IDS 566 Project Report

Sentimental Analysis and Topic Modeling of COVID – 19 Vaccine Tweets

by

Group 17

Tarun Sura (UIN - 654229662)

Kavya Vijayandhran (UIN - 658400629)

Sanjay Madesha (UIN - 662505955)

Problem Statement

A new word was trending on Twitter as the year 2020 began. Between February 2020 and January 2021, Twitter users sent over 1 million tweets per day about Covid – 19 (SARS-CoV-2). Not only citizens but also government officials, have used Twitter to share COVID-19 policies and news on a regular basis.

The goal of this study is to identify topics and sentiments in public COVID-19 vaccine-related Twitter discussions, as well as to discern significant changes in topics and sentiments over time, in order to better understand public perceptions, concerns, and emotions that may influence herd immunity goals.

Methods Used

Data Extraction and Cleaning:

The primary data source for COVID – 19 Vaccine Related Tweets was a Kaggle repository containing tweets scraped from Twitter. There were 11020 tweets about the covid-19 vaccine in the dataset.

The following data cleaning steps were performed before processing the tweets to remove noise and unnecessary data:

- Removed twitter handles, hashtags, URLs, special characters, single characters, and stop-words
- Replaced multiple spaces with a single space.

The following methods and metrics were used:

- Tokenization
- Vectorization
- Lemmatization

Fig 1: Project Flow

Sentiment Analysis:

We used VADER (Valence Aware Dictionary and Sentiment Reasoner), a pre-trained sentiment analyzer, to perform Sentiment Analysis. VADER has been trained to detect sentiments in social media languages, such as short sentences containing slang and abbreviations. As a result, it's an excellent choice for detecting sentiments in tweets.

Each text's polarity (positive/negative) and intensity (strength) of emotion are detected by VADER. Given a text, it generates three scores that can be used to determine the text's sentiment: negative, positive, and neutral, all of which add up to 1. Using the scores, we found the sentiment for each tweet. A word cloud was used to depict the words from the most negative and positive sentiment tweets.

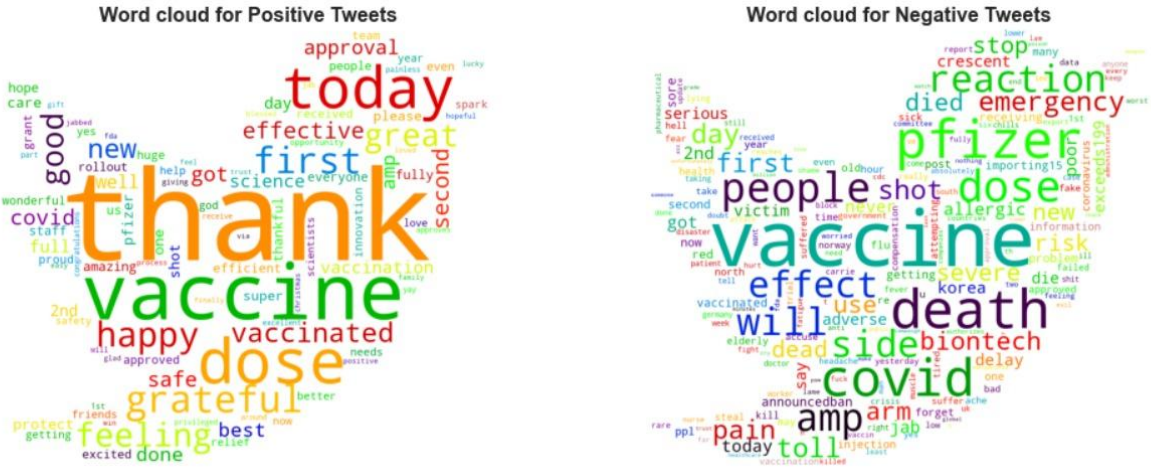

Fig 2: Word Cloud for Positive and Negative Tweets

The Kernel Density estimate graph was plotted to determine the sentiment for the entire dataset as well as for the top two regions - India and Malaysia. From the KDE plot of the entire dataset, we can say that positive and negative sentiments are nearly identical and show ambiguity after a certain point, while neutral sentiment has an asymmetric distribution.

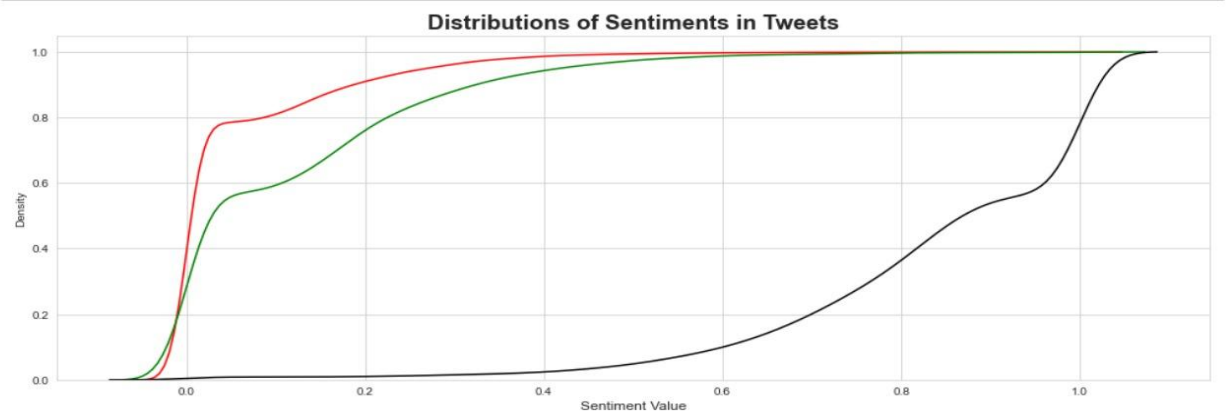

Fig 3: KDE Plot

By knowing the sentiment in a particular region, we can identify the regions where there exists a majority of negative sentiment towards COVID - 19 vaccines and increase efforts to spread awareness about the vaccines.

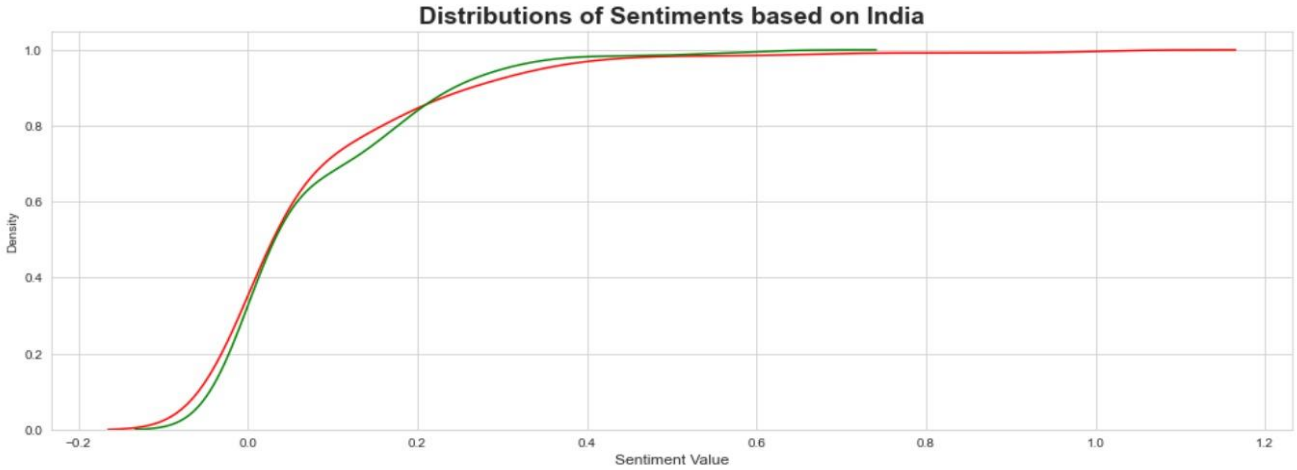

Fig 4: KDE Plot for India

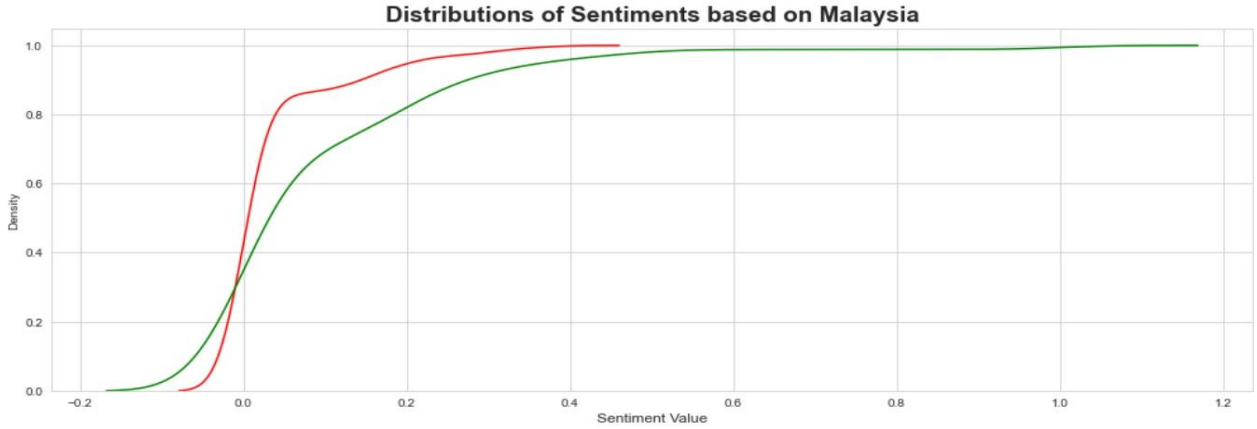

Fig 5: KDE Plot for Malaysia

We can conclude from the above distribution that sentiments in India are less positive and more negative towards the Covid -19 vaccines, whereas Malaysia has a more positive and less negative sentiment towards the Covid -19 vaccines.

Topic Modeling:

We used the Latent Dirichlet allocation (LDA) technique for modeling the topics obtained from the tweets. LDA is a supervised model that generates topics based on word frequency from a set of documents. It uses a generative probabilistic model and Dirichlet distributions to achieve this.

To begin with, we tokenized the text, removing stop words, and numbers but avoided removing words containing numbers (eg: covid-19, 2nd). Furthermore, we removed one-word characters before applying lemmatization. Then we made a Bigram and a trigram, along with a dictionary and corpus for Topic modeling.

We started by setting k to 15 and fitted an LDA model using a gensim package with chunk sizes of 500 and 400 iterations. Then, using c_v as a measure, we computed coherence scoring to determine the coherence score. The best k value is then determined by plotting a line graph. We took the best k value from the graph and ran the LDA model. We extracted the top 5 words from each topic using the new LDA model, interpreted the results based on the Topics, and plotted a word cloud to show the top 3 words from each topic.

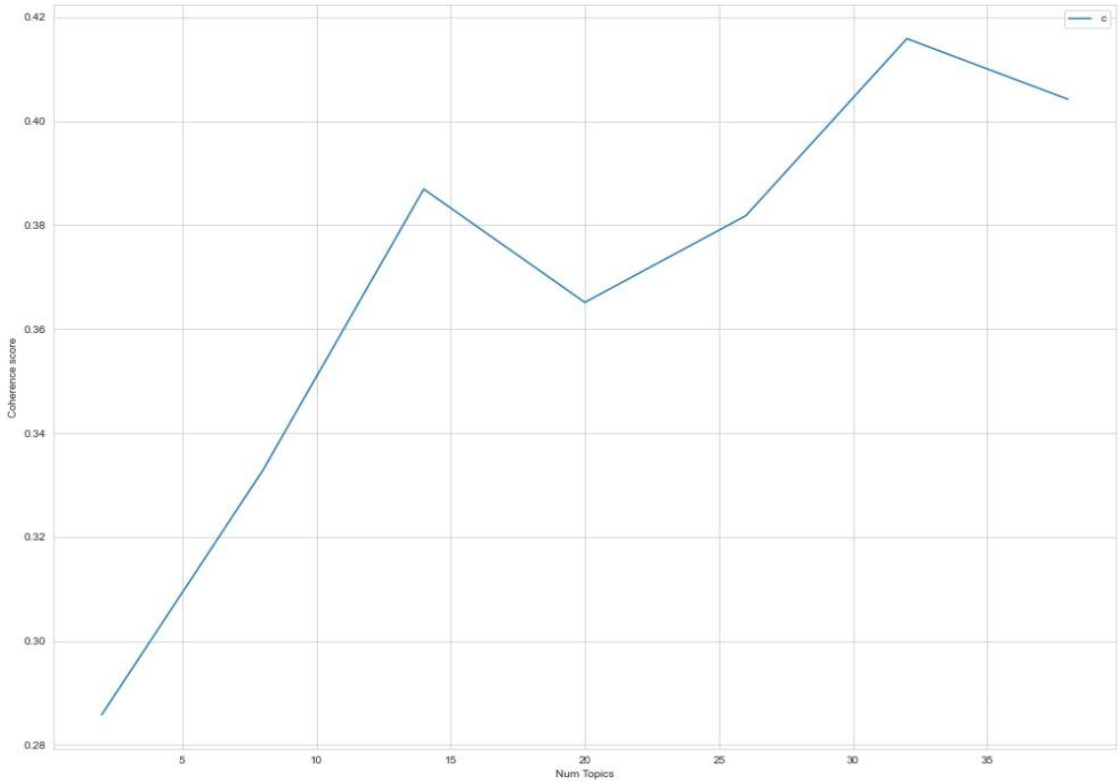

Fig 6: Line graph for best ‘k’ value

```
Topic 1:  grateful feeling incredibly happy opportunity
Topic 2:  jab cases health following hours
Topic 3:  38 day belgium zavemtem vaccinem
Topic 4:  new approval variant fda allergic
Topic 5:  use approved emergency coronavirus know
Topic 6:  clinic walk effect help patients
Topic 7:  given tested positive really time
Topic 8:  headache make support national announced
Topic 9:  ceo weeks says committee stop
Topic 10:  data approves european astrazeneca medicines
Topic 11:  way making blame wondering great
Topic 12:  giving record fab va market
Topic 13:  thanks days country hard immune
Topic 14:  fully million protect workers healthcare
Topic 15:  jab getting effective effects 000
Topic 16:  lame carrie ivaccinated led lam
Topic 17:  science approved german calling message
Topic 18:  2021 number cases safety remember
Topic 19:  time finally team study great
Topic 20:  10 lives disease control cdc
Topic 21:  correct john alongside alison general
Topic 22:  president effects post questions health
Topic 23:  world real study effective booster
Topic 24:  push online 5g eligible vaccinat
Topic 25:  deaths 23 norway elderly shots
Topic 26:  receive 100 storage including process
Topic 27:  purchase uk million drug plans
Topic 28:  best taking man supply ugur
Topic 29:  cov sars ve ofvaccine event
Topic 30:  yesterday having hope looking forward
Topic 31:  year good old news happy
Topic 32:  arm hours sore site injection
```

Fig 7: List of 32 topics obtained from the tweets

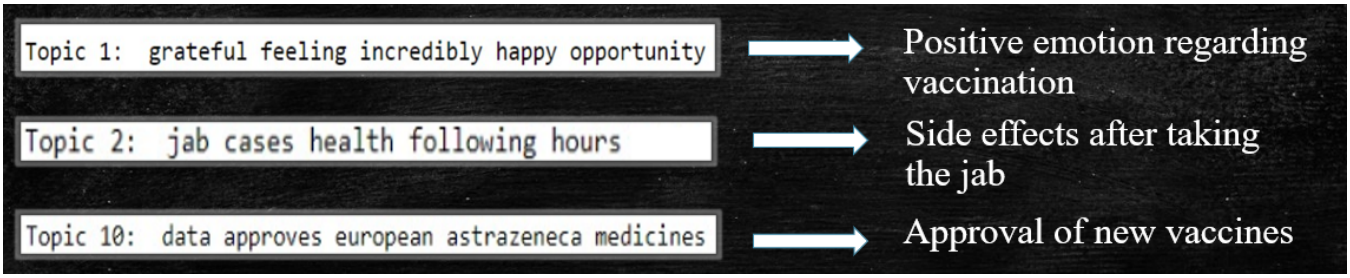

Fig 8: Interpreting the topics obtained from topic modeling

Results:

1) We were able to find the sentiments of the top 2 locations using the contents and location of the tweets.
    a) India - Most negative sentiments related to Covid -19 vaccines and hence requires more efforts to spread awareness towards Covid -19 vaccines
    b) Malaysia - Most positive sentiments related to Covid -19 vaccines.
2) Using LDA, we were able to carry out topic modeling, which along with coherence measure led to 32 optimum topics with a coherence score of 0.55
3) Word clouds were created for the most positive and negative words in the tweets, as well as for the top 5 words in each topic.

Contributions:

- Tarun Sura - Visualizations and Topic Modeling
- Kavya Vijayandhran - Data Cleaning
- Sanjay Madesha - Sentiment Analysis