Analysing Public Sentiment During the COVID-19 Pandemic

Insights from Twitter (X) Data

Final Dissertation

Student ID: xxx

Submission Date: xxx

Supervisor: xxx

This report is submitted in partial fulfilment of the requirement for the degree of M.Sc. Data Science by (Student Name).

Table of Contents

Abstract

The Covid-19 outbreak, initially identified as pneumonia and later declared a pandemic by the World Health Organization (WHO) in 2020, has spread rapidly, leading to 74.3 million cases worldwide and 1.65 million deaths within the span of a year. The pandemic also had a significant impact on the economies of various countries around the world. According to the International Monetary Fund (IMF), its estimates show that the global economy shrunk by 4.4% in 2020, marking the worst decline since the Great Depression of the 1930s (BBC).

As a result of these challenges various conversations started on social media. In this study we aim to analyse these conversations on twitter, focusing on tweets shared by individuals during the COVID-19 outbreak using sentiment analysis with VADER and BERT from Hugging Face.The study focused on using hashtags such as #COVID19, #COVID, and #Vaccine, for mining a total of 35,000 tweets from the platform between July 27 and July 30, 2020.

We used VADER to classify COVID-19 and vaccine related tweets into three sentiment categories; Positive, Negative, and Neutral, based on their compound scores. Word Cloud and charts were used to visualise the consistency of various text data, while Unigram, Bigrams, Trigrams helped analyse tweet content. Unigrams gave an overview of the most often stated words, bigrams showed usual word pairs, and trigrams emphasised frequently used three-word phrases.

This study reveals the impact of the pandemic on public sentiment and the reactions towards the virus and vaccination efforts.

1. Introduction

COVID-19, a contagious disease caused by the coronavirus SARS-CoV-2 had a huge impact on the world, disrupting businesses and people's everyday life. In the education sector alone, about 1.6 billion learners around the world had to adjust the method of learning (Weforum, 2020). In addition to this the expected growth of the employment-to-population ratio by 0.45 percent was overshadowed by a decline of 2.05 percent, resulting in a negative gap of 2.50 percentage points in advanced countries. The impact was also severe in emerging and developing regions, with Latin America and the Caribbean experiencing an even greater decline of 6.52 percentage points (St. Louis Fed, 2023).

More than the direct health impact of the disease itself, another issue government and health-focuses corporations had to deal with was the perception of people about the pandemic. Uncertainty and fear were some of the consequences of the pandemic, affecting individuals' mental health around the world (Wells, 2006). This psychological effect of the disease made it important to control the narrative and get only verified truth and opinions to spread. Social media platforms were one of the major means of communication and expression during the pandemic playing a crucial role in enabling government organisations to disseminate information with Twitter playing a key role.

Twitter's (now X) real-time nature, handling up to 500 million tweets daily, provided a valuable resource for extracting large volumes of user-generated content. From January to May 2020 alone, over 120 million COVID-19-related tweets were posted on the platform. This amount of data about a particular topic, in this case the Covid-19 pandemic, offered unfiltered access to understand what the average person thinks of the situation. With the ability to mine data (tweets) off the platform, the possibility of running a sentiment analysis leveraging natural language processing (NLP) became pertinent. NLP is a data science method of assessing and interpreting the emotional tone of content and to classify sentiments in opinions by using various machine learning and deep learning algorithms, making it possible to classify an author's attitude based on whether it is positive, neutral, or negative (Liu, 2012).

In this research, the aim is to analyse tweets and interactions from individuals, organisations, and government agencies on Twitter during the COVID-19 outbreak; It will specifically examine sentiments related to the vaccine and how those opinions shaped the latter days of the pandemic. Two major research questions will be answered;

(1) What were the major concerns regarding the pandemic as well as the issuance of the vaccine?
(2) What are the most prevalent themes and discourses regarding COVID-19?

Understanding these dynamics is essential for assessing the impact of the outbreak and informing strategies to address future public health crises. To begin, past research related to the study will be examined to understand what scholars and researchers have already established and build on that to further the application of data science to the problem. The study will clean the data and perform an exploratory data analysis as part of the methodology. It will then introduce the cleaned data to various machine learning models including Bert, Vader, Support Vector Machine, Logistics Regression, one-hot encoding and Long short-term memory (LSTM). The results of various models will be explored and discussed then a conclusion put-forward from an overarching perspective.

2. Literature Review

Sentiment analysis across various social media platforms is an efficient way to track trends and get an extensive grasp of individual feedback and sentiment. For example, In an article by Vasu Jain (2013) he predicted the success of a movie based on twitter user sentiments by categorising the labels into four distinct categories positive, negative, neutral and irrelevant. The study's findings suggested that movies with a higher level of positive sentiment were more likely to be successful at the box office, but those with positive sentiment performed poorly. The indifferent and irrelevant sentiments had little bearing on the film's popularity. Covid-19 had such a huge impact on the world that scholars from all over are working to analyse sentiments from different perspectives.

In their papers Mansoor et al. (2020) collected 165,116 tweets on the keywords “coronavirus”, the tweets were collected between Jan 1st, 2020 and June 29th, 2020 from around the world focused on the most impacted countries, including the USA, India, and Brazil. The authors used sentiment classification models including LSTM and ANN, to analyse binary sentiment as positive and negative including emotional responses, such as fear and trust.

The accuracy of the sentiment classification models were assessed using a dataset of #coronavirus related tweets labelled with VADER.The results showed that LSTM models had an accuracy of 84.5%, outperforming ANN models, which had 76%. Their findings utilising time-series graphs revealed an

increase in cases after March 2020, with a significant disparity between negative and positive sentiment in the early months and persistently greater levels of fear compared to trust throughout the epidemic. Regional differences were also seen, with nations such as Bangladesh and South Africa exhibiting stronger positive emotion and the United States and Brazil displaying higher negative sentiment. Thailand, Vietnam, and Poland had high fear levels, while Oman, Syria, and Kazakhstan had high trust values, indicating a range of global responses to the pandemic.

Focusing their research on measuring the impact of disease outbreak, Rameshwer Singh et al., (2023) explore the application of sentiment analysis and machine learning techniques in the prediction of disease outbreaks. The research underscores the escalating significance of these technologies within the domain of public health, where their integration holds potential for advancing the accuracy of outbreak predictions. The abstract also briefly touches upon the methodologies employed, highlighting how the synergy between sentiment analysis and machine learning can augment predictive capabilities, offering a novel approach to disease surveillance.

The results of their surveys suggest that the public perception in times of pandemic is largely controlled by microblogging internet opinions and therefore carrying out sentiment analysis can yield valuable insights into public sentiment and behaviour, both of which are pivotal for understanding and anticipating disease spread. The implications of these insights are far-reaching, particularly in informing public health strategies that aim to enhance preparedness and response to outbreaks. Governments can also use insights from the analysis to curb the spread of fake news about diseases and consequently identify internet hotspots that can help spread verified medical updates during such a period.

In another study exploring the collaboration between private and public organisations, Aditya Singhal et al., (2022) examined the role of social media platforms in health communication, with a focus on the COVID-19 pandemic, by analysing Twitter activity from various healthcare organisations. The study proposed to understand the nature of the information shared, public engagement, and sentiment forecasting by collecting data from the Twitter accounts of five pharmaceutical companies, ten public health agencies in the U.S. and Canada, and the World Health Organization (WHO) between 2017 and 2021. The analysis employs content analysis using Natural Language Processing (NLP) techniques and sentiment forecasting through machine learning models.

The analysis revealed a shift in focus towards more relevant topics during the COVID-19 pandemic, as evidenced by a change in the coherence score (c_umass) from -3.6530 before COVID-19 to -3.7944

during the pandemic. In terms of user impact among public health institutions, the World Health Organization (WHO) achieved the highest score, followed by the Centers for Disease Control and Prevention (CDC) and the National Institutes of Health (NIH), with Pfizer leading among private organisations. Sentiment forecasting was conducted using ARIMA and SARIMAX models and they both demonstrated strong predictive accuracy, with Mean Absolute Error (MAE) ranging from 0.027 to 0.084, Mean Squared Error (MSE) from 0.001 to 0.011, and Root Mean Squared Error (RMSE) from 0.031 to 0.105. Additionally, the research found that public engagement was notably higher on topics related to 'COVID-19' compared to others, indicating how swifting public interest can change in precarious times like the during the pandemic.

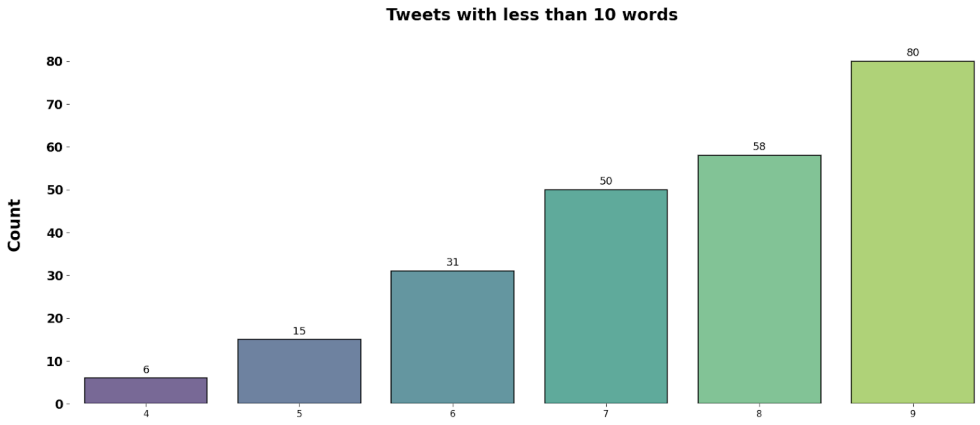
In conclusion, monitoring and managing public sentiments during global health crises, such as a pandemic, is crucial for effective public health response and communication. The COVID-19 pandemic highlighted how rapidly misinformation can spread and influence public behaviour, exacerbating challenges for healthcare systems and policy makers. Had there been more robust sentiment analysis and control mechanisms in place, it could have significantly mitigated the spread of fear and misinformation, improving compliance with health guidelines and vaccination efforts. This research emphasises the importance of integrating sentiment analysis into public health strategies, aiming to enhance preparedness and response in future pandemics by leveraging real-time data to guide communication and policy decisions.

3. Methodology

3.1 Dataset Features

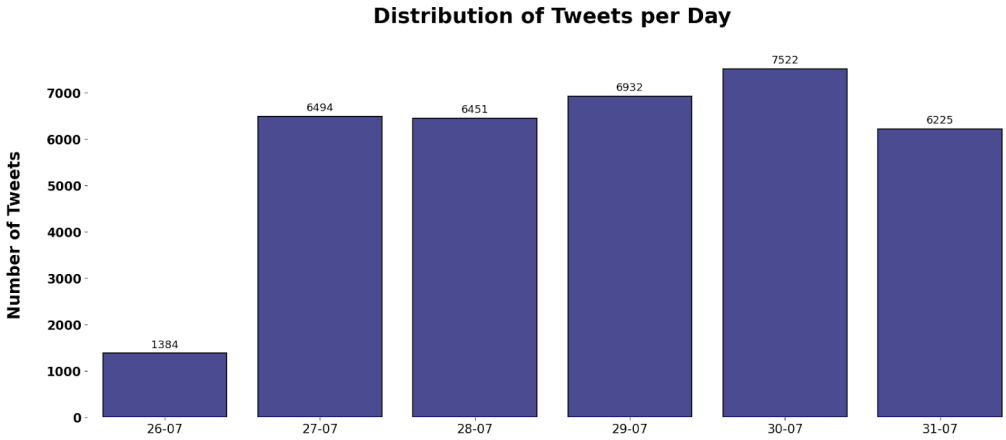
The dataset used for this study was mined from Twitter, following all possible legal and ethical standards and procedures. The dataset includes a total of 35,000 tweets divided into four datasets with varying quantities of words per tweet, with the highest word count being 80 and the lowest being 6.

Figure 1 - Tweets with less than 10 words



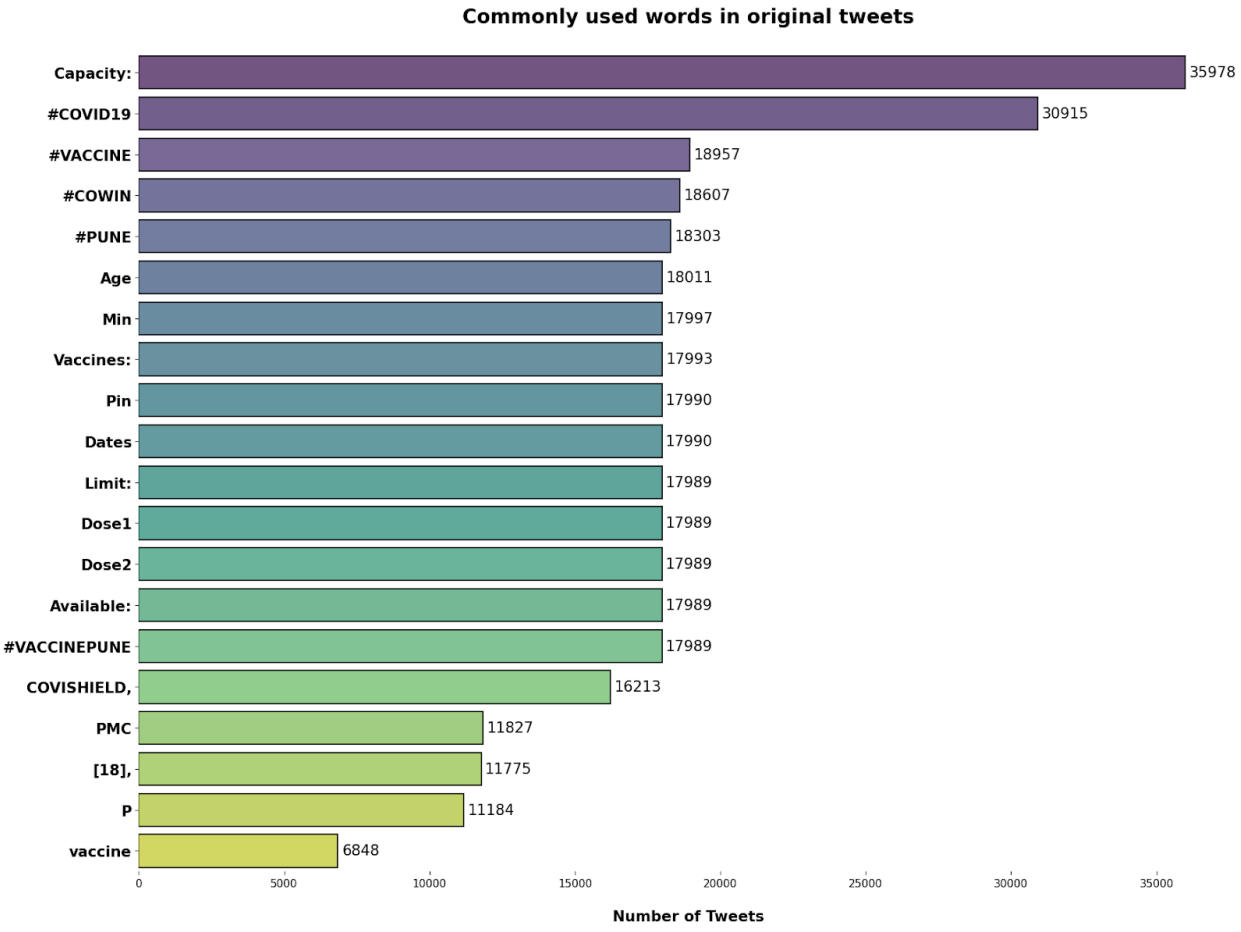
The distribution of tweets by day reveals that the largest number of tweets sent in a single day about this issue was 7,522, while the lowest was 1,384.

Figure 2 - Distribution of tweets per day



The most frequently used words in the dataset were 'capacity,' 'COVID19,' and 'Vaccine,' with each occurring over 18,000 times. Their high frequency emphasises their importance in the analysis.

Figure 3 - Tweets with less than 10 words

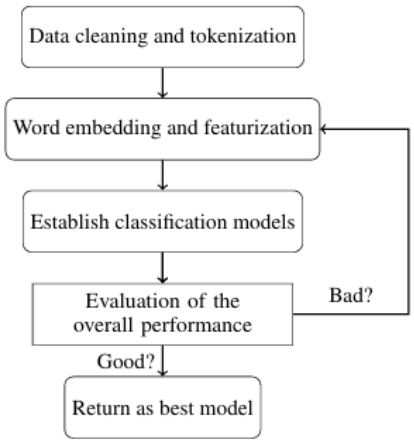


3.2 Text Preprocessing

After gathering the data, the first step was to preprocess and clean it so that it could be properly evaluated. First Duplicates of tweets and usernames were removed to ensure each entry was unique. Twitter identities and URLs were removed to focus entirely on tweet content. To reduce data noise, special letters, punctuation, and symbols like underscores ('_') and quotes

(") were removed, as well as numeric values. The content was then changed to lowercase to ensure uniformity and avoid case sensitivity concerns. Contractions were enlarged to their full forms (for example, "don't" to "do not") in order to further standardise the text. Tweets were tokenized into individual words, and common stop words were deleted to focus on the most important topics.

Figure 4 - Flowchart of the analysis



After preprocessing the data all tweets were primarily divided into tokens for the sentiment analysis. In this study we used Text Blob and VADER to classify tweets. TextBlob is an open-source python library for natural language processing and sentiment analysis. It provides a variety of techniques for dealing with textual data, including sentiment analysis. While VADER is a sentiment analysis tool introduced in 2014 based on words and rules, it is specifically customised for sentiments expressed on social media. It's an open-source tool that considers word order and degree modifiers (Chauhan et al., 2018).

Textblob was used to calculate sentiment polarity, and classify them into categories based on their polarity scores.

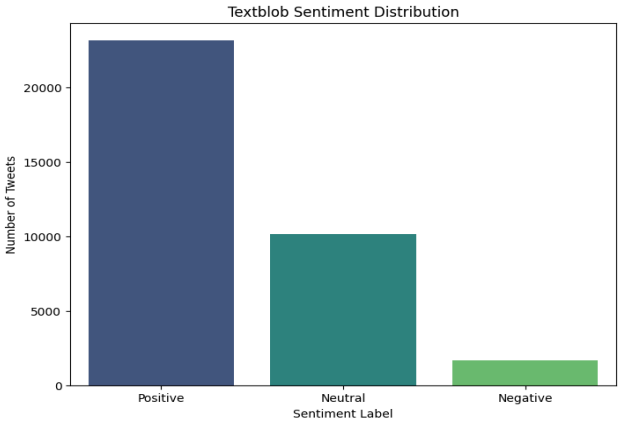
Positive: Polarity score > 1.0

Negative: Compound score < -0.1

Neutral: Scores in-between

Tweets with polarity scores greater than 0.1 have been classified as 'Positive', scores below -0.1 as 'Negative', and scores in the middle range as 'Neutral'.

Figure 5 - Sentiment distribution in textblob



VADER relies on a dictionary which maps out lexical features to emotion and intensities known as sentiment scores. The sentiment scores are gotten by summing up the intensity of each word in a text (Calderon, 2017). We use the following threshold values and classes in this study, to classify tweets as positive, negative and neutral:

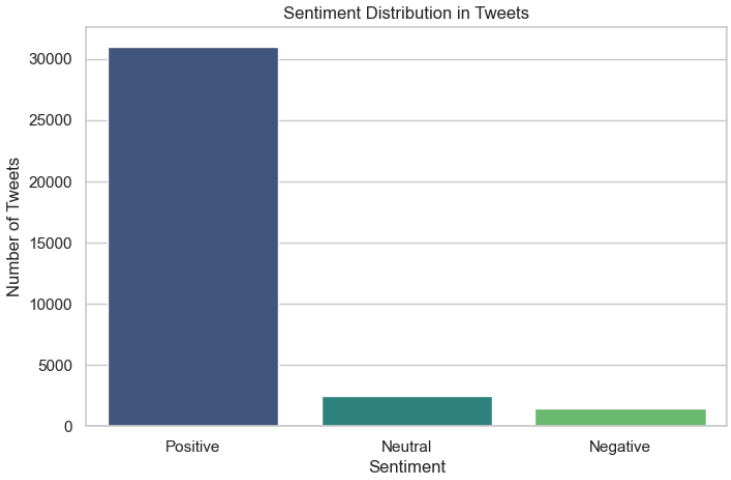
Positive sentiment: (compound score ≥ 0.05), assign score = 1
Negative sentiment: (compound score ≤ -0.05), assign score = -1
Neutral sentiment: (compound score > -0.05 and < 0.05), assign score = 0

We used word clouds to visualise the frequency of text data and N-grams were applied to further understand the content of the tweets.

4. Result and Discussion

BERT, VADER, and TextBlob were used to categorise the data, and they all produced the same result, demonstrating that positive sentiments outweighed negative or neutral sentiments. The result using Textblob shown in Figure 5, suggested more neutral sentiment compared to the result obtained via BERT shown in Figure 6.

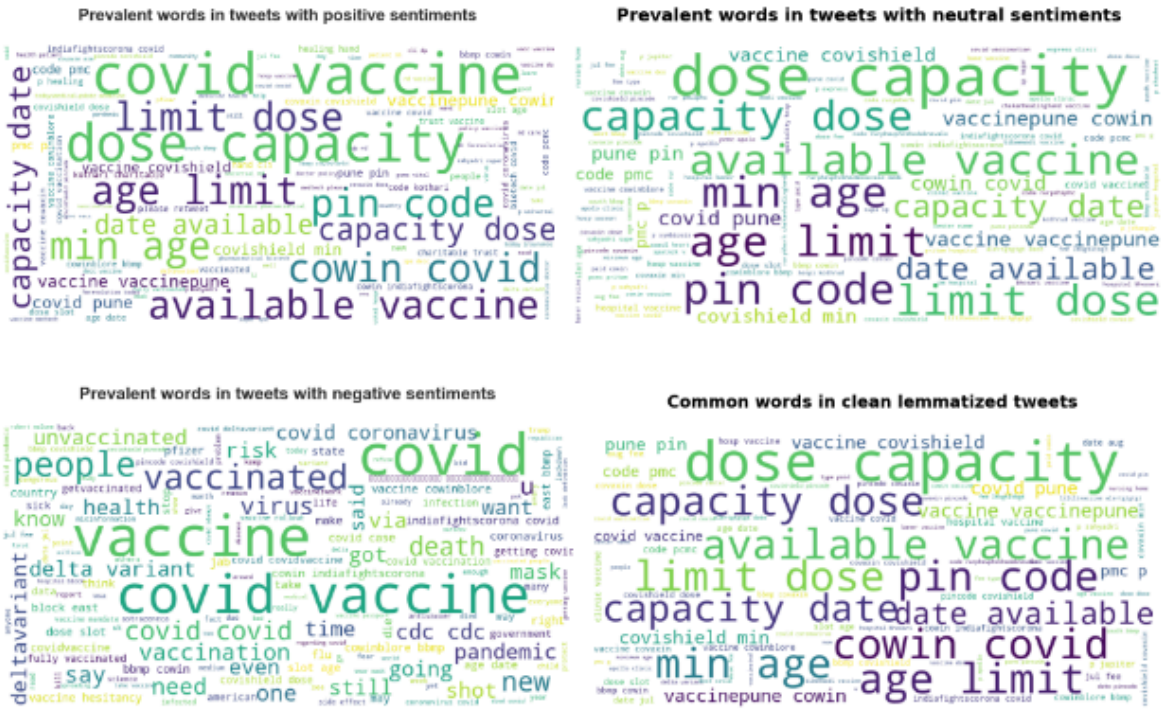
Figure 6 - Bert Sentiment Distribution in Tweets



Word cloud was used to display the frequency of various words and sentiments in the dataset. Word cloud also known as Tag cloud is a data visualisation technique used to display the frequency or importance of each word in a dataset. In the dataset we were able to display the prevalent words for each sentiment classification. For the positive sentiments the most prevalent words includes *Covid, dose, Vaccine, Capacity* as seen in Figure 6.

While the most prevalent words in the negative sentiments include *death, delta variant, vaccinated,* As seen in the negative sentiment word cloud which indicates a significant focus on concerns about the pandemic's intensity, the advent of new varieties, and immunisation difficulties.

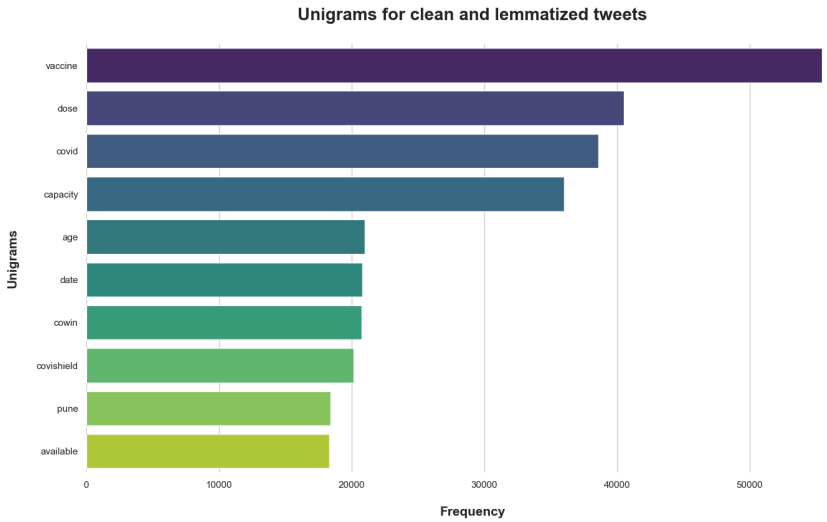
Figure 7 - Word Cloud showing sentiments of tweets across various processing stages



Unigrams is a single word that represents an entire concept, biagrams are two words that contain a distinct meaning when put together while trigrams are three words that contain a distinct meaning when used together.(Dr. W.J.B. Mattingly, 2021) In this study unigrams, bigrams, and trigrams were used to analyse and capture the various aspects of sentiment in text data. By examining these different levels of word combinations, the study aimed to uncover deeper insights into the patterns and themes within the text, thereby improving the understanding of sentiment trends and causes.

The most common unigrams for the clean and lemmatized tweets include *vaccine, dose, covid, capacity, and age*.

Figure 8 - Unigrams for clean and lemmatised tweets



Similarly from the Bigrams the words *dose capacity, available vaccine, age limit, pin code* were some of the words that attracted more attention during the start of the vaccine issuance. While the top Trigrams included *min age limit, age limit dose, limit dose capacity.*

Figure 9 - Biagrams for clean and lemmatised tweets

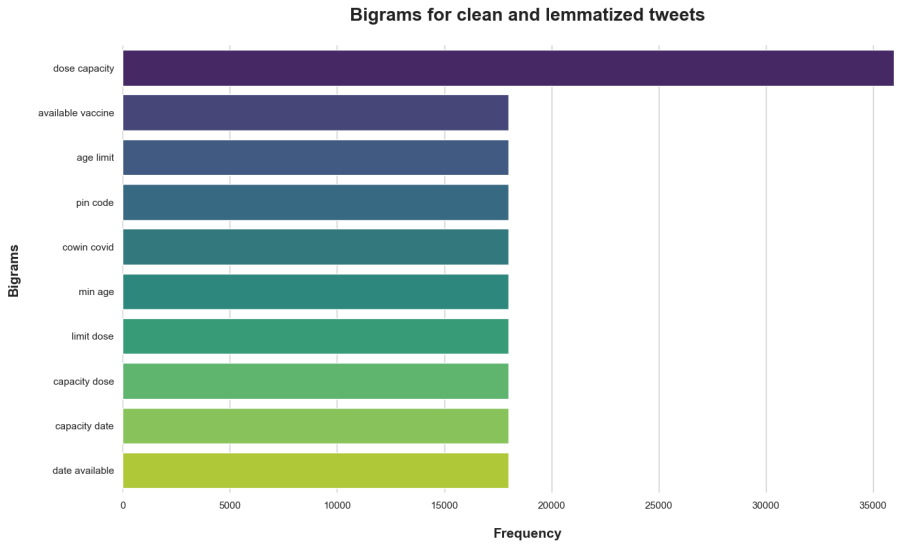
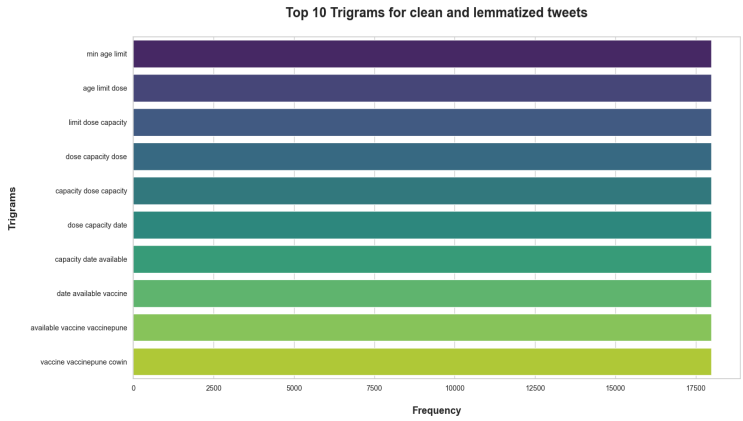
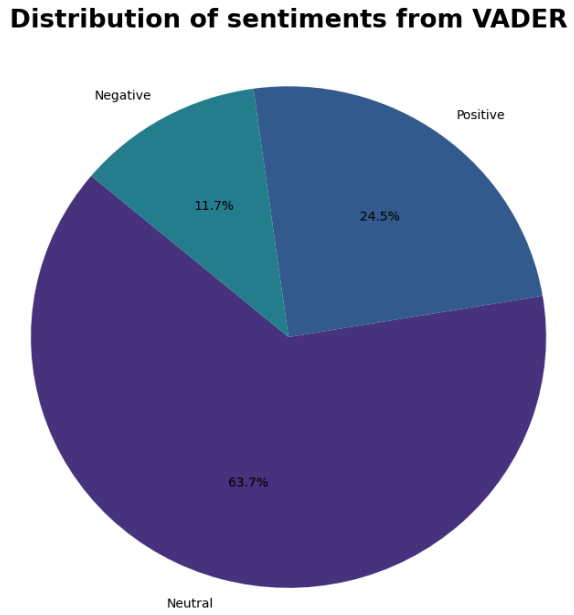


Figure 10 - Top 10 Triagrams for clean and lemmatised tweets



VADER was used to classify sentiments into three categories: positive, negative, and neutral. The findings indicate that the bulk of tweets expressed neutral feelings, making up 63.7% of all content. A sizable proportion of 24.5% expressed good attitudes, whereas only 11.7% expressed negative feelings. This distribution shows that the most common sentiment in the tweets was neutral, with positive sentiments being large and negative ones being very little.

Figure 11 - Distribution of sentiments from Vader



As a result of these findings…

References

Coronavirus: How the pandemic has changed the world economy. (2020). BBC. https://www.bbc.com/news/business-52317545

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. https://www.morganclaypool.com/doi/abs/10.2200/S00416ED1V01Y201204HLT016

Liu, J., & Wu, Z. (2020). Comparative study of Twitter sentiment on COVID-19 tweets. IEEE Conference Publication. https://ieeexplore.ieee.org/document/9418320

Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modelling study. (2020). JMIR Public Health and Surveillance. https://publichealth.jmir.org/2020/3/e19945/

Rodrigues de Andrade, F. M., Barreto, T. B., Herrera-Feligreras, A., Ugolini, A., & Lu, Y.-T. (2020). Twitter in Brazil: Discourses on China in times of coronavirus. Social Science Research Network. https://ssrn.com/abstract=3619019

Singh, M., Jakhar, A. K., & Pandey, S. (2020). Sentiment analysis on the impact of coronavirus in social life using the BERT model. Social Network Analysis and Mining. Springer. https://link.springer.com/article/10.1007/s13278-020-00685-3

Singh, R. & Singh, R. 2023, "Applications of sentiment analysis and machine learning techniques in disease outbreak prediction – A review", Materials Today: Proceedings, vol. 81, pp. 1006–1011.

Singhal, A., Baxi, M.K. and Mago, V., 2022. Synergy between public and private health care organisations during COVID-19 on twitter: Sentiment and engagement analysis using forecasting models. JMIR Medical Informatics, 10(8), p.e37829.

The economic impact of COVID-19 around the world. (2023). Federal Reserve Bank of St. Louis. https://files.stlouisfed.org/files/htdocs/publications/review/2023/04/10/the-economic-impact-of-covid-19-around-the-world.pdf

Tien, Y.-J., & Zhao, Y. (2020). Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak. Humanities and Social Sciences Communications. Nature. https://www.nature.com/articles/s41599-020-00633-1

Wells, A. (2006). The metacognitive model of worry and generalised anxiety disorder. In Worry and its psychological disorders: Theory, assessment and treatment (pp. 179–199). Wiley Publishing.