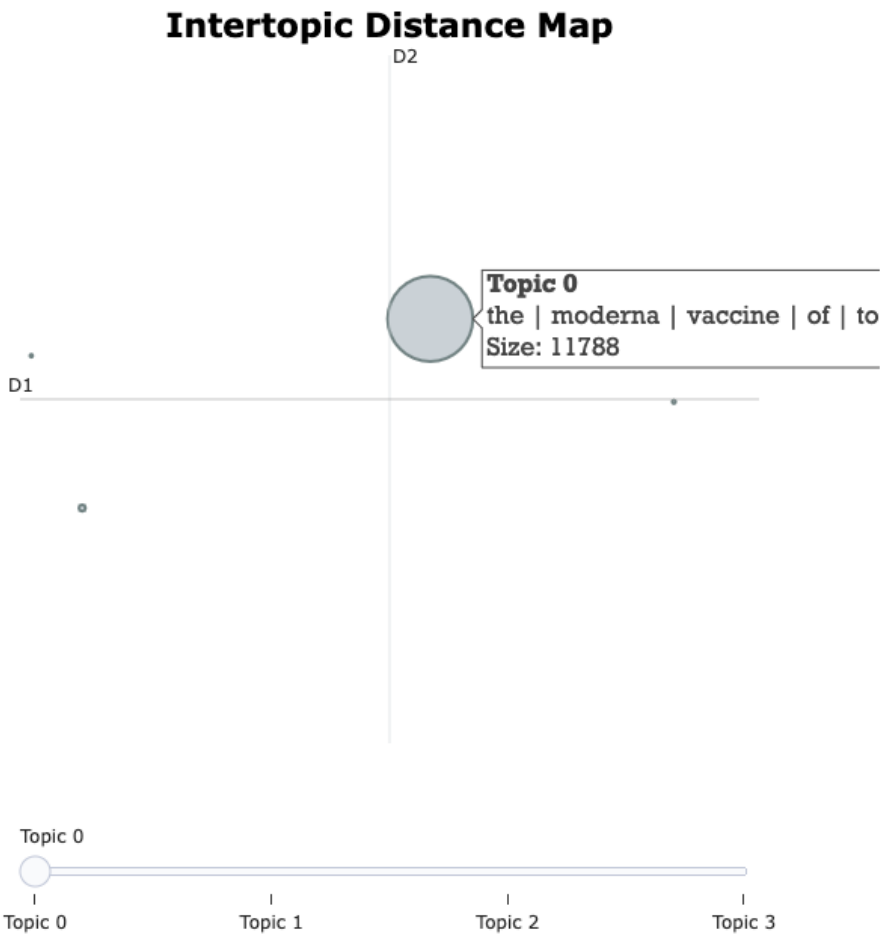

Figure 15: Negative Intertopic Map

Similar to the Positive intertopic map, vaccination was the talk of the town that greatly dwarfs the rest of the trending topics. However, moderna pushback would warrant a deeper investigation.

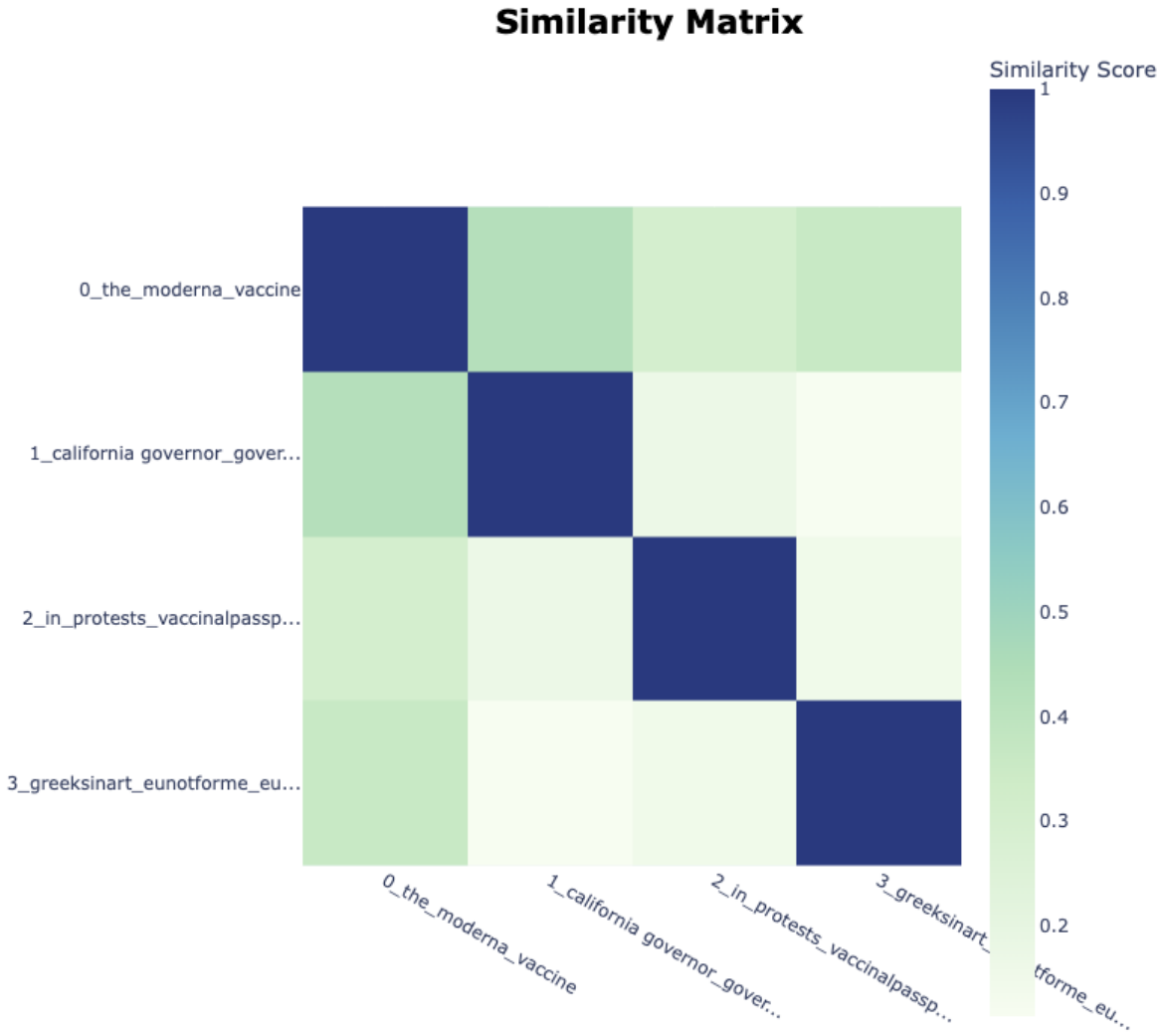

Figure 16: Negative Similarity Matrix

Similar to the positive similarity matrix, the topics were clearly defined with similarity scores < 0.5. This indicates that the topics greatly differs from one another, forming individual conversational topics.

```
# Set TOKENIZERS_PARALLELISM to false
os.environ["TOKENIZERS_PARALLELISM"] = "false"


# Get the topic-word matrix from BERTopic
topic_word_matrix = topic_model.get_topic_info()

# Extract the list of lists of topic keywords
topics_keywords = topic_word_matrix["Name"].apply(lambda x: x.split()).tolist()

# Create a Gensim Dictionary from the topic keywords
dictionary = Dictionary(topics_keywords)

# Create a Gensim Corpus from the topic keywords
corpus = [dictionary.doc2bow(topic) for topic in topics_keywords]

# Compute the coherence score using C_v coherence measure (can also try other coherence measure
coherence_model = CoherenceModel(model=topic_model, texts=topics_keywords, dictionary=dictionar
coherence_score = coherence_model.get_coherence()

# Print the coherence score
print("Coherence Score:", coherence_score)
```

```
Coherence Score: 1.0
```

Figure 17: Coherence Evaluation

From the above, the BERTopic model has a perfect coherence score of 1. This would indicate that the topics are clearly defined and that the phrases are semantically cohesive. It implies that the themes inside the text data are represented by the subjects, making them easier to grasp and interpret.

Conclusion

In summary, the BERT model is an effective technique for topic modelling as depicted in this report. This is supported by the respective similarity matrix where topics are clearly segregated. In the sentiment section, we can conclude that the general populace has a positive sentiment excluding those of a neutral stance. However, findings from the word cloud where Canada was portrayed negatively should be further examined, along with duplicate users, as fake news and bots are an issue in the current digital age.

It should be duly noted that BERTopic modelling relies on pre-trained BERT models, which may be expensive to compute. Therefore, it might be slow in real-world application. Furthermore, areas of specialisation text data or languages with limited resources may provide difficulties for BERTopic since pre-trained BERT models may not be easily accessible or may not operate at their best.

REFERENCES

Chaffey, D. (2023, June 7). *Global social media statistics research summary 2023 [June 2023]*. Smart Insights. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

Liddy, E. D. (n.d.). *Natural language processing*. SURFACE at Syracuse University. https://surface.syr.edu/istpub/63/

Vayansky, I., & Kumar, S. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582