Sentiment Analysis on COVID-19 Vaccine Tweets using Machine Learning and Deep Learning Algorithms

Tarun Jain[1], Vivek Kumar Verma[2], Akhilesh Kumar Sharma[3]*, Bhavna Saini[4], Nishant Purohit[5], Bhavika[6], Hairulnizam Mahdin[7], Masitah Ahmad[8]*, Rozanawati Darman[9]*, Su-Cheng Haw[10], Shazlyn Milleana Shaharudin[11], Mohammad Syafwan Arshad[12]

Manipal University Jaipur, Dehmi Kalan, Off Jaipur-Ajmer Expressway, Jaipur, Rajasthan, 303007 India[1, 2, 5, 6]
School of Information Technology, Manipal University Jaipur, Jaipur, Rajasthan[3]
Central University Rajasthan, India[4]
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia[7, 9]
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Selangor Malaysia[8]
Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, 63100, Cyberjaya, Malaysia[10]
Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjung Malim, Perak, Malaysia[11]
Department of Statistics, Columbia University, New York, N.Y., USA[11]
MZR Global Sdn Bhd, 5A, Jalan Kristal K7/K, Seksyen 7, 40000 Shah Alam, Selangor, Malaysia[12]

***Abstract*—One of the main functions of NLP (Natural Language Processing) is to analyze a sentiment or opinion of the text considered. In this research the objective is to analyze the sentiment in the form of tweets towards the Covid-19 vaccination. In this study, the collected tweets are in the form of a dataset from Kaggle that have been categorized into positive and negative depending on the polarity of the sentiment in that tweet, to visualize the overall situation. The reviews are translated into vector representations using various techniques, including Bag-Of-Words and TF-IDF to ensure the best result. Machine learning algorithms like Logistic Regression, Naïve Bayes, Support Vector Machine (SVM) and others, and Deep Learning algorithms like LSTM and Bert were used to train the predictive models. The performance metrics used to test the performance of the models show that Support Vector Machine (SVM) achieved the highest accuracy of 88.7989% among the machine learning models. Compared to the related research papers the highest accuracy obtained using LSTM is 90.59 % and our model has predicted with the highest accuracy of 90.42% using BERT techniques.**

Keywords*—Covid-19 vaccine; sentiment analysis; machine learning; deep learning; natural language processing

I. Introduction

In the wake of COVID-19, the healthcare sector has received considerable attention. Safety regulations such as wearing masks, keeping a good hygiene by washing hands regularly, and maintaining a safe distance from people are especially important now. Nevertheless, these measures can only decrease the spread of the virus, not eliminate it. In this case, vaccination proved to be the sole solution that had the greatest effectiveness in eradicating the coronavirus. But from the very beginning, the acceptance and public sentiment surrounding the COVID-19 vaccine have been subject to varying opinions and concerns. People have had mixed feelings about vaccinations; we have even seen the reluctance of our own family members towards it. Since it is very new to the market, people are not ready to trust the invention and are hesitant about it. This hesitancy and skepticism have highlighted the need to delve deeper into understanding the sentiments of individuals towards the vaccine.

Nowadays, the Internet is the best source for any company to learn about public perceptions of their products and services. For its rich knowledge, the business community is tapping into social media content. It has been utilized to carry out marketing and branding initiatives for organizations in the areas of innovation, product design, and stakeholder relations. It is a useful means of communicating and sharing information with the public for government and non-profit organizations. Every day, people use social media platforms such as Facebook, Twitter, and Instagram to voice their opinions and thoughts. These platforms have emerged as powerful sources for expressing opinions and thoughts, making them ideal for capturing and analyzing public sentiments. People began voicing their concerns on the COVID-19 vaccination process as soon as it began. And since COVID 19 has affected so many lives but the vaccine showed up as a ray of hope amidst these extreme conditions, it has been extremely important to analyze the sentiments of people towards the COVID-19 vaccine [1].

Sentiment analysis is one major task of NLP (Natural Language Processing). It is also called Opinion mining. It is done to capture the author's feelings, emotion towards an entity. Sentiment analysis tries to capture this information by analyzing unstructured text data in the form of reviews and comments [2]. By harnessing the potential of sentiment analysis, the aim is to gain valuable insights into the perceptions and emotions of individuals regarding the COVID-19 vaccine. Sentiment analysis particularly is helpful when it comes to negative reviews. It helps discover the exact

shortcomings of the products. This requires the text to be classified into two sentiment polarities that are positive and negative (or neutral). In this research study, various textual and numerical features from tweets are extracted, evaluated on how they correlate, and used to predict sentiment of people related to the COVID-19 vaccine. The goal is to contribute to the existing knowledge on sentiment analysis and its applicability in the context of COVID-19 vaccines, ultimately aiding in the development of informed strategies and interventions for public health initiatives.

Further, the paper is divided into different sections to bring out all parts of the study properly. Section II shows the works of other people on sentiment analysis related to the COVID vaccine and how people have reciprocated to it. Section III gives information about the background of this study, highlighting a contrast between machine learning and deep learning techniques. Next is Section IV that highlights a major part of the study. It starts with the analysis of the data used in the study. Then it talks the complete process followed to carry out the study- the machine learning and deep learning algorithms used, the pre-processing techniques and feature extraction methods for both the type of algorithms, the performance metrics used to analyze the result of the model and lastly a comparison between the way machine learning and deep learning algorithms work [3][4][5][6]. Section V gives a deep analysis of the results obtained on the data with machine learning and deep learning techniques. It shows the results for all the five machine learning algorithms used with different feature extraction methods and then bar plots comparing their accuracy. Then, it shows the results for the two deep learning models used and their plots for accuracy and loss. Section VI highlights the conclusions and key takeaways from the study, along with the discussion of future plans and scope.

II. Literature Review

Sentiment analysis is being used in a large spectrum of fields right now. And a lot of people are increasingly interested in researching it. In the past year there have been much research on the sentiment of people towards the COVID-19 vaccine that came into existence. One such research employing tweets collected between December 21 and July 2, 2021, had information on the most prevalent vaccines that had just become available around the world. It used a tool called VADER to analyze people's sentiment towards certain vaccines. The tool found that 33.96% responses were positive, 17.55% were negative and the left 48.49% were neutral responses. It applied the basic data preprocessing steps and feature extraction algorithms on the tweets in the dataset. Then it finally used a recurrent neural network (RNN) that included LSTM and Bi-LSTM where LSTM secured an accuracy of 90.59% and Bi-LSTM obtained an accuracy of 90.83%. This study contributes to a better knowledge of public opinion on COVID-19 vaccinations and advances the goal of removing the virus worldwide [7]. Another such study used tweets in general and then only from four countries with the most tweets on the COVID-19 vaccine: India, USA, Canada, and England. It consumed two text mining methods that are LDA and VADER to extract the sentiment from those tweets. The overall analysis showed that there were almost twice people that had a positive feeling towards the COVID vaccine than those having a negative feeling. However, the country-specific analysis showed that the people's sentiment remained consistent for the vaccines that were approved in their country, while most people had some fear towards other vaccines [8]. Another research performing sentiment analysis had tweets that were taken from the 14th to the 18th of January 2021. Covishield and Bharat Biotech's Covaxin were two vaccines employed in this work. The purpose of this study was to examine the sentiments expressed in tweets about these two vaccines in India. It used the Syuzhet package version 1.0.1 to classify tweets based on sentiments into positive and negative as well as eight other emotions (fear, joy, anticipation, anger, disgust, sadness, surprise, trust). It used the NRC Emotion Lexicon to analyze the tweets. The study showed that while most of the population has positive feelings about these vaccines, there are also negative feelings about them, according to the analysis, associated with the sentiments such as fear and anger [9]. Another study introduces a lexicon-based framework for sentiment classification of tweets, categorizing them as positive, negative, or neutral. The results indicate that the proposed system surpasses existing systems in terms of performance and accuracy [10].

This research makes use of a dataset containing tweets about the opinion of people about vaccines like Pfizer/BioNTech, Sinopharm, etc. in the Kaggle data repository. Tweepy, a Python package, was used to capture the data. It synthesizes the dialogue surrounding worldwide immunization attempts and progress using an API called TextBlob and using word cloud visualizations. TextBlob classified roughly half of the tweets in the dataset as neutral, and the other half comprised of 75% positive tweets and 25% tweets depicting a negative sentiment [11]. This study used web scrapping to extract the data from online news and blogs to work on. TextBlob library was used to analyze the sentiments of the public opinion collected. They gave a result that more than 90% of the articles had positive sentiments towards the vaccination drive [12]. From thirteen Reddit communities, data was collected regarding COVID-19 vaccines. LDA topic modelling was done on this data, and it was found that most of the communities had a positive sentiment towards the drive and found no major change in the opinions of people since December 2020 [13]. A RapidMiner software for data science was used for classifying English and Filipino tweets with the help of Naïve Bayes to conduct opinion mining. The results showed that the research had an accuracy of 81.77%. Their conclusion was that majority of people were enthusiastic and supportive of the vaccination drive [14]. This study takes about 1.2 million tweets to perform NLP and sentiment analysis on them to find out useful insights about the approach of people towards the COVID-19 vaccine and the measures to stay safe from the virus. This research used TextBlob and Vader as sentiment analysis tools and performed time series forecasting at a later stage. The result showed that many people have a positive view of the vaccination drive than negative. But more than that, people were highly conscious about maintaining hygiene and social distancing to combat the spread of the virus [15].

A study that was conducted in Indonesia was focused on analyzing the opinion of Indonesian people towards the newly

introduced vaccine. It collected the data from twitter using Rapid miner tool. The results of this study were slightly different. They showed about 39% of positive sentiment and 56% of negative sentiment and 1% of neutral sentiment. The people did not really trust that the vaccine was safe for them to consume [16]. Following this one, another study was conducted to evaluate the opinion of people of Indonesia about the two most prevalent vaccines, Sinovac and Pfizer in the country and understand people's view on both. The best performing model came out to be Support Vector machine and the study concluded that people were more positive for the Pfizer vaccine as compared to Sinovac. While about 77% of the tweets indicated a positive sentiment towards Sinovac, this number shot up to 81% in the case of Pfizer vaccine [17].

In contrast to a country-specific approach, some studies were conducted to analyze the sentiments of people towards the vaccine on a global level, for different countries. This study collected about 820,000 tweets and analyzed the sentiment of those tweets in two stages. In the first stage, the sentiments of people towards the vaccine around the globe was considered and the findings showed which countries had an overall positive attitude and which countries had a negative one. This stage also included gender-based analysis about the sentiments of people to address those issues in a different way. The second phase included the tweets to be organized into word clouds to analyze the most used words and sentiments by people of different countries [18]. This study made use of 928,402 tweets collected from different countries and the six most popular vaccines' tweets were picked from them to perform the analysis. They conducted an aspect-based analysis considering health, policy etc. and used four Bert models. The total accuracy was found out to be 87% and the F1 score lied between 84% to 88% [19]. This study used two different approaches to understand people's hesitancy towards the vaccine. These were machine learning based and lexicon based. It divided the dataset into two cultures English and Arabic and studied them separately. The study analyzed the performance of both the approaches on the datasets and then used the better performing approach for the spatiotemporal analysis [20]. This research collected the English language tweets posted over a course of 3 months and applied the Vader tool to classify the tweets as positive, negative, and neutral. The results revealed that out of the 2,678,372 tweets in consideration, 42.8% were positive, 26.9% were neutral and 30.3% were negative. The important topics from the positive and negative tweets were drawn out using latent Dirichlet allocation analysis, and these topics were then subjected to a geographical and temporal analysis. The study concluded that the highest positive sentiment tweets came from United Arab Emirates and the lowest positive sentiment tweets came from Brazil. Also, the sentiment score increased a good amount at the start, then slowly decreased and finally remained almost the same till the end of the period of the tweets [21].

The works discussed above indicate that previous studies have focused on sentiment analysis using various techniques and datasets. Furthermore, these studies highlight the use of various machine learning and deep learning models for sentiment classification but do not delve deep into comparative studies that evaluate the performance of different models, feature extraction techniques, and sentiment analysis tools to identify the most effective approaches for sentiment analysis in the context of COVID-19 vaccination. This study employs a comprehensive range of techniques and models, including both machine learning algorithms (Logistic Regression, Naïve Bayes, Support Vector Machine) and deep learning algorithms (LSTM and BERT). It compares the performance of different feature extraction techniques, namely Bag of Words and TF-IDF, to showcase their impact on sentiment analysis accuracy. By utilizing these diverse approaches, it provides a better understanding of the most effective feature extraction methods and evaluation of the performance of different models, highlighting the strengths and weaknesses of each in the context of COVID-19 vaccination sentiment analysis.

III. BACKGROUND

Machine learning and Deep learning both fall under the umbrella of Artificial Intelligence, but they are more efficient in serving different purposes. Deep learning involves the use of something called a neural network which replicates how a human brain works to solve complex problems. But it requires a large amount of data to function with great accuracy, unlike machine learning which can work on lesser amounts of data. Machine learning learns from the data that is provided and makes intelligent predictions on the new data that is fed to it, with some human intervention. Both machine learning and deep learning models have been used on this dataset, but Deep learning has been preferred since the results it gave had better accuracy. Both machine learning and deep learning have a slight difference in how it classifies the data into sentiments.

For machine learning algorithms, firstly input data is fed into the system and pre-processing is performed on it to make it easier for the classification algorithm to classify it. Pre-processing includes converting the whole text into upper or lowercase, removal of extra words such as special characters or words that add no sentiment to the sentence. Then feature extraction is performed onto this data that extracts the important features from the tweet and converts it into vectors for the algorithm to be able to process it. This makes it easier for classification and improves the accuracy of the model in consideration. The next step is the model training which is when the model is trained onto the given data to classify the sentiment of the tweets and then it is tested using the test dataset to give the output or the sentiment of the tweets fed to the model. Here, the sentiment of the tweets can be positive or negative [22].

For deep learning algorithms, the process is slightly different. The input data goes through pre-processing to remove the extra words from the tweets and then it is fed to the deep learning algorithm which takes care of both the feature extraction and model training phase for classifying the sentiment of the data provided. Further the model is tested to check its performance on a test dataset, like how the machine learning algorithms do it [23].

Deep learning solves the problem end-to end whereas machine learning first fragments the problem into smaller statements and then solves it incrementally. In machine learning the data is undergone through feature extraction first and then classification is performed but in deep learning,

feature extraction and classification are performed simultaneously [24]. Deep learning works adequately on large amounts of data by giving higher accuracies. High end systems are required to run deep learning algorithms as it mainly focuses on GPU of the system. Whereas a domain expert is required in Machine learning to spot and reduce the complexity of the data for the traditional algorithms to work. When the amount of data which is fed to the model is less, machine learning performs better than the deep learning models. But as the amount of data increases, the rate with which the performance of a machine learning model was increasing rapidly falls and remains almost the same with further increase in the amount of data. However, for a deep learning model the performance steadily increases with the increase in the amount of data fed to the model. For larger amounts of data, deep learning models perform a lot better than machine learning models. There are two deep learning models used in this study: BERT and LSTM. Both are discussed in detail in the following section.

A. BERT

Bidirectional Encoder Representations from Transformers also known as BERT is a deep learning model which was published by researchers at Google in 2018. It works on the encoder-decoder network where self-attention is used on the encoder side and attention is used on the decoder side. Large text corpus is used to train the Bert model, this gives the model the ability to understand better and grasp variability in data patterns on several NLP tasks. Being bidirectional it gives the model the freedom to learn and understand the context of a word from both the left and the right sides while training the model. This nature of the model helps it to understand the language deeply. Also, for the model to work well, some amount of pre-processing is done on the data. This makes the BERT model suitable for a variety of NLP tasks.

B. LSTM

LSTM also known as Long Short-Term Memory network are a part of a unique kind of RNN that has the potential to learn long-term dependencies. LSTM can remember information for long periods of time without any struggle and reduces the impact of short-term memory. Recurrent neural networks have chain type structure where each module is intertwined several times. LSTMs have a similar structure but instead of caring a single neural network there are four that are connected to each other in a unique way. LSTM networks retain the relevant information from the prior data in the sequence that helps in processing the incoming data points. There are three things that are important to determine the output of LSTM: the cell state, the previous hidden layer, and the input data at the current timestamp. The cell state is like the memory of the network. An LSTM cell has three gates: One is the forget gate, which allows it to forget the irrelevant information from the prior timestamp. The equation for the forget gate is:

$$f_t = \sigma(x_t * U_f + H_{t-1} * W_f) \quad (1)$$

Here, x_t is the input to the current timestamp, U_f is the weight, H_{t-1} is the hidden state of previous timestamp and W_f is the weight matrix of the hidden state.

Next is the input gate, which decides which information must be kept from the current timestamp. The equation for the input gate is:

$$i_t = \sigma(x_t * U_i + H_{t-1} * W_i) \quad (2)$$

Here, x_t is the input to the current timestamp, U_i is the weight matrix of input, H_{t-1} is the hidden state of previous timestamp and W_i is the weight matrix of input corresponding with hidden state.

The last one is the output gate, which determines what the hidden state will be for the next timestamp. The equation for the input gate is:

$$o_t = \sigma(x_t * U_o + H_{t-1} * W_o) \quad (3)$$

Here, x_t is the input to the current timestamp, U_o is the weight matrix of output, H_{t-1} is the hidden state of previous timestamp and W_i is the weight matrix of output corresponding with hidden state [25][26].

IV. Methodology

A. Data Collection and Analysis

In the initial stages of the research, a dataset comprising 10,000 tweets regarding people's opinions on the COVID-19 vaccine was sourced from Kaggle. However, it was observed that a significant portion of the dataset consisted of neutral tweets. Recognizing the potential impact of these neutral tweets on the data consistency and the subsequent model performance, a decision was made to remove them from the dataset. This cleaning process resulted in a refined dataset of approximately 3,700 tweets. Out of the total tweets, approximately 2,000 exhibited a positive sentiment towards the COVID-19 vaccine, while around 1,700 displayed a negative sentiment. This balanced distribution of positive and negative sentiments provides a suitable foundation for training and evaluating machine learning models.

B. Data Cleaning and Preprocessing

To start with, data pre-processing steps were applied to the given dataset to ensure the data's quality and consistency. These pre-processing steps are crucial in preparing the dataset for accurate model training and reliable outcomes. The dataset was thoroughly cleaned by removing any extraneous data or unnecessary elements that could introduce irregularities. This involved eliminating punctuation, symbols like "#," and Twitter handles such as "@user." Further, this involved removing URLs from the tweet texts, converting the text to lowercase to eliminate case sensitivity, and applying tokenization to break down the text into individual words or tokens. Stopwords, which are commonly used words in a language like "the", "and", "is", were removed from the text. These words are often irrelevant for sentiment analysis and can introduce noise into the data. After that an important step was lemmatization. This was applied to reduce words to their base or root form. This helps in standardizing the text data by converting variations of a word. These pre-processing steps were implemented to ensure the dataset's consistency and to avoid poor model training and inaccurate outcome due to inconsistencies in the dataset [27]. After applying data pre-processing techniques to enhance the quality of the dataset, the

next step involved splitting the dataset into two distinct parts: a training dataset and a testing dataset. The dataset was split in a ratio of 80:20, with 80% of the tweets allocated to the training dataset and the remaining 20% reserved for the testing dataset. This ensures that a substantial portion of the data is utilized for training the model while still leaving a sizable portion for evaluation.

C. Feature Extraction and Model Training

When it comes to the machine learning models, the tweets in the training set undergo a process of vector representation using techniques like Bag of Words and TF-IDF. These techniques filter out irrelevant words and convert the tweets into numerical representations. By utilizing these vectorized representations, the classification algorithms are trained and tested on the given dataset. The results obtained from the testing dataset provide insights into the performance of these models in analyzing the sentiments [28]. The next step involves the classification of the data. In this process, several machine learning algorithms are utilized to train classifiers that can accurately predict the sentiment of the tweets. The following algorithms are applied to the training dataset: Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. Each algorithm learns from the training data, capturing patterns and relationships between the tweet features and their corresponding sentiments. Once the classifiers are trained, they are evaluated using the testing dataset. By comparing the predicted sentiments with the actual sentiments of the tweets in the testing dataset, various performance metrics such as accuracy, precision, recall, and F1-score are calculated. Finally, the ROC AUC score for each model is compared. It is used to assess the degree of separability between the different classes. A higher ROC AUC score indicates that the model performs well in terms of classification [29]. Thus, these metrics provide insights into how well each classifier can performs in classifying the sentiments of the COVID-19 vaccine-related tweets.

On the other hand, the process for deep learning models differs slightly. In this case, the pre-processed data is directly fed into the deep learning algorithm. The deep learning algorithm itself takes care of both the feature extraction and model training phases. This feature extraction process is performed by the hidden layers of the neural network. It automatically learns and extracts relevant features from the data during the training process, eliminating the need for explicit feature extraction. Deep learning models consist of multiple layers of interconnected artificial neurons. Each neuron receives input signals, applies a mathematical operation, and produces an output signal. The outputs from one layer serve as inputs to the next layer, forming a hierarchical representation of the data. Once the model is trained, it can be tested on a separate test dataset, like how the machine learning algorithms are evaluated. The performance of the deep learning model on the test dataset helps analyze its effectiveness in sentiment classification.

V. Results

The results for the machine learning and deep learning models have been separately illustrated. Deep learning models perform better than the machine learning models with a maximum accuracy of 90.42%.

A. Machine Learning Models

The results of the proposed models: Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), Decision Tree and Random Forest are shown in this section. Different assessment criteria, including Accuracy, Precision, Recall, F-score, Confusion matrix, and ROC curve, were utilized to evaluate the models reviewed here. First, the results for each model using Bag of Words feature extraction algorithm are shown and then using TF-IDF feature extraction techniques are shown. Then, using a unique feature extraction technique that works well, the comparison between the results for each classification model is displayed.

B. Support Vector Machine

Support vector machine model is used for classification, and this algorithm works on the concept of finding a hyperplane that provides the best separability between different classes.

Support Vector Machine model with Bag of Words analyzed and registered 274 positive tweets correctly, 74 positive tweets incorrectly, 362 negative tweets correctly and 31 negative tweets incorrectly as per the confusion matrix. Based on this, Table I shows the major classification metrics precision, recall and f1-score for both the classes individually. Here, 1 is for the positive tweets and 0 is for the negative ones. For the positives, the precision score is 0.83, the f1-score is 0.87 and the recall is 0.92. For the negatives, the precision score is 0.90, the f1-score is 0.84 and the recall is 0.79. The total accuracy of the model comes out to be 0.86.

TABLE I. Result of Support Vector Machine using Bag of Words

	precision	recall	f1-score	support
0	0.90	0.79	0.84	348
1	0.83	0.92	0.87	393
accuracy			0.86	741

Support Vector Machine with TFIDF analyzed and registered 307 positive tweets correctly, 41 positive tweets incorrectly, 351 negative tweets correctly and 42 negative tweets incorrectly as per the confusion matrix. Based on this, Table II shows the major classification metrics precision, recall and f1-score for both the classes individually. Here, 1 is for the positive tweets and 0 is for the negative ones. For the positives, the precision score is 0.90, the f1-score is 0.89 and the recall is 0.89. For the negatives, the precision score is 0.88, the f1-score is 0.88 and the recall is 0.88. The total accuracy of the model comes out to be 0.89.

TABLE II. Result of Support Vector Machine using TF-IDF

	precision	recall	f1-score	support
0	0.88	0.88	0.88	348
1	0.90	0.89	0.89	393
accuracy			0.89	741

C. Naïve Bayes

Naïve Bayes classifier is based on probability. It assumes that each variable input to the classifier is independent but gives good accuracy when applied. It uses conditional probability for obtaining the result. Conditional probability is basically calculating the probability of completing a certain task given a certain condition must always be satisfied.

Naïve Bayes Model with Bag of Words analyzed and registered 281 positive tweets correctly, 67 positive tweets incorrectly, 355 negative tweets correctly and 38 negative tweets incorrectly in the confusion matrix. Based on this, Table III shows the major classification metrics precision, recall and f1-score for both the classes individually. Here, 1 is for the positive tweets and 0 is for the negative ones. For the positives, the precision score is 0.84, the f1-score is 0.87 and the recall is 0.90. For the negatives, the precision score is 0.88, the f1-score is 0.84 and the recall is 0.81. The total accuracy of the model comes out to be 0.86.

TABLE III. RESULT OF NAIVE BAYES USING BAG OF WORDS

	precision	recall	f1-score	support
0	0.88	0.81	0.84	348
1	0.84	0.90	0.87	393
accuracy			0.86	741

Naïve bayes with TFIDF analyzed and registered 271 positive tweets correctly, 77 positive tweets incorrectly, 371 negative tweets correctly and 22 negative tweets incorrectly in the confusion matrix. Based on this, Table IV shows the major classification metrics precision, recall and f1-score for both the classes individually. Here, 1 is for the positive tweets and 0 is for the negative ones. For the positives, the precision score is 0.83, the f1-score is 0.85 and the recall is 0.78. For the negatives, the precision score is 0.83, the f1-score is 0.88 and the recall is 0.94. The total accuracy of the model comes out to be 0.87.

TABLE IV. RESULT OF NAIVE BAYES USING TF-IDF

	precision	recall	f1-score	support
0	0.92	0.78	0.85	348
1	0.83	0.94	0.88	393
accuracy			0.87	741

D. Logistic Regression

Decision tree classifier is used for classification and regression. It forms a tree-like structure and learns simple decision rules to predict the target class value.

Logistic Regression with Bag of Words analyzed and registered 287 positive tweets correctly, 61 positive tweets incorrectly, 358 negative tweets correctly and 35 negative tweets incorrectly in the confusion matrix. Based on this, Table V shows the major classification metrics precision, recall and f1-score for both the classes individually. Here, 1 is for the positive tweets and 0 is for the negative ones. For the positives, the precision score is 0.85, the f1-score is 0.88 and the recall is 0.91. For the negatives, the precision score is 0.89, the f1-score is 0.86 and the recall is 0.82. The total accuracy of the model comes out to be 0.87.

TABLE V. RESULT OF LOGISTIC REGRESSION USING BAG OF WORDS

	precision	recall	f1-score	support
0	0.89	0.82	0.86	348
1	0.85	0.91	0.88	393
accuracy			0.87	741

Logistic Regression with TFIDF analyzed and registered 295 positive tweets correctly, 53 positive tweets incorrectly, 350 negative tweets correctly and 43 negative tweets incorrectly in the confusion matrix. Based on this, Table VI shows the major classification metrics precision, recall and f1-score for both the classes individually. Here, 1 is for the positive tweets and 0 is for the negative ones. For the positives, the precision score is 0.87, the f1-score is 0.88 and the recall is 0.89. For the negatives, the precision score is 0.87, the f1-score is 0.86 and the recall is 0.85. The total accuracy of the model comes out to be 0.87.

TABLE VI. RESULT OF LOGISTIC REGRESSION USING TF-IDF

	precision	recall	f1-score	support
0	0.87	0.85	0.86	348
1	0.87	0.89	0.88	393
accuracy			0.87	741

E. Decision Tree Classifier

Decision tree classifier is used for classification and regression. It forms a tree-like structure and learns simple decision rules to predict the target class value.

Decision Tree Classifier analyzed and registered 257 positive tweets correctly, 91 positive tweets incorrectly, 347 negative tweets correctly and 46 negative tweets incorrectly in the confusion matrix. Based on this, Table VII shows the major classification metrics precision, recall and f1-score for both the classes individually. Here, 1 is for the positive tweets and 0 is for the negative ones. For the positives, the precision score is 0.79, the f1-score is 0.84 and the recall is 0.88. For the negatives, the precision score is 0.85, the f1-score is 0.79 and the recall is 0.74. The total accuracy of the model comes out to be 0.82.

On plotting its confusion matrix, Decision Tree classifier with TFIDF analyzed and registered 259 positive tweets correctly, 69 positive tweets incorrectly, 322 negative tweets correctly and 72 negative tweets incorrectly in the confusion matrix. Based on this, Table VIII shows the major classification metrics precision, recall and f1-score for both the classes individually. Here, 1 is for the positive tweets and 0 is for the negative ones. For the positives, the precision score is 0.78, the f1-score is 0.80 and the recall is 0.82. For the negatives, the precision score is 0.78, the f1-score is 0.76 and the recall is 0.74. The total accuracy of the model comes out to be 0.78.

TABLE VII. RESULT OF DECISION TREE USING BAG OF WORDS

	precision	recall	f1-score	support
0	0.85	0.74	0.79	348
1	0.79	0.88	0.84	393
accuracy			0.82	741

TABLE VIII. RESULT OF DECISION TREE USING TF-IDF

	precision	recall	f1-score	support
0	0.78	0.74	0.76	348
1	0.78	0.82	0.80	393
accuracy			0.78	741

F. Random Forest Classifier

Random Forest classifier uses many decision trees and finds the average of the results from these trees to obtain improved accuracy for prediction.

Table IX shows the major classification metrics precision, recall and f1-score for both the classes individually. Here, 1 is for the positive tweets and 0 is for the negative ones. For the positives, the precision score is 0.81, the f1-score is 0.86 and the recall is 0.92. For the negatives, the precision score is 0.89, the f1-score is 0.81 and the recall is 0.75. The total accuracy of the model comes out to be 0.84. On plotting its confusion matrix, the random forest model analyzed and registered 258 positive tweets correctly, 90 positive tweets incorrectly, 363 negative tweets correctly and 30 negative tweets incorrectly as shown in Table IX.

TABLE IX. RESULT OF RANDOM FOREST USING BAG OF WORDS

	precision	recall	f1-score	support
0	0.89	0.75	0.81	348
1	0.81	0.92	0.86	393
accuracy			0.84	741

Table X shows the major classification metrics precision, recall and f1-score for both the classes individually. Here, 1 is for the positive tweets and 0 is for the negative ones. For the positives, the precision score is 0.80, the f1-score is 0.87 and the recall is 0.94. For the negatives, the precision score is 0.92, the f1-score is 0.82 and the recall is 0.74. The total accuracy of the model comes out to be 0.85. On plotting its confusion matrix, the random forest classifier with TFIDF analyzed and registered 255 positive tweets correctly, 93 positive tweets incorrectly, 364 negative tweets correctly and 29 negative tweets incorrectly as shown in Table X.

TABLE X. RESULT OF RANDOM FOREST USING TF-IDF

	precision	recall	f1-score	support
0	0.92	0.74	0.82	348
1	0.80	0.94	0.87	393
accuracy			0.85	741

G. Comparing Results

The presented data in Table XI provides a comprehensive overview of the performance metrics evaluated across various models, specifically focusing on accuracy, precision, recall, and F1-score. These metrics were meticulously analyzed using the Bag of Words technique as the chosen method for feature extraction. By examining these performance indicators, valuable insights are gained into the effectiveness and efficiency of each model in the context of the analyzed dataset.

TABLE XI. COMPARING RESULTS OF ALL THE MODELS USING BAG OF WORDS

Model	Accuracy	Precision	Recall	F1-score
SVM	0.8583	0.8733	0.8303	0.9211
Naive Bayes	0.8583	0.8712	0.8412	0.9033
Logistic Regression	0.8704	0.8818	0.8544	0.8906
Decision Tree	0.7841	0.8010	0.7834	0.8193
Random Forest	0.8556	0.8715	0.8250	0.9236

The following graph in Fig. 1 shows a comparison of accuracy score of different models with Bag of Words as the feature extraction method.

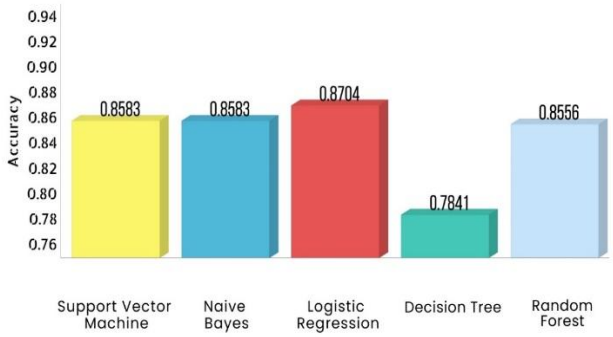

Fig. 1. Accuracy bar plot for machine learning models with bag of words as feature extraction method.

The following graph in Fig. 2 shows a comparison of ROC AUC score of different models with Bag of Words as the feature extraction method.

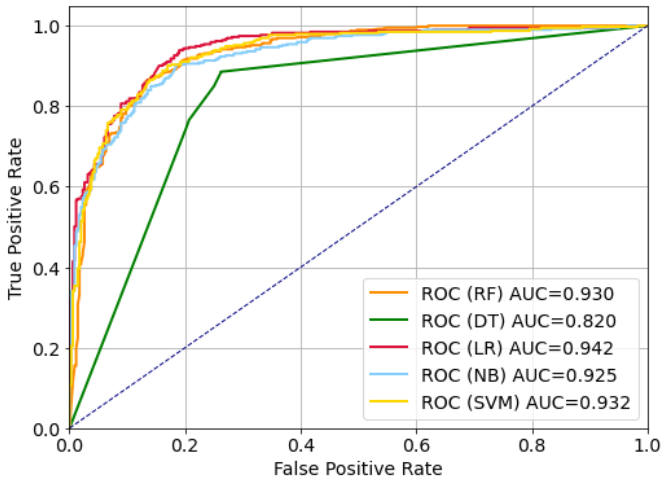

Fig. 2. ROC curve machine learning models with bag of words as feature extraction method.

Here, 'RF' is Random Forest, 'DT' is Decision Tree, 'LR' is Logistic Regression, 'NB' is Naïve Bayes and 'SVM' is Support Vector Machine.

From the ROC curve and the accuracy bar plot, it can be observed from these plots that most of the classifiers perform decently, and Logistic Regression classifier with Bag of Words feature extraction method performs the best with AUC score of 0.942 and an accuracy of 87.0445%. Close to it is the Support Vector Machine and Naive Bayes classifier with an accuracy of 85.8300%.

The presented data in Table XII provides a comprehensive overview of the performance metrics evaluated across various models that are accuracy, precision, recall, and F1-score. These metrics have been diligently analyzed for all the models, with a specific focus on the utilization of the TF-IDF technique as the chosen approach for feature extraction.

TABLE XII. COMPARING RESULTS OF ALL THE MODELS USING TF-IDF

Model	Accuracy	Precision	Recall	f1-score
SVM	0.8879	0.8943	0.8954	0.8931
Naive Bayes	0.8664	0.8823	0.8281	0.9440
Logistic Regression	0.8704	0.8794	0.8685	0.8906
Decision Tree	0.7841	0.8010	0.7834	0.8193
Random Forest	0.8367	0.8585	0.7944	0.9338

The following graph in Fig. 3 shows a comparison of accuracy scores of different models with TF-IDF as the feature extraction method.

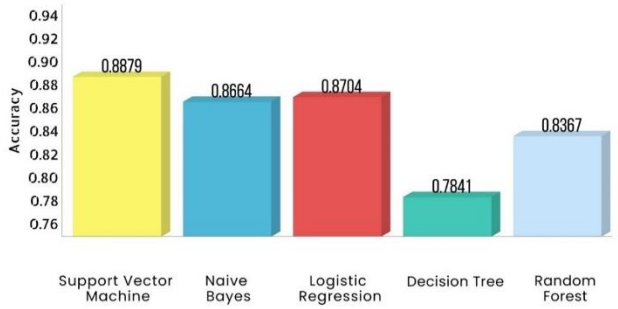

Fig. 3. Accuracy bar plot for machine learning models with TF-IDF as feature extraction method.

The following graph in Fig. 4 shows a comparison of ROC AUC scores of different models with TF-IDF as the feature extraction method. Here, 'RF' is Random Forest, 'DT' is Decision Tree, 'LR' is Logistic Regression, 'NB' is Naïve Bayes and 'SVM' is Support Vector Machine.

From the ROC curve and the accuracy bar plot, it can be observed from these plots that all the classifiers perform decently, and Support Vector Machine classifier with TF-IDF feature extraction method performs the best with an AUC score of 0.95 and an accuracy of 88.7989%. Close to it is the Logistic Regression and Naive Bayes classifier with an accuracy of 87.0445% and 86.6397% respectively.

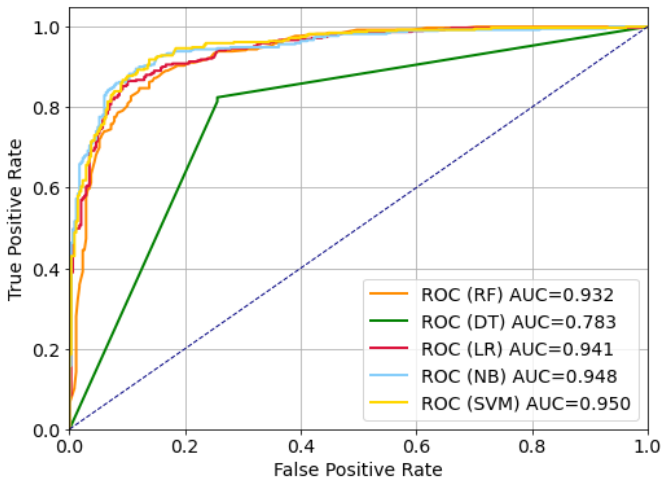

Fig. 4. ROC curve machine learning models with TF-IDF as feature extraction method.

H. Deep Learning Models

1) BERT: The deep learning model BERT worked efficiently with the dataset and gave a validation accuracy of 90.42%. The mode ran for three epochs where it gave an accuracy of 89.88% in first epoch, 89.20% in the second epoch and finally 90.42% in the third epoch which was the highest.

The above graph in Fig. 5 shows the relationship between loss and the learning rate. The model experienced the minimum loss when the learning rate was around 10^ (-4).

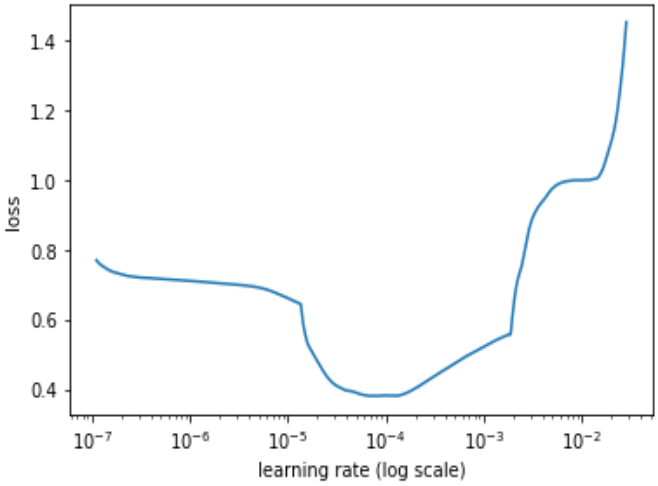

Fig. 5. A plot between learning rate and loss.

2) LSTM: The deep learning model LSTM worked decently with the dataset and gave a validation accuracy of 88.26. The mode ran for 5 epochs where it gave an accuracy of 70.04 in first epoch, 86.50 in the second epoch, 87.58 in the third epoch, 87.72 in the fourth and finally 88.26 in the fifth epoch which was the highest. The following graph in Fig. 6 showcases the plot between the accuracy and the number of epochs with the training and the validation set.

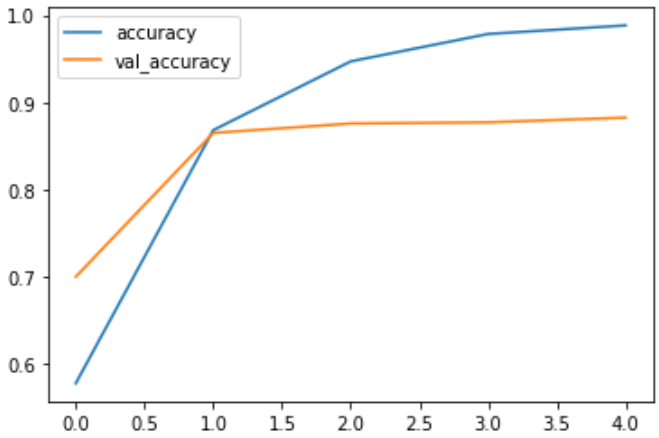

Fig. 6. Accuracy of LSTM model for both training and test set. Here accuracy refers to the training set accuracy and val_accuracy refers to the testing set accuracy.

Here the validation accuracy goes nearly constant after intersecting with accuracy at 0.88 whereas the accuracy plot keeps on increasing and takes over after the intersection.

The following graph in Fig. 7 showcases the plot between the loss and the number of epochs with the training and the validation set.

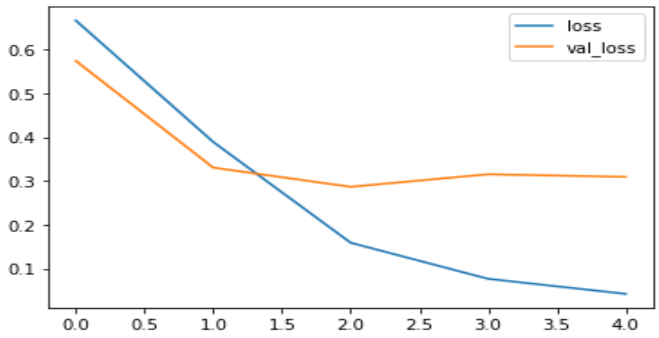

Fig. 7. Loss of LSTM model for both training and test set. Here loss refers to the training set loss and val_loss refers to the testing set loss.

VI. Conclusions

The aim of this study is to analyze the sentiments of people about the COVID-19 vaccine that has been introduced recently through the social media platform 'Twitter'. To be able to evaluate the opinion of the public, a dataset with the most recent tweets was taken and applied two word embedding techniques to them. Five machine learning algorithms and two deep learning algorithms have been utilized for classification of tweets into positive and negative. Experimental results suggest that out of the machine learning models used, Support vector machine when used with TF-IDF as word embedding technique gives the highest accuracy. However, deep learning models give a higher accuracy. LSTM model when used with some preprocessing gave the accuracy 88.26% after four epochs. They helped in analyzing that most people have a positive outlook for the COVID-19 vaccine, while some part of the population is still hesitant about it. The possible reasons for the same can be that people fear that the vaccine might have side effects, or they might not be open to accept a new vaccine introduced to the market, or they are not aware enough about the consequences of not taking the COVID vaccine. Compared to the related research papers the highest accuracy obtained using LSTM is 90.59 % and our model has predicted with the highest accuracy of 90.42% using BERT techniques. This study can be of utmost importance to organizations analyzing the sentiment of a large population towards the COVID-19 vaccine in turn acting as a tool to find out ways to cope with the problem. It can help them find what section of society is hesitant and why, so that they can probably change something or improve the quality of services they provide.

However, it should be noted that this study uses only two feature extraction methods, Bag of Words and TF-IDF. Future work might consider utilizing alternative feature extraction methods such as Word2Vec and GloVe to further improve the effectiveness of the models. Another important aspect to consider can be the geographic and cultural context. While this study analyzed sentiments on a global level, further research could focus on sentiment analysis within specific regions or countries. This would allow for a better understanding of the variations in public opinion and can help identify country-specific challenges, such as vaccine hesitancy, misinformation, or unique socio-political factors that influence sentiment.

Acknowledgment

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through Matching Grant (vot H995).

References

[1] Khakharia, A.; Shah, V.; Gupta, P. Sentiment Analysis of COVID-19 Vaccine Tweets Using Machine Learning. Rochester, NY June 18, 2021. [Google Scholar]

[2] Liu, B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies 2012, 5, 1–167. [Google S]

[3] Bhavya Joshi, Akhilesh Kumar Sharma, Narendra Singh Yadav & Shamik Tiwari (2022) DNN based approach to classify Covid'19 using convolutional neural network and transfer learning, International Journal of Computers and Applications, 44:10, 907-919, DOI: 10.1080/1206212X.2021.1983289

[4] Ramani, P., Pradhan, N., Sharma, A.K. (2020). Classification Algorithms to Predict Heart Diseases—A Survey. In: Gupta, M., Konar, D., Bhattacharyya, S., Biswas, S. (eds) Computer Vision and Machine Intelligence in Medical Image Analysis. Advances in Intelligent Systems and Computing, vol 992. Springer, Singapore.

[5] A. K. Sharma, K. I. Lakhtaria, A. Panwar and S. Vishwakarma, "An Analytical approach based on self organized maps (SOM) in Indian classical music raga clustering," 2014 Seventh International Conference on Contemporary Computing (IC3), Noida, India, 2014, pp. 449-453, doi: 10.1109/IC3.2014.6897215.

[6] Shrivastava, D.K., Sharma, A.K., Bhardwaj, S. (2021). Prediction of COVID'19 Outbreak by Using ML-Based Time-Series Forecasting Approach. In: Singh, P.K., Polkowski, Z., Tanwar, S., Pandey, S.K., Matei, G., Pirvu, D. (eds) Innovations in Information and Communication Technologies (IICT-2020). Advances in Science, Technology & Innovation. Springer, Cham

[7] Alam, K.N.; Khan, M.S.; Dhruba, A.R.; Khan, M.M.; Al-Amri, J.F.; Masud, M.; Rawashdeh, M. Deep Learning-Based Sentiment Analysis of COVID-19 Vaccination Responses from Twitter Data. Computational and Mathematical Methods in Medicine 2021, 2021, 1-4. [Publisher Site] [Google Scholar]

[8] Yin, H.; Song, X.; Yang, S.; Li, J. Sentiment Analysis and Topic Modeling for COVID-19 Vaccine Discussions. World Wide Web 2022, 25, 1067–1083. [Google Scholar]

[9] Dubey, A. D. Public Sentiment Analysis of COVID-19 Vaccination Drive in India. Rochester, NY January 24, 2021. [Google Scholar] [CrossRef]

[10] Asghar, Dr. M.; Kundi, F.; Khan, A.; Ahmad, S. Lexicon-Based Sentiment Analysis in the Social Web. Journal of basic and applied scientific research 2014, 4, 238–248. [Google Scholar]

[11] Dua, S. Sentiment Analysis of COVID-19 Vaccine Tweets. Medium. https://towardsdatascience.com/sentiment-analysis-of-covid-19-vaccine-tweets-dc6f41a5e1af (accessed 2023-01-17).

[12] Bhagat, K. K.; Mishra, S.; Dixit, A.; Chang, C.-Y. Public Opinions about Online Learning during COVID-19: A Sentiment Analysis Approach. Sustainability 2021, 13, 3346. [Google Scholar] [CrossRef]

[13] Melton, C. A.; Olusanya, O. A.; Ammar, N.; Shaban-Nejad, A. Public Sentiment Analysis and Topic Modeling Regarding COVID-19 Vaccines on the Reddit Social Media Platform: A Call to Action for Strengthening Vaccine Confidence. Journal of Infection and Public Health 2021, 14, 1505–1512. [Google Scholar] [CrossRef]

[14] Villavicencio, C.; Macrohon, J. J.; Inbaraj, X. A.; Jeng, J.-H.; Hsieh, J.-G. Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes. Information 2021, 12, 204. [Google Scholar] [CrossRef]

[15] Sattar, N. S.; Arifuzzaman, S. COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA. Applied Sciences 2021, 11, 6128. [Google Scholar] [CrossRef]

[16] Pristiyono; Ritonga, M.; Ihsan, M. A. A.; Anjar, A.; Rambe, F. H. Sentiment Analysis of COVID-19 Vaccine in Indonesia Using Naïve Bayes Algorithm. IOP Conf. Ser.: Mater. Sci. Eng. 2021, 1088, 012045. [Google Scholar] [CrossRef]

[17] Nurdeni, D. A.; Budi, I.; Santoso, A. B. Sentiment Analysis on Covid19 Vaccines in Indonesia: From The Perspective of Sinovac and Pfizer. In 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT); 2021; pp 122–127. [Google Scholar] [CrossRef]

[18] Ansari, M. T.; Khan, N. Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content. Electronic Journal of General Medicine 2021, 18, em329. [Google Scholar] [CrossRef]

[19] Aygün, İ.; Kaya, B.; Kaya, M. Aspect Based Twitter Sentiment Analysis on Vaccination and Vaccine Types in COVID-19 Pandemic With Deep Learning. IEEE Journal of Biomedical and Health Informatics 2022, 26, 2360–2369. [Google Scholar] [CrossRef]

[20] Alsabban, M. Comparing Two Sentiment Analysis Approaches by Understand the Hesitancy to COVID-19 Vaccine Based on Twitter Data in Two Cultures. In 13th ACM Web Science Conference 2021; WebSci '21; Association for Computing Machinery: New York, NY, USA, 2021; pp 143–144. [Google Scholar] [CrossRef]

[21] Liu, S.; Liu, J. Public Attitudes toward COVID-19 Vaccines on English-Language Twitter: A Sentiment Analysis. Vaccine 2021, 39 (39), 5499–5505. [Google Scholar] [CrossRef]

[22] Ikonomakis, M.; Kotsiantis, S.; Tampakas, V. Text Classification Using Machine Learning Techniques. WSEAS TRANSACTIONS on COMPUTERS 2005, 4, 966-974. [Google Scholar]

[23] Tang, D.; Wei, F.; Qin, B.; Liu, T.; Zhou, M. Coooolll: A Deep Learning System for Twitter Sentiment Classification. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014); Association for Computational Linguistics: Dublin, Ireland, 2014; pp 208–212. [Google Scholar]

[24] Zulfiker, Md. S.; Kabir, N.; Biswas, A. A.; Zulfiker, S.; Uddin, M. S. Analyzing the Public Sentiment on COVID-19 Vaccination in Social Media: Bangladesh Context. Array 2022, 15, 100204. [Google Scholar]

[25] Nyawa, S.; Tchuente, D.; Fosso-Wamba, S. COVID-19 Vaccine Hesitancy: A Social Media Analysis Using Deep Learning. Ann Oper Res 2022. [Google Scholar]

[26] Nuser, M.; Alsukhni, E.; Saifan, A.; Khasawneh, R.; Ukkaz, D. Sentiment analysis of covid-19 vaccine with deep learning. Journal of Theoretical and Applied Information Technology 2022, 100, 1-3. [Google Scholar]

[27] Didi, Y.; Walha, A.; Ben Halima, M.; Wali, A. COVID-19 Outbreak Forecasting Based on Vaccine Rates and Tweets Classification. Computational Intelligence and Neuroscience 2022, 2022, e4535541. [Google Scholar]

[28] Soni, K. M.; Gupta, A.; Jain, T. Supervised Machine Learning Approaches for Breast Cancer Classification and a High Performance Recurrent Neural Network. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA); 2021; pp 1–7. [Google Scholar]

[29] Yacouby, R.; Axman, D. Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems; Association for Computational Linguistics: Online, 2020; 79–91. [Google Scholar]