# Covid-19 booster dose hesitancy analysis: Twitter-based sentiment analysis and machine learning on COVID-19 Booster dose vaccination Twitter dataset.

## Abstract

The global COVID-19 pandemic, originating in Wuhan, China in 2019, has led to the development and authorization of various vaccines as a primary means to mitigate its spread. In light of concerns related to emerging variants and potential waning immunity, the adoption of booster shots has gained substantial attention as a measure to reinforce protection against the virus and its variants. However, the proliferation of rumors and controversies surrounding vaccines, including potential side effects, has had a significant impact on individuals' decisions to accept or decline vaccination. Due to the limited availability of comprehensive data, individuals often turn to Twitter, a popular social media platform, to express their emotions and opinions on various topics, including COVID-19. Therefore, we collected publicly available tweets from Twitter to conduct an analysis of public sentiment regarding COVID-19 booster dose hesitancy. In our study, we compiled a total of 208,252 tweets for analysis. According to the findings of the sentiment analysis study, the majority of tweets exhibited either a neutral or positive sentiment. Specifically, 46.3% of the analyzed tweets were classified as having neutral sentiments, while 40.7% were determined to convey positive sentiments. These findings indicate a gradual shift in public sentiment towards optimism concerning COVID-19 booster doses, primarily driven by concerns surrounding new variants and waning immunity. To assess vaccine hesitancy, we employed the TextBlob computation method in conjunction with the TF-IDF vectorization method and the LinearSVC learning algorithm. Our experimental results demonstrate the effectiveness of employing TextBlob, TF-IDF, and LinearSVC in precisely classifying public sentiment into positive, neutral, or negative categories. Remarkably, our model achieved elevated levels of accuracy, precision, recall, and F1 scores, recording values of 0.9786, 0.9712, 0.9648, and 0.9679, respectively.

**Keywords**: Twitter, Sentiment analysis, Machine Learning, Booster dose hesitancy, COVID-19

## 1. Introduction

According to multiple studies and reports, the emergence of Coronavirus Disease 2019 (COVID-19) was initially detected and observed in Wuhan, China in 2019. This novel disease rapidly spread across numerous countries, leading to the implementation of various preventive measures such as mandatory face mask usage, hand sanitization, and social distancing to mitigate its transmission and impact. The global pandemic resulted in the closure of educational institutions, the prohibition of large gatherings, and the suspension of international travel, significantly impacting over 200 countries worldwide. To combat the spread of the disease and safeguard individuals on a global scale, several COVID-19 vaccines were developed, including Pfizer-BioNTech's Comirnaty, Moderna's Spikevax, Oxford-AstraZeneca's Vaxzevria, Johnson & Johnson's Janssen, Sinovac's CoronaVac, and Sinopharm's vaccine. These vaccines were subsequently administered through widespread vaccination campaigns.

In response to the emergence of a new variant of the SARS-CoV-2 virus, many countries worldwide have begun considering the implementation of additional booster doses of the COVID-19 vaccine. Initially targeted at immunocompromised individuals, these booster shots have since been expanded to the general population, aiming to enhance their immune response. The efficacy of these booster doses in providing protection against the evolving variants has led researchers, healthcare professionals, and policymakers to recognize them as a valuable strategy in the fight against COVID-19. However, despite the positive outlook surrounding booster doses, there has been persistent opposition from anti-vaccine groups, who actively propagate conspiracy theories and falsehoods. These groups often criticize the vaccine development process and raise concerns about potential long-term side effects associated with combating this infectious virus. The dissemination of vaccine misinformation, often amplified by media coverage, has played a significant role in fostering vaccine hesitancy, ultimately impacting vaccination rates. Covid-19 vaccination hesitancy, diplomatic challenges, and unequal access to vaccines, especially in low- and middle-income countries, remain significant factors contributing to obstacles in the global vaccination efforts, including the administration of booster shots. Therefore, researchers and public health experts have been actively compiling datasets to gain a deeper understanding of vaccine hesitancy and its underlying factors.

In the present era of digitalization, the internet has emerged as the primary source of information, and popular social media platforms serve as influential tools for the instantaneous dissemination of information. Since the onset of the pandemic, numerous assertions concerning the COVID-19 vaccine have been circulating, which have the potential to impact public confidence in vaccination efforts. The dissemination of false rumors about vaccines can significantly contribute to vaccine hesitancy, thereby posing a substantial risk to global health. Text mining, a valuable technique for extracting insights from vast data repositories, can be utilized to analyze public sentiment towards COVID-19 vaccines using social media databases. Social media platforms provide individuals with a platform to express their opinions, making them a valuable resource for data collection. Thus, the collection and ongoing surveillance of data from platforms such as Twitter can offer a wide array of perspectives on vaccines. This invaluable information can assist public health organizations in discerning the factors that influence vaccine confidence by leveraging historical and geographical data. Embracing a comprehensive approach becomes imperative in bolstering global vaccination coverage and effectively tackling these challenges.

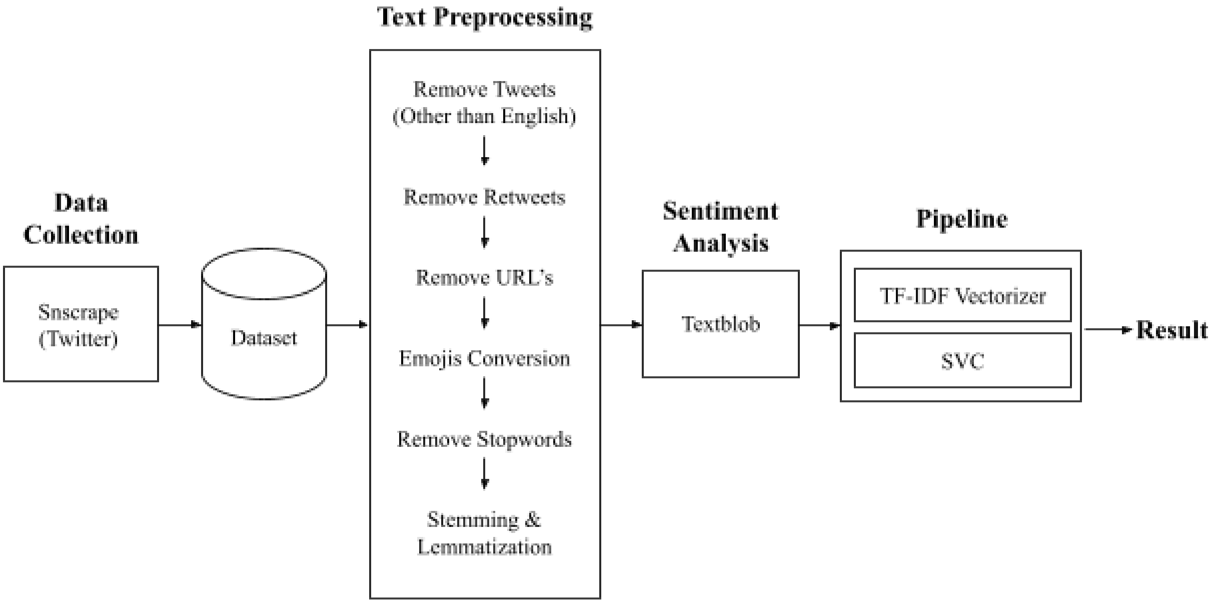

*Figure 1. Experimental Design*

The organization of this study is delineated as follows: Section 2 comprises a review of the existing literature on COVID-19 hesitancy. Subsequently, Section 3 expounds on the methodology adopted for this investigation, and Section 4 delves into the presentation of the acquired results. Sections 5, 6, and 7 are dedicated to discussions, limitations, and conclusions, respectively. Furthermore, Section 6 offers an overview of prospective future work, while our acknowledgments are conveyed in Section 8.

## 2. Literature Review

Subsequent to the outbreak of the COVID-19 pandemic, numerous studies have been conducted on COVID-19 vaccination hesitancy through sentiment analysis. These studies aimed to analyze public sentiments regarding COVID-19 using Twitter data, employing various sentiment computation techniques. Among the most widely used sentiment analysis methods, TextBlob is commonly employed in such studies followed by VADER. Praveen SV, Jose Manuel Lorenz, and Rajesh Ittamalla (2022) examined the sentiments of Indians on the COVID-19 booster vaccine utilizing the TextBlob sentiment computation technique. Miftahul Qorib, Timothy Oladunni, and Max Denis (2023) found that the performance of TextBlob is consistent and the top highest model performances were achieved using the TextBlob sentiment computation while studying the daily tweets to analyze COVID-19 vaccine resistance in a combination of different sentiment computation methods, vectorization techniques, and classification algorithms. Areeba Umair & Elio Masciari (2022) used TextBlob  to analyze the people's sentiment for the vaccine campaign. S.V. Praveen, R. Sundar, and Vajratiya Vajrobol (2023) studied the sentiments of individuals with comorbidities towards COVID-19 booster vaccine shots utilizing the TextBlob sentiment computation technique. While Han Xu, Ruixin Liu, and Ziling Luo (2022) employed VADER to analyze the COVID-19 vaccine topic from Twitter. COVID-19 vaccine topic from Twitter. However, in some research studies sentiment analysis with model classification was performed. Amerah Alabrah, Husam M. Alawadh, and Ofonime Dominic Okon (2022) analyzed the response of citizens of gulf countries on COVID–19 vaccination utilizing Ratio, TextBlob, and VADER sentiment computation method along with LSTM and machine learning classifiers. They highlighted the performance of VADER as having the best classification results using KNN and Ensamble boost.  Anushtha Vishwakarma & Mitali Chugh (2023) used Vader-lexicon for calculating the sentiments, performed EDA, and analyzed the results using machine learning and deep learning algorithms. Papri Ghosh, Ritam Dutta, and Nikita Agarwal (2023) performed sentiment analysis along with machine learning algorithms to analyze the user tweet responses on the third booster dosage for COVID-19 vaccination. Noralhuda N. Alabid & Zainab Dalaf Katheeth (2021) implemented ML classifiers to classify the collected tweets into positive, negative, and neutral.

Previous investigations regarding COVID-19 vaccine hesitancy have primarily focused on the reluctance of individuals to receive vaccination within specific regions, geographical areas, and communities. However, we contend that there is potential for further exploration by studying the reactions of individuals not constrained by specific regions or geographical areas. In the majority of research studies, the analysis of worldwide perceptions towards widely administered vaccines has been conducted through the utilization of natural language processing (NLP) and supervised machine learning algorithms. Additionally, investigations have explored the variances in vaccine perspectives across different cultures and continents. Moreover, topic modeling techniques have been applied to tweets to identify the key aspects that individuals discuss when engaging in such online conversations. Our aim was to analyze the global context and public response to vaccination as a measure of protection against COVID-19 variants, both prior to the implementation of booster doses and subsequent to their introduction. Additionally, we believe that monitoring the vaccination situation over a specific period of time can provide us with more comprehensive insights into the global COVID-19 situation. Furthermore, it is worth noting that the scope of the performance study was limited, which imposed constraints on the breadth and depth of the analysis. We hold the view that incorporating performance evaluation techniques has the potential to enhance the outcomes of research on COVID-19 vaccine hesitancy.

Natural Language Processing (NLP) is an area of computer science that focuses on the examination and interpretation of text. Over time, notable progress has been achieved by leveraging deep learning (DL), machine learning (ML), and NLP methodologies, as evidenced by a prominent study. Furthermore, ML and NLP have found applications in the realm of medical reviews to analyze public behavior. Moreover, such an approach will offer valuable guidance for future studies, enabling them to make additional contributions to the subject matter. It is essential to acknowledge that the landscape of COVID-19 vaccine hesitancy is dynamic and subject to change. Consequently, it is of utmost importance to address the present state of COVID-19 vaccine hesitancy in order to provide timely and up-to-date insights.

By examining and comparing previous studies, we can expand the current research by delving into unexplored areas that have not been addressed previously. Expanding upon prior investigations that have identified effective strategies and methodologies yielding favorable outcomes, we have incorporated the TextBlob sentiment analysis technique alongside TF-IDF vectorization and the LinearSVC classification algorithm in order to delve into public sentiments surrounding the COVID-19 booster dose vaccine. Drawing upon TextBlob's demonstrated efficacy, the efficiency of TF-IDF vectorization, and the promising outcomes achieved by LinearSVC, we have specifically chosen these methodologies for our study. Our principal objective entails evaluating the general public's perception regarding the effectiveness of booster dose vaccinations in combating COVID-19 variants, while also assessing the performance of our model.

## 3. Methodology

For our study, we utilized a social media platform (Twitter) to gain insights into people's perspectives on COVID-19 booster doses and the concerns they expressed regarding booster doses. To collect relevant data, we utilized the Python library Snscrape to scrape public tweets, which enabled us to understand people's experiences, emotions, and perspectives on health policies. The study commenced with a preprocessing stage in which we eliminated extraneous text from the scraped tweets to ensure suitability for sentiment analysis.

In the preprocessing step, the data was normalized by removing stopwords, URLs, and retweets, and translating emoji into words from the collected tweets. Subsequently, we tokenized the tweets and applied stemming and lemmatization techniques. Following the preprocessing, we employed TextBlob to calculate sentiment scores, and the dataset was vectorized using the TF-IDF vectorization technique. Lastly, we utilized the LinearSVC machine learning algorithm to classify the COVID-19 Twitter dataset into positive, neutral, or negative categories, thus evaluating our model's performance.

### 3.2. *Data collection*

We employed the Python library Snscrape to scrape tweets posted between May 2021 and December 2022 that contained the phrase "COVID-19 Booster Vaccine". Snscrape offers users the capability to extract various elements such as user profiles and hashtags. This robust tool enables users to retrieve different types of search results, including live tweets, top tweets, and user information. In our study, we utilized Snscrape, a sophisticated Python Twitter scraping tool, to collect the necessary data, eliminating the need for an Application Programming Interface (API). After removing non-English and duplicate tweets, we obtained a total of 184,224 English tweets and saved them as CSV(Comma-separated value). The exclusion of non-English tweets was essential to align with the nature of our analysis.

### 3.3. *Text preprocessing*

Text preprocessing is an essential stage in natural language processing (NLP) where unprocessed textual data is cleaned and converted into a suitable format for analysis and machine learning algorithms. This process significantly improves the accuracy and efficiency of NLP tasks by eliminating irrelevant information, standardizing text, and reducing complexity. In the scope of our study, we implemented several preprocessing techniques, encompassing the removal of retweets, URLs, and punctuations, conversion of emojis into words, tokenization, elimination of stop words, as well as applying stemming and lemmatization.

#### *3.3.1. Retweets*

When a Twitter user shares another user's tweet, it is called a retweet. Retweeting is a way to amplify and spread interesting or relevant content across the platform, allowing users to share information and opinions with a wider audience. Due to the prevalence of content duplication in retweets, we took measures to eliminate duplicate tweets. This was necessary to prevent distortion of word frequency, ensure the model's accuracy, and reduce the computational resources needed to conduct the experiment.

### 3.3.2. URLs and punctuation

In the subsequent phase, the removal of URLs is undertaken. This involves eliminating URLs from the tweets. The elimination of URLs is essential as they lack inherent meaning and do not impact the sentiment value of the tweets. Retaining URL links could distort the word frequency analysis since each tweet contains a link. To clean up the tweet texts, non-essential symbols and punctuations such as #, ?, /, , and ! are removed using the Python regular expression (re) package.

### *3.3.3. Emoji*

In the next step we proceeded with the task of transforming emojis into words using the "emoji.demojize()" function from the Python library “emoji”. Emojis are commonly employed by individuals to convey their emotions, and converting these emojis into phrases can potentially enhance the sentiment analysis of the tweets. By converting emojis into words, a more comprehensive understanding of the sentiment expressed in the tweets can be achieved.

### *3.3.4. Tokenization*

Tokenization refers to the process of splitting a string into a list of tokens or individual units. In this particular step, we performed tokenization on our text to achieve two objectives: to reduce code length and to effectively parse the text. As part of this process, all words within the tweets were converted to lowercase.Subsequently, every individual word from the tweets was meticulously stored in a list, creating a comprehensive collection of words that facilitated the calculation of polarity for each word.

### *3.3.5. Stop words*

Stopwords are frequently used words in a language that contribute minimal meaning to a sentence. Once the tweets have been converted into tokens, the subsequent task involves eliminating these stopwords. Common examples of stopwords include 'the', 'is', 'at', 'which', 'on', and others. Removing these stopwords proves beneficial in calculating tweet sentiments. Since these words do not significantly contribute to the analysis and may distort the phrase composition within the tweets, their removal is recommended.

### *3.3.6. Stemming and lemmatization*

Stemming serves as a normalization technique employed to reduce an inflected word to its word stem or base form. It involves the process of reducing words to their fundamental or root form. For instance, words like "spreading" or "spreads" would be reduced to "spread." It is important to note that some stemmed words may not be valid within the language. In this experiment, we utilized the 'PorterStemmer' stemming algorithm.

In contrast, lemmatization is a different approach that aims to reduce inflected words to their root form by considering the vocabulary and form of the words. It takes into account the context of the word, determining the appropriate part of speech and its intended meaning. In this particular experiment, we utilized the 'WordNetLemmatizer' method for lemmatization.

### 3.4. *Sentiment Analysis*

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique that aims to discern the emotional tone conveyed in the text. The sentiment analysis process involves the automatic collection and analysis of subjective judgments regarding different aspects of an item. Textual sentiment analysis holds a prominent position within the field of NLP, and when applied appropriately, it has the potential to organize extensive volumes of unstructured text data and quantify textual sentiments proficiently. The sentiment score of text documents is determined by counting the occurrences of positive, neutral, and negative words within each document. The sentiment scores are then calculated on a scale ranging from -1 to 1. In this scale, a score of -1 indicates the utmost negative sentiment, a score of 1 represents the utmost positive sentiment, and a score of 0 signifies a neutral sentiment.

In our research, we employed sentiment analysis to gain insights into people's perceptions of COVID-19 emerging variants and booster dose hesitancy as expressed on the social media platform Twitter. For this purpose, we utilized the TextBlob computation method to compute the sentiments of people described in texts.

#### *3.4.1. TextBlob*

TextBlob is a Python library designed for Natural Language Processing (NLP), offering comprehensive capabilities for analyzing and manipulating textual data. With TextBlob, it becomes possible to extract the polarity and subjectivity of a given sentence. The polarity score ranges between [-1,1], where -1 signifies a negative sentiment and 1 denotes a positive sentiment. By utilizing the Pandas data frame, we generated polarity values for tweets and associated them with the tweets. Positive polarity values were associated with a positive reaction, negative polarity values were associated with a negative reaction, and polarity values close to zero were classified as neutral sentiments.

Based on the findings obtained through TextBlob Sentiment Analysis, a significant increase in monthly positive sentiments from 27.58% to 37.44% was observed following the announcement of booster doses for the general public. This represents a substantial change of 35.75% in positive sentiments. Additionally, there was a notable decrease in neutral sentiment and negative sentiment, indicating a reduction in booster dose hesitancy by 13.86% and 12.49% respectively.

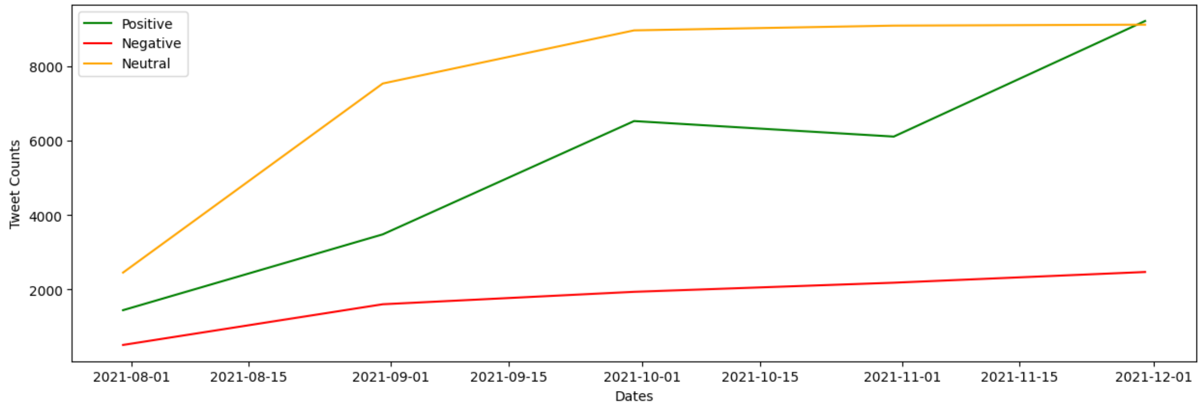

*Figure 2. Monthly Sentiments Breakdown (Booster Dose Announcement)*

We further observed a consistent increase in positive sentiments regarding the COVID-19 booster dose following the emergence of the omicron variant in 2021-22. The monthly positive sentiments reached 40.13% to 46.12%, while the negative sentiments showed a continuous decrement with a massive drop of 14.79% in negative reactions. Furthermore, the neutral sentiments also decreased up to 7.07% during that period.

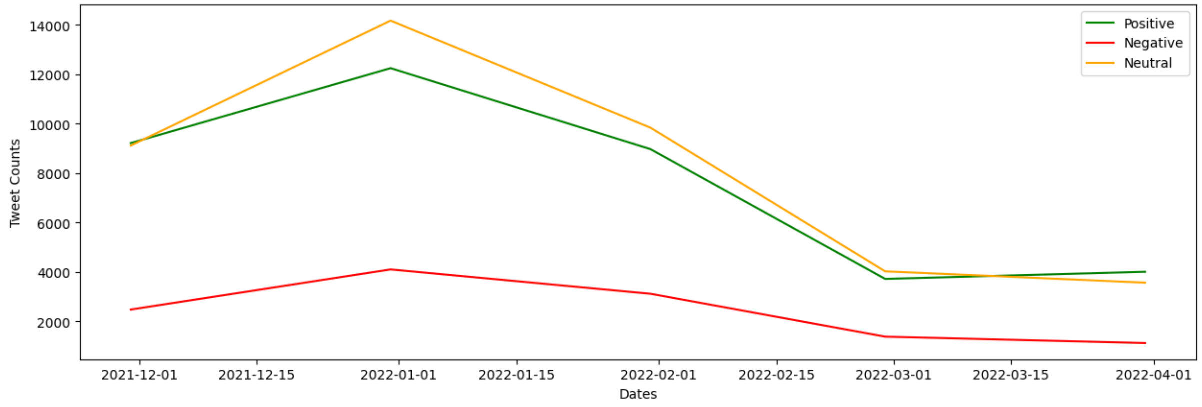

*Figure 3. Monthly Sentiments Breakdown (Booster Dose & Omicron Variant)*

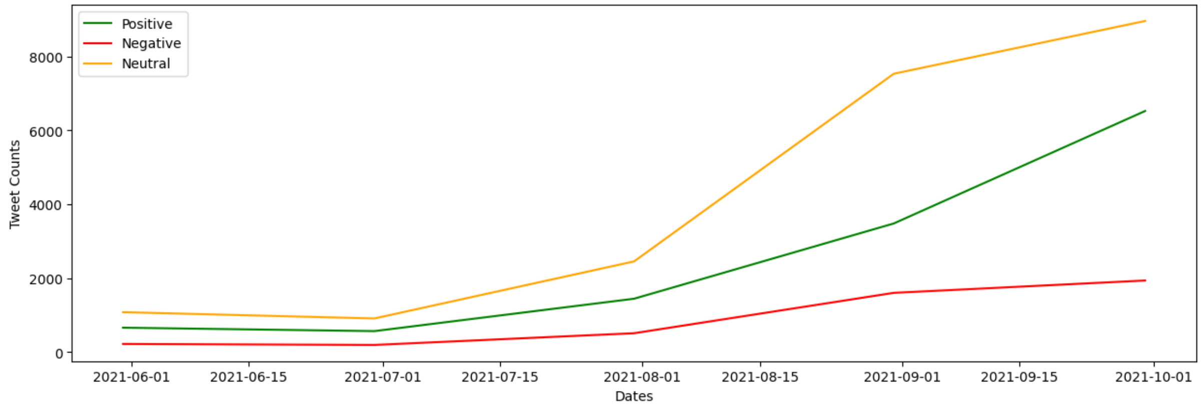

*Figure 4. Monthly Sentiments Breakdown (Second Dose & Delta Variant)*

Moreover, it was observed that prior to the announcement of booster doses, when the delta variant emerged as a prominent COVID-19 strain, people's sentiments regarding COVID-19 vaccination exhibited a slight decrease in positive sentiments (change up to 15.78%) and a rise in negative sentiments (change up to 9.90%). However, the changes in sentiments align with the announcement of booster doses to the general public during the third quarter (in September 2021). It indicates a correlation between the introduction of booster doses and studies claiming the effectiveness of booster doses in enhancing immunity against COVID-19 variants. The data is depicted in *Figure 5*. below provides a positive indication that individuals are increasingly optimistic about receiving booster dose vaccinations implying a massive decrement in vaccine hesitancy over time.

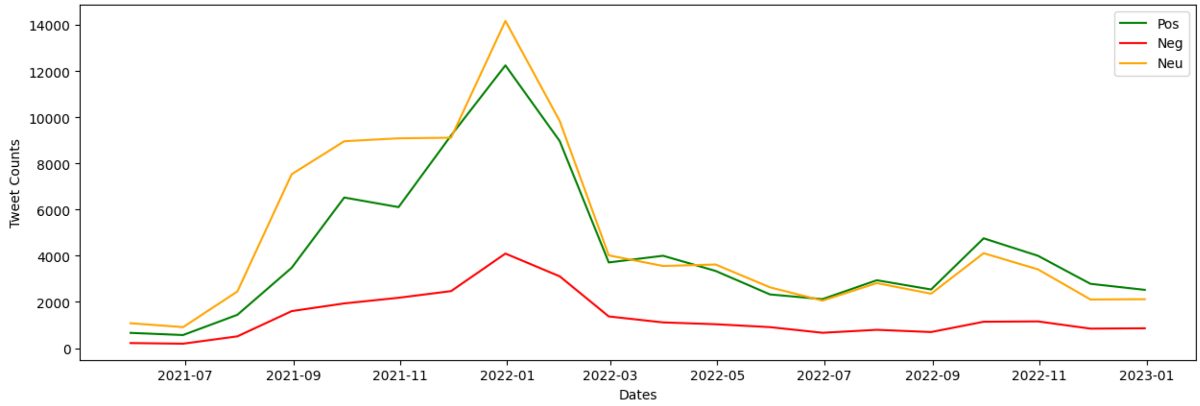

*Figure 5. Overall Reactions on Booster Dose*

### 3.5. *Vectorizations*

During the vectorization stage, the processed text was converted into numerical vectors that solely encapsulate the vital information. This conversion facilitated effective pattern recognition and knowledge discovery, enhancing the overall performance of the algorithms. Text vectorization plays a crucial role in transforming textual data into a numerical format that can be comprehended and processed by machine learning models. In the experiment, the TF-IDF vectorization technique was employed.

#### *3.5.1. TF-IDF*

The TF (Term Frequency) component of TF-IDF measures the frequency of a term within a document. It indicates how often a specific term appears in a document compared to the total number of terms in that document. A higher TF value indicates that a term is more important within the document.

$$TF \;=\; \frac{Frequency\ of\ a\ word\ in\ the\ document}{Total\ number\ of\ words\ in\ the\ document} \qquad (1)$$

The IDF (Inverse Document Frequency) component of TF-IDF measures the rarity of a term in the entire corpus. It calculates the logarithmically scaled inverse fraction of the total number of documents that contain the term. The IDF value is higher for terms that appear in fewer documents across the corpus.

$$IDF = log(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ including\ the\ word}) \qquad (2)$$

By combining TF and IDF, TF-IDF assigns a weight to each term in a document. This weight reflects the term's relevance to the document compared to the entire corpus. Terms that occur frequently within a specific document but rarely in other documents receive higher TF-IDF scores, indicating their significance in distinguishing the document's content.

$$TF\_IDF\ (t, d) = TF(t, d) * IDF(t) \quad (3)$$

*where t is the number of terms that appear in document d.*

### 3.6. *Classification model*

A classification model pertains to a predictive modeling task wherein it aims to determine a class label for a given input data instance. In the context of classification, it is imperative to possess a training dataset wherein each data point is meticulously assigned a distinct label or category. In this study, we used the LinearSVC classification model as the learning algorithm.The data was labeled based on the sentiment analysis results obtained after conducting the sentiment analysis calculation. Subsequently, the dataset was divided into a 70:30 ratio for training and testing purposes. The assessment of model performance was carried out by measuring accuracy, precision, recall, and F1 scores.

Accuracy is calculated by dividing the number of correctly predicted observations by the total number of observations. Precision, on the other hand, is computed by dividing the number of accurately predicted positive observations by the total count of predicted positive observations. Recall, also known as sensitivity, represents the proportion of correctly predicted positive observations within the actual class. The F1 Score is a comprehensive metric that amalgamates Precision and Recall, furnishing a weighted average of both measures.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1\ Score = \frac{2*(Recall * Precision)}{Recall + Precision} \quad (7)$$

#### *3.6.1. Linear SVC*

Linear Support Vector Classification (Linear SVC) is a versatile algorithm employed for support vector classification, capable of effectively handling both dense and sparse input data, thereby offering enhanced flexibility in the choice of penalties and loss functions. It exhibits excellent performance when dealing with larger datasets, as it tends to converge faster with a higher number of samples. It is a valuable tool for various applications, such as text classification, image recognition, and sentiment analysis, offering reliable and accurate predictions. In the study, the default hyperparameter values provided by the sklearn library were utilized for LinearSVC().

## 4. Results

In our analysis, our primary focus has been on assessing public sentiments regarding booster dose vaccinations. To accomplish this, we have employed sentiment analysis to identify and comprehend the sentiments expressed by individuals in their textual content on the Twitter platform. Through the utilization of TextBlob, each word within a tweet is meticulously scrutinized to discern whether the overall sentiment conveyed by the text is positive, negative, or neutral in nature. This methodological approach allows us to gain insight into the prevailing sentiments expressed by individuals pertaining to booster dose vaccinations on Twitter. Our study encompassed a comprehensive analysis of 208,252 publicly available tweets that pertained to the COVID-19 booster dose. Among these tweets, 95,956 (46.3%) were found to have neutral sentiments, 84,250 (40.7%) conveyed positive sentiments, and 26,896 (13.0%) expressed negative sentiments. These quantitative findings provide valuable insights into the prevailing sentiments expressed by Twitter users regarding the COVID-19 booster dose. The overall sentiments of individuals regarding the COVID-19 booster dose vaccine, based on Twitter data, are depicted in *Figure 6.* given below.

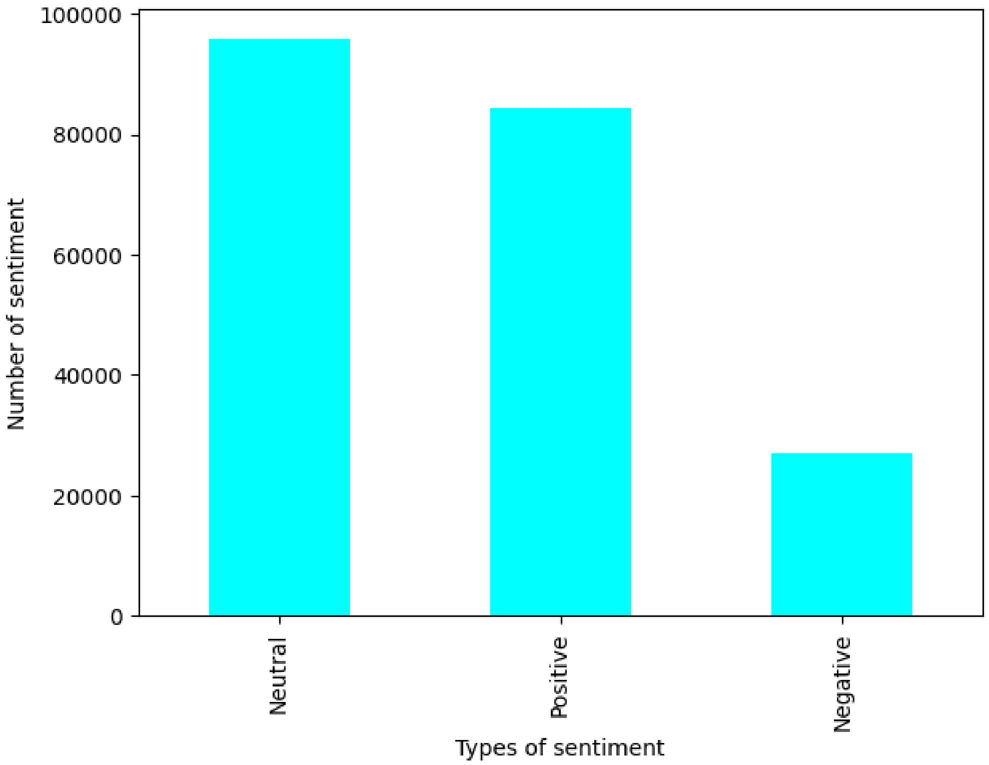

*Figure 6. Overall Sentiments Representation*

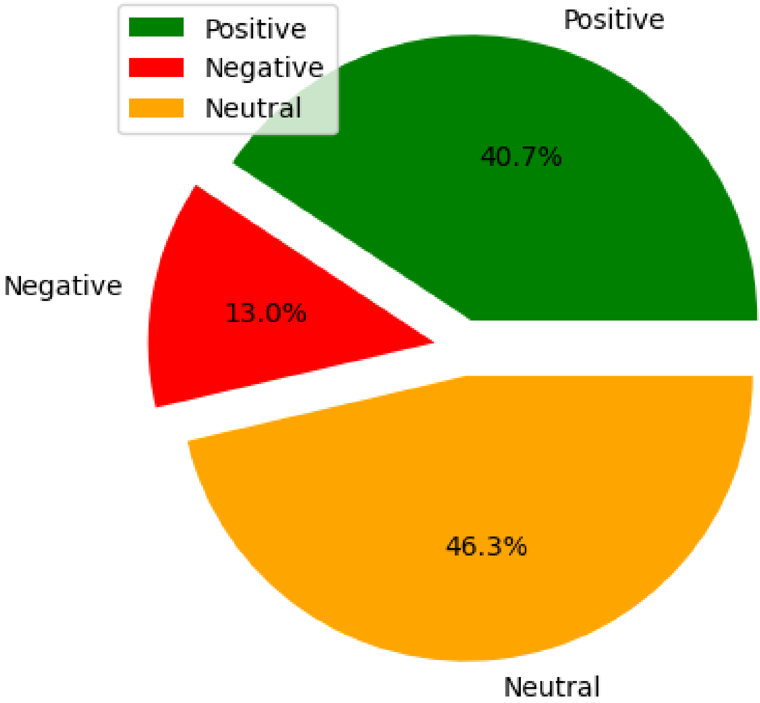

*Figure 7. Pie-Chart representation of Sentiments*

The distribution of sentiments from May 2021 to December 2022 is presented in Table 1

*Table 1. Month-wise Sentiments Distribution (in Percentage)*

| **Date** | **Positive** | **Negative** | **Neutral** | **Total** | **Positive(%)** | **Negative(%)** | **Neutral(%)** |
|---|---|---|---|---|---|---|---|
| 2021-05-31 | 659 | 219 | 1079 | 1957 | 33.67 | 11.19 | 55.13 |
| 2021-06-30 | 566 | 192 | 909 | 1667 | 33.95 | 11.51 | 54.52 |
| 2021-07-31 | 1443 | 509 | 2453 | 4405 | 32.75 | 11.55 | 55.68 |
| 2021-08-31 | 3480 | 1602 | 7533 | 12615 | 27.58 | 12.69 | 59.71 |
| 2021-09-30 | 6524 | 1936 | 8961 | 17421 | 37.44 | 11.11 | 51.43 |
| 2021-10-31 | 6107 | 2181 | 9088 | 17376 | 35.14 | 12.55 | 52.30 |
| 2021-11-30 | 9213 | 2469 | 9114 | 20796 | 44.30 | 11.87 | 43.82 |
| 2021-12-31 | 12248 | 4095 | 14174 | 30517 | 40.13 | 13.41 | 46.44 |
| 2022-01-31 | 8968 | 3110 | 9838 | 21916 | 40.91 | 14.19 | 44.88 |
| 2022-02-28 | 3710 | 1372 | 4018 | 9100 | 40.76 | 15.07 | 44.15 |
| 2022-03-31 | 4000 | 1114 | 3558 | 8672 | 46.12 | 12.84 | 41.02 |
| 2022-04-30 | 3343 | 1035 | 3621 | 7999 | 41.79 | 12.93 | 45.26 |
| 2022-05-31 | 2326 | 909 | 2633 | 5868 | 39.63 | 15.49 | 44.87 |
| 2022-06-30 | 2122 | 663 | 2057 | 4842 | 43.82 | 13.69 | 42.48 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2022-07-31 | 2939 | 792 | 2815 | 6546 | 44.89 | 12.09 | 43.00 |
| 2022-08-31 | 2539 | 695 | 2357 | 5591 | 45.41 | 12.43 | 42.15 |
| 2022-09-30 | 4756 | 1143 | 4111 | 10010 | 47.51 | 11.41 | 41.06 |
| 2022-10-31 | 4000 | 1158 | 3411 | 8569 | 46.67 | 13.51 | 39.80 |
| 2022-11-30 | 2784 | 843 | 2106 | 5733 | 48.56 | 14.70 | 36.73 |
| 2022-12-31 | 2523 | 859 | 2120 | 5502 | 45.85 | 15.61 | 38.53 |

Upon analyzing the aforementioned statistics, we observed a slight decrease in positive responses during the mid-2021 period. This decline in positivity can be attributed to concerns surrounding the effectiveness of COVID-19 vaccines, which can be influenced by factors such as emerging variants, waning immunity over time (such as the Delta variant), and the overall vaccination coverage within the population. During this timeframe, a noticeable decrease in positive sentiments was observed, declining from 33.95% to 27.58%. Conversely, there was an increase in negative sentiments from 11.51% to 12.69%. However, with the introduction of booster doses to the general public and studies supporting the notion that each additional dose enhances vaccine efficacy against severe disease, a significant upturn in positive sentiments was witnessed. Notably, positive sentiments reached a notable peak with a remarkable growth rate of 35.75%.

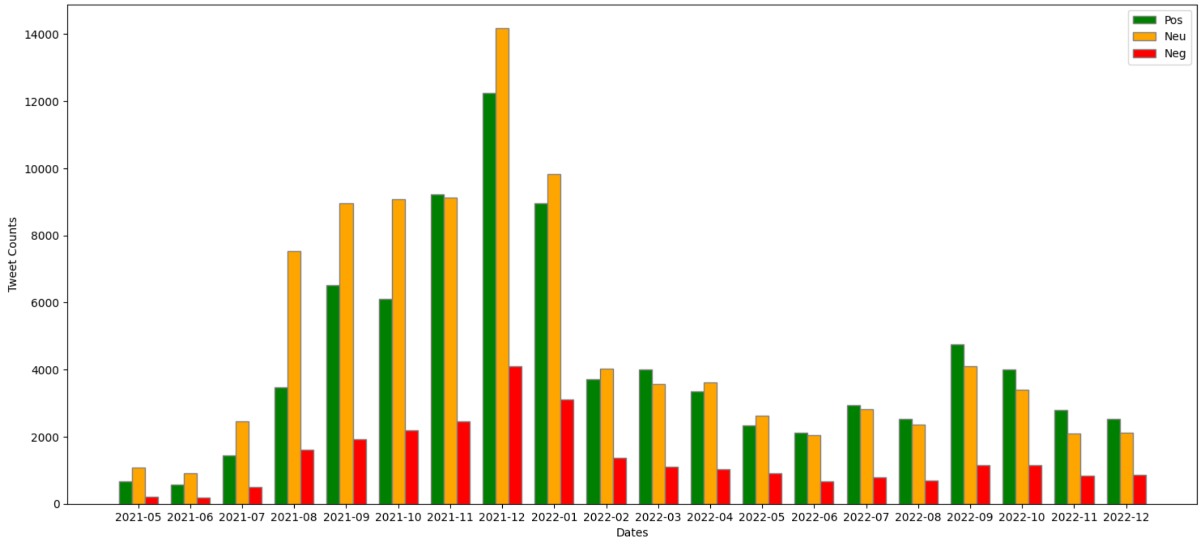

*Figure 8. Monthly Sentiments Spikes*

Our analysis also revealed a notable increase in the monthly volume of tweets, which surged from 4,405 in the third quarter of 2021 (July-September) to 17,421. This trend continued with a further rise to 30,517 tweets in the last quarter of 2021 (November-December). These figures indicate a substantial surge in individuals expressing their emotions and opinions on Twitter pertaining to the COVID-19 vaccination, specifically the booster dose.

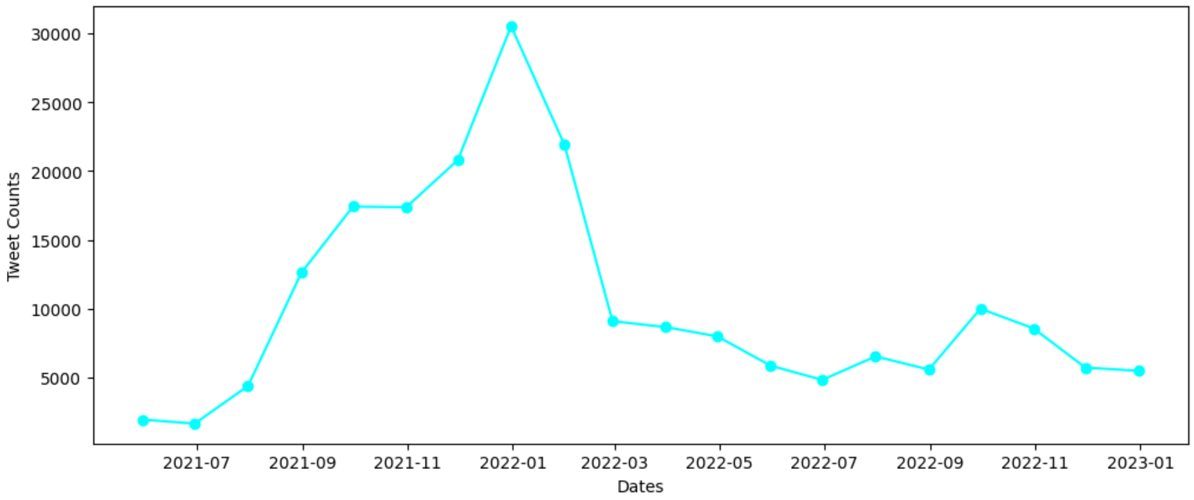

*Figure 9. Overall Tweets Counts*

Towards the end of 2021, there was a noticeable convergence between positive and neutral sentiments concerning the COVID-19 booster dose. However, in the subsequent year of 2022, the overall response towards the booster dose demonstrated a predominantly positive inclination. This shift in sentiment is evident by the discernible impact of the change in the percentage of positive sentiments on the corresponding shift in neutral sentiments. This sustained increase in positive sentiments, coupled with the diminishing disparity between positive and neutral sentiments over the span of several months, indicates a gradual reduction in vaccine hesitancy. These findings suggest that the general public is progressively developing a favorable and optimistic attitude toward COVID-19 vaccination.

*Table 2. Change in sentiments (in Percentage)*

| Date | Positive (% Change) | Negative (% Change) | Neutral (% Change) |
|---|---|---|---|
| 2022-01-31 | 1.95 | 5.75 | -3.35 |
| 2022-02-28 | -0.36 | 6.24 | -1.63 |
| 2022-03-31 | 13.13 | -14.79 | -7.07 |
| 2022-04-30 | -9.39 | 0.72 | 10.33 |
| 2022-05-31 | -5.15 | 19.72 | -0.87 |
| 2022-06-30 | 10.56 | -11.60 | -5.32 |
| 2022-07-31 | 2.44 | -11.63 | 1.22 |
| 2022-08-31 | 1.14 | 2.74 | -1.96 |
| 2022-09-30 | 4.62 | -8.14 | -2.58 |
| 2022-10-31 | -1.75 | 18.34 | -3.07 |
| 2022-11-30 | 4.02 | 8.80 | -7.71 |
| 2022-12-31 | -5.57 | 6.171 | 4.89 |

Additionally, we conducted an analysis of the word cloud, which provided insights into individuals' preferences and concerns regarding various vaccines. Through the frequent usage of specific words, individuals justified their emotions and sentiments towards vaccination. Among the commonly mentioned vaccines were Pfizer, Moderna, and Johnson, indicating their prominence in public discourse surrounding COVID-19 vaccination.

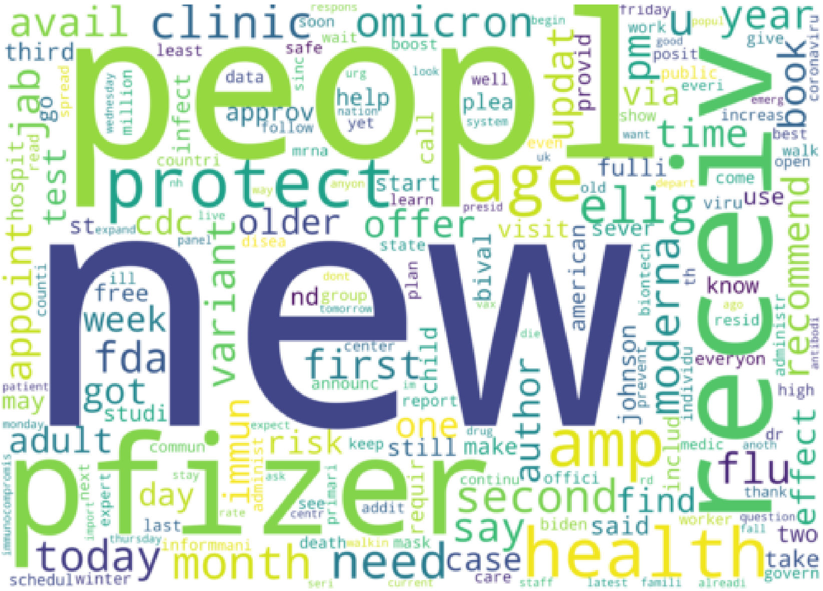

Figure 10. Commonly used words regarding Booster Dose

When individuals aimed to emphasize the positive aspects of vaccines and provide information to others, they frequently employed words such as "strong," "perfect," “full”, and "quick". These expressions signify that certain individuals perceived the vaccination process as a significant lifesaver during this challenging phase. On the other hand, there were instances where individuals held opposing viewpoints. They shared experiences of encountering serious health issues even after receiving their second vaccine dose. Through social media as a platform, these individuals sought to raise awareness among a wider audience, conveying that vaccination alone is not a comprehensive solution to combat the deadly virus. To express their concerns, they commonly used words such as "serious," "bad," "weak," and "hard."

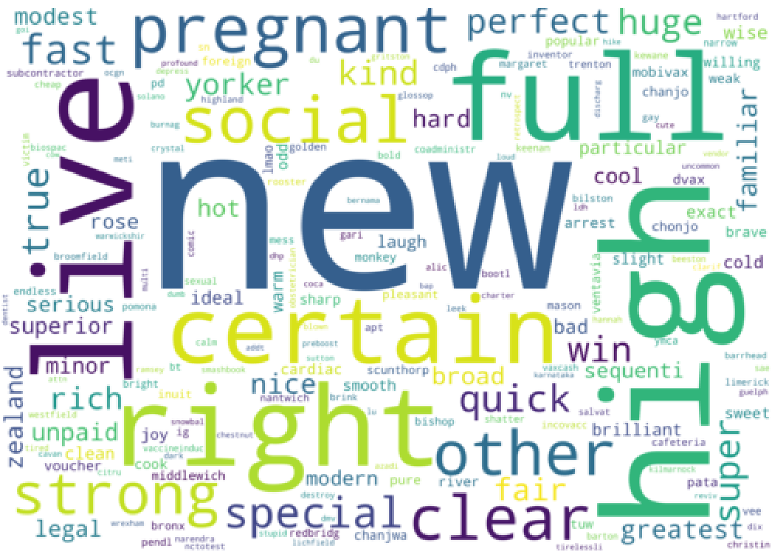

*Figure 11. Positive Words*

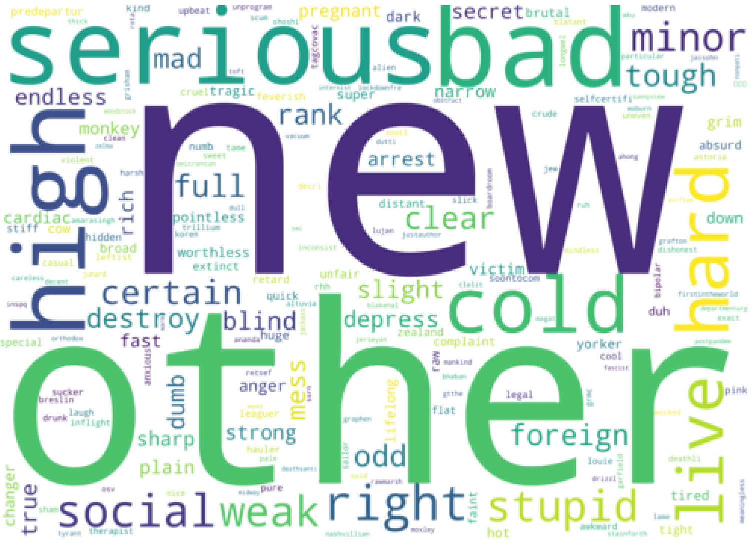

*Figure 12. Negative Words*

On evaluating and analyzing the performance of our model it reveals that the TextBlob sentiment analysis, coupled with TF-IDF vectorization and the LinearSVC model classifier (TextBlob + TF-IDF + LinearSVC), demonstrates superior model performance, showcasing accuracy, precision, recall, and F1 scores of 0.9784, 0.9711, 0.9617, and 0.9663, respectively. Notably, our findings indicate that the incorporation of both stemming and lemmatization techniques has significantly contributed to the overall enhancement of the model's performance.

## 5. Discussion

Our investigation yielded significant findings regarding the overall sentiment surrounding the COVID-19 booster dose, with a predominant inclination towards neutrality or positivity. The optimistic sentiment observed can be attributed to the sense of urgency imposed by the circumstances and the expectation that the vaccine would enhance immunity against emerging variants. According to our study, approximately 87% of the collected tweets exhibited either neutral or positive sentiments. As the vaccination campaign progressed and a larger segment of the population received vaccinations, attitudes towards vaccination demonstrated increased complexity.

The introduction of the booster dose vaccination directly influenced public reactions towards vaccination and the overall COVID-19 situation. Notably, our analysis of monthly tweets during 2021 highlighted a noticeable increase in activity during the period when booster dose vaccinations were introduced. Furthermore, we observed a consistent growth in positive reviews regarding the booster dose vaccination, consequently impacting the distribution of negative and neutral sentiments. The proportion of neutral sentiments decreased from 59.71% to 36.73%, while positive sentiments rose from 27.58% to 48.56%. Conversely, no significant fluctuations were observed in the negative sentiment category, which remained within the range of 11-15%. Over time, the gap between the number of positive and neutral sentiments gradually diminished, indicating a convergence in public opinion.

As depicted in *Figure 8*, the percentage of individuals expressing positive or neutral sentiments regarding COVID-19 booster doses exhibited fluctuations over the observed time period. Notably, there was a slight increase in the percentage of positive reactions toward booster doses in the last quarter of 2021 compared to the midpoint of the same year. Subsequent to the announcement of booster doses, a notable decrease in the percentage of individuals expressing neutral sentiments towards booster doses was observed. Our findings indicate a discernible trend of increased polarization in public opinion regarding booster dose vaccines when compared to the initial months of 2021. Remarkably, the data also indicates that in 2022, despite the continuous global vaccination efforts, sentiment towards vaccination predominantly maintained a positive outlook. This positive trend might be ascribed to an increasing familiarity with the vaccination process and the tangible benefits it offers.

Overall, the analysis presented in the paper emphasizes the significance of comprehensively examining the global vaccination situation and effectively implementing the vaccination process. Furthermore, it emphasizes the imperative of ongoing surveillance of sentiments regarding vaccination to promptly detect and address any challenges or apprehensions that may arise throughout the course of the vaccination campaign. In our experiment, the highest achieved model performance was 0.9784, accompanied by a precision score of 0.9711, utilizing the combination of TextBlob sentiment analysis, TF-IDF vectorization, and the LinearSVC classification algorithm. Notably, this performance surpassed the results reported in a previous study conducted by Qorib et al. in 2023, which attained a performance of 0.9675. It is worth mentioning that the previous study also employed TextBlob sentiment analysis and LinearSVC with TF-IDF vectorization.

## 6. Limitations and Future Works

It is imperative to acknowledge and address several limitations inherent in this study. Firstly, our analysis concentrated solely on perceptions of booster doses expressed on the Twitter platform over an 18-month period. Consequently, viewpoints articulated on other platforms such as Instagram or Facebook were not taken into account, potentially resulting in slight variations in results across different timeframes and platforms. Secondly, the study did not account for the influence of subcultures, which can exert a significant impact on individuals' perceptions of COVID-19 vaccines. Future research endeavors could explore the extent to which subcultures modify or shape individuals' attitudes and perceptions toward COVID-19 vaccines, thereby providing a more comprehensive understanding.

Moreover, our experiment employed a limited set of techniques for sentiment analysis and performance evaluation, primarily relying on classification models with accuracy as the primary metric, aligning with previous studies. Expanding the repertoire of techniques and algorithms utilized could enhance the quality of results and enable a more comprehensive sentiment analysis and performance assessment. Additionally, it is crucial to acknowledge that our study solely focused on English-language tweets, thereby limiting the generalizability of our findings to English-speaking individuals worldwide. Future research could encompass an examination of perceptions across different languages to gain a more nuanced understanding of attitudes toward COVID-19 vaccines within diverse linguistic communities.

## 7. Conclusion

The COVID-19 pandemic, initially detected in Wuhan, China in 2019, prompted the implementation of preventive measures such as compulsory face mask usage, hand sanitization, social distancing, closure of educational institutions, restrictions on large gatherings, and the suspension of international travel, affecting numerous countries worldwide. In light of emerging variants and concerns about diminishing immunity, the administration of booster shots has garnered significant attention as a strategy to enhance protection against the virus and its variants. Nonetheless, the proliferation of rumors and controversies pertaining to vaccines, including potential adverse effects, has had a notable influence on individuals' decisions regarding vaccination acceptance or refusal. Given the scarcity of comprehensive data, individuals often turn to the popular social media platform Twitter to express their sentiments and viewpoints on various subjects, including the COVID-19 situation.

The implementation of booster dose vaccination has resulted in an increase in favorable responses and a reduction in the disparity between neutral and positive sentiments. These observations indicate a growing acceptance of booster doses and a decline in vaccine hesitancy associated with booster doses over time. The utilization of positive descriptors such as 'strong', 'super', and 'high' in reference to booster doses, along with the graphical representation demonstrating a predominance of positive reactions over neutral and negative ones in successive months, indicates a prevailing positive sentiment towards COVID-19 booster dose vaccination. In conclusion, our classification model, employing TextBlob in conjunction with TF-IDF vectorizer and LinearSVC classifier, achieved an accuracy of 0.9784. This accuracy surpasses the previous research study, which utilized various sentiment computation methods, vectorization techniques, and classification models, including the one we employed, resulting in the highest accuracy of 0.96752.

### References

Sv, P., Lorenzo, J. M., Ittamalla, R., Dhama, K., Chakraborty, C., Kumar, D. V. S., & Mohan, T. (2022). Twitter-Based Sentiment Analysis and Topic Modeling of Social Media Posts Using Natural Language Processing, to Understand People's Perspectives Regarding COVID-19 Booster Vaccine Shots in India: Crucial to Expanding Vaccination Coverage. *Vaccines*, *10*(11), 1929. https://doi.org/10.3390/vaccines10111929

Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2022). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems With Applications*, *212*, 118715. https://doi.org/10.1016/j.eswa.2022.118715

Vishwakarma, A., & Chugh, M. (2023). COVID-19 vaccination perception and outcome: society sentiment analysis on Twitter data in India. *Social Network Analysis and Mining*, *13*(1). https://doi.org/10.1007/s13278-023-01088-7

Alabrah, A., Alawadh, H. M., Okon, O. D., Meraj, T., & Rauf, H. T. (2022). Gulf Countries' Citizens' Acceptance of COVID-19 Vaccines—A Machine Learning Approach. *Mathematics*, *10*(3), 467. https://doi.org/10.3390/math10030467

Alabid, N. N. (2021, December 1). *Sentiment analysis of Twitter posts related to the COVID-19 vaccines*. Alabid | Indonesian Journal of Electrical Engineering and Computer Science. https://ijeecs.iaescore.com/index.php/IJEECS/article/view/25942/15822#

Umair, A., & Masciari, E. (2022). Sentimental and spatial analysis of COVID-19 vaccines tweets. *Journal of Intelligent Information Systems*, *60*(1), 1–21. https://doi.org/10.1007/s10844-022-00699-4

Ghosh, P., Dutta, R., Agarwal, N., Chatterjee, S., & Mitra, S. (2023). Social Media Sentiment Analysis on Third Booster Dosage for COVID-19 Vaccination: A Holistic Machine Learning Approach. In *Lecture notes in electrical engineering* (pp. 179–190). Springer Science+Business Media. https://doi.org/10.1007/978-981-19-8477-8_14

Ong, S., Pauzi, M. B. M., & Gan, K. H. (2022). Text Mining and Determinants of Sentiments towards the COVID-19 Vaccine Booster of Twitter Users in Malaysia. *Healthcare*, *10*(6), 994. https://doi.org/10.3390/healthcare10060994

Catelli, R., Pelosi, S., Comito, C., Pizzuti, C., & Esposito, M. (2023). Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy. *Computers in Biology and Medicine*, *158*, 106876. https://doi.org/10.1016/j.compbiomed.2023.106876

Xu, H., Liu, R., Luo, Z., & Xu, M. (2022b). COVID-19 vaccine sensing: Sentiment analysis and subject distillation from twitter data. *Telematics and Informatics Reports*, *8*, 100016. https://doi.org/10.1016/j.teler.2022.100016
Sv, P., Sundar, R., Vajrobol, V., Ittamalla, R., Srividya, K., Farahat, R. A., Chopra, H., Rehman, M. E. U., Chakraborty, C., & Dhama, K. (2023). The Perspectives of Individuals with Comorbidities Towards COVID-19 Booster Vaccine Shots in Twitter: A Social Media Analysis Using Natural Language Processing, Sentiment Analysis and Topic Modeling. *Journal of Pure and Applied Microbiology*. https://doi.org/10.22207/jpam.17.1.54

World Health Organization: WHO. (2022, May 17). Interim statement on the use of additional booster doses of Emergency Use Listed mRNA vaccines against COVID-19. *World Health Organization*. https://www.who.int/news/item/17-05-2022-interim-statement-on-the-use-of-additional-booster-doses-of-emergency-use-listed-mrna-vaccines-against-covid-19

Nezhad, Z. M., & Deihimi, M. A. (2022). Twitter sentiment analysis from Iran about COVID 19 vaccine. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, *16*(1), 102367. https://doi.org/10.1016/j.dsx.2021.102367

Marcec, R., & Likić, R. (2021). Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. *Postgraduate Medical Journal*, *98*(1161), 544–550. https://doi.org/10.1136/postgradmedj-2021-140685

Ansari, T. J., & Khan, N. (2021). Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content. *Electronic Journal of General Medicine*, *18*(6), em329. https://doi.org/10.29333/ejgm/11316

Alam, K. M. R., Khan, M. S., Dhruba, A. R., Khan, M. M., Al-Amri, J. F., Masud, M., & Rawashdeh, M. (2021). Deep Learning-Based Sentiment Analysis of COVID-19 Vaccination Responses from Twitter Data. *Computational and Mathematical Methods in Medicine*, *2021*, 1–15. https://doi.org/10.1155/2021/4321131

*Indonesian Journal of Electrical Engineering and Computer Science*. (n.d.). https://ijeecs.iaescore.com/index.php/IJEECS

Chintalapudi, N., Mittal, M., & Amenta, F. (2021). Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models. *Infectious Disease Reports*, *13*(2), 329–339. https://doi.org/10.3390/idr13020032