National College of Ireland

Sentiment Classification of Public Perception regarding Covid-19 Vaccine: Deep Learning and Stacking Ensemble of Machine Learning Approach

MSc Research Project
MSc in Data Analytics

Dimple
20176759

School of Computing
National College of Ireland

Supervisor: Prof. Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Dimple
Student ID:	20176759
Programme:	MSc Data Analytics
Year:	2021-22
Module:	Research Project
Supervisor:	Prof. Vladimir Milosavljevic
Submission Due Date:	16th Dec 2020
Project Title:	Sentiment Classification of Public Perception regarding Covid-19 Vaccine: Deep Learning and Stacking Ensemble of Machine Learning Approach
Word Count:	7682
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Dimple
Date:	30th January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Classification of Public Perception regarding Covid-19 Vaccine: Deep Learning and Stacking Ensemble of Machine Learning Approach

Dimple
20176759

Abstract

World Health Organization (WHO) identified Coronavirus as an illness disease, can be named Covid-19. In March 2020, the Covid-19 (define as 'CO' stands for Corona, 'VI' for the Virus, and 'D' for Disease) has been declared a global pandemic that had a significant impact on the economy of the world and a major impact on the lifestyle of people. The entire pharmaceutical industry was focusing on the development of a safe and efficient vaccine. In early December 2020, several vaccinations have been developed and licensed by WHO. Mass production has been started and delivered in several countries. The announcement of a 90 Percent effective rate vaccine raised hope in people, and they have found social media to be a valuable source of sharing their opinions related to vaccines. These perspectives enable researchers to engage in data mining based on public reviews and opinions and provide important insights to help government make better decisions. The purpose of this paper is to examine the evolution of public opinions on the tweets regarding the Covid-19 vaccine on social media like Twitter. For this analysis, the dataset is collected from the Kaggle website. Various libraries of the Natural Language Processing toolkit are used to clean the impurities and noise from data. Cleaned data is not labelled. So, the sentiment analysis approach is used to label the data and categorize it into positive, negative, and neutral sentiments. Text Blob is a python library used to find the polarity and subjectivity score of the labelled dataset. To achieve the objectives, the dataset is split into train and test and then both deep learning models and stack models of machine learning are applied. The results of the models are compared and find that Convolutional Neural Network (CNN) and Long term short memory (LSTM) model has the highest accuracy of 66 Percent to predict the sentiments of the public related to the Covid-19 vaccine on tweets.

1 Introduction

Coronavirus is not a contagious condition that may be shared by touch or tiny droplets formed when coughing, sneezing, or talking. On 8th December 2019, the first two cases were detected as Coronavirus patients in Wuhan, China. This illness rapidly spread worldwide that increased the number of infected coronavirus cases and death rates. WHO declared the Covid-19 pandemic in March 2020 and then lockdown has been announced in some parts of the world. These unexpected conditions have had a significant impact

on the lives of billions of people all over the world. The Covid-19 vaccine with 90 Percent effective has been licensed by WHO in December 2020 and started delivered in some countries.

Other hand, people were not able to get connected physically due to lockdown and social distancing. Here now, modern technology plays a vital role and social media became the surviving platform for people (Chakraborty et al.; 2020). Twitter is one of the social media platforms that is used by a considerable number of users. Through a global status update message called tweet, users often broadcast information on personal or public real-life occurrences, or just express their thoughts, ideas, or opinions on a particular topic, product, service, or event (*Monitoring the public opinion about the vaccination topic from tweets analysis*; 2019). After the approval of vaccination, people started sharing their opinions, fears, thoughts related to the Covid-19 vaccine on Twitter by posting stories and posts. Many researchers have examined the influence of various social media initiatives on vaccine reluctance and public opinion on the vaccination procedure. Thus, this study aims to examine and analyze the Covid-19 vaccine-related tweets over Twitter. This paper deals with the sentiments of the public towards the Covid-19 vaccine by using sentiment analysis. This research starts with the collection of dataset from the Kaggle and then, Sentiment analysis and Natural Language Processing (NLP) techniques are used to make the extracted data more cleaned and qualitative for further analysis.

• Motivation and Background

The sentimental analysis is a complex procedure that includes sentiment analysis and subjective analysis, among other things. Machine learning (ML), natural language processing (NLP), and computational linguistics are all areas where this new research is taking place. Positive, neutral, and negative feelings are all subclasses of one another (Saad and Saberi; 2017). In this paper, tweets are collected from Twitter using the Twitter API and categorize the data based on the sentiments related to the Covid-19 vaccine.

COVID-19 has been quickly spreading over the world since December 2019, producing serious respiratory infections and potentially fatal outcomes for many people. There has been an outbreak of the pandemic since March 2020. Only a fraction of the more than 260 vaccinations that have been proposed have been given the go-ahead (Yang and Sornlertlamvanich; 2021). The safety of many people is still a concern, and some people have refused to welcome them because of this. Social isolation, quarantine, and travel restrictions enforced by the government make it impossible to gather evidence on the vaccine's safety. That's the reason people be afraid to get vaccinated. In **Figure 1**, the total number of people who received at least one dose of the Covid-19 vaccine is divided by the total population of mentioned countries. To study the text classification of tweets and sentiment analysis on their emotions, many researchers have been approached Natural language processing techniques. For the evaluation of results, machine learning models like support vector machine SVM and K-Nearest Neighbor (KNN) have been used with great accuracy (Adamu, Bin Mat Jiran, Gan and Samsudin; 2021).

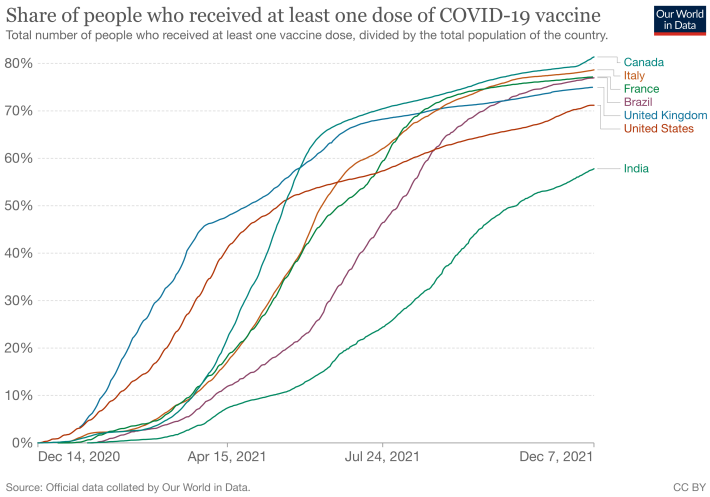

Figure 1: Vaccinated people by time throughout the world

A variety of methods exist for analyzing sentiment in neural networks. Modelling word-level sentiment information with an MLP and tweet-level sentiment information with CNN is called Multi-level Sentiment-enriched Word Embedding (MSWE), created by Xiong et al. SVM is utilized for sentiment classification in their model, which also learns sentiment-specific word embeddings (Xiong et al.; 2018). The problem with these algorithms is that they don't understand the context of the text because they employ the Bag of Words technique. This is a supervised learning issue, which implies that the full dataset must be explicitly labelled before these techniques can be used. The subjective nature of the labelling procedure makes it more likely that mistakes will be introduced into the dataset.

Focus - To overcome the limitation of manually labelling the dataset, an approach of Text Blob python library will be used in this study. Then, stacking ensemble methodology of machine learning models and a few models of deep learning will be used to evaluate the result and will be compared their results based on their accuracy.

- **Novelty**

The novelty of this paper aims to compare the deep learning and hybrid method of machine learning algorithms to analyze the public sentiments towards the Covid-19 vaccine over Twitter.

- **Research Question**

"How can sentiments of public related to Covid-19 vaccine using natural language processing technique and deep learning classifier (RNN, CNN, MLP and LSTM) and machine learning stack modelling of Decision tree Classifier, XGBoost Classifier and Random Forest Classifier be analyzed over Twitter?"

- **Objectives**

Several objectives have been developed, implemented, assessed, and ultimately findings have been presented to answer this research question. The main aim and objectives of the study are:

1. Identify previous work on sentiment analysis of public opinions towards the Covid-19

vaccine.
2. the public opinions on the Covid-19 vaccine from the Kaggle website.
3. the extracted data by cleaning unwanted and hollow text using a natural language processing toolkit.
4. Classify the data into three sentiments categories- positive, negative, and neutral using a Text Blob approach and visualize them using python libraries.
5. Compare the deep learning (CNN, MLP, RNN and LSTM) and machine learning stack model of Decision tree Classifier, XGBClassifier and Random Forest Classifier.

2 Literature Review

An essential part of the research paper's literature evaluation is the detailed study of previous research on the topic. Here, the study question and hypothesis are laid forth in detail. Several algorithmic models are compared to analyze the influence of Machine Learning and deep learning techniques in this review. There has been a lot of studies done on the topic of sentiment analysis.

• Review of the Sentiment Analysis on Tweets by approaching different techniques

A lot of people are connected to different platforms of social media. Twitter is one of the applications of social media where a million users included from the main characters of the world to the general people share their thoughts, opinions, expressions about the highlighted topic or field. Sentiment analysis of Twitter data determines the opinion of the public. (Wagh and Punde; 2018) used the approach of tweet collection from Twitter. Python or R programming language was used to collect the data based on the keywords that are used as input. Preprocessing of the text data is also called the filter of unstructured data by removing Retweets, URLs, Stop words, numbers. Stemming and tokenization techniques of natural language processing were used to clean the data. Data retrieved after preprocessing on public opinion tweets further labelled into three categories positive, negative, and neutral using sentiment analysis.

With the use of Hadoop, author (Trupthi et al.; 2017) created an interactive automated system that predicts the sentiments of social media reviews and tweets, which can analyze a massive quantity of information. Real-time tweets, which are a rich source of data for opinion mining and sentiment analysis, were examined and analyzed sentiments of public opinion. Using SNS services, data was extracted by the streaming API of Twitter. Hadoop technique was used to store the extracted data. It is possible to extract many words with their positive and negative probability using MapReduce, which is an efficient method. The result of the reduction was a list of terms, each with a good and negative score. Using MongoDB, these scores were saved in a database. Natural language processing technique was used to remove the words with the help of POS tagging. POS tagging is the part of the NLP toolkit that is used to distinguish the words of a sentence into grammar-based parts of speech. An adjective like "young" may have a different level of feeling than an adverb like "old". The tweets were classified using the categorization module into positive, negative, or neutral and mapper code split the data into two files – integer and tweets (string). Then, reduce checked each word if it is a string and does this

word occurs in more than one tweet. If it occurs, then word scored negative else positive.

• **Review of Natural Language Processing techniques for sentiment analysis**
Natural Language Processing technique is the subset of AI that is used to transform the human language-based data into machine-level language and derive the meaning of the text. To get the insight of sentiments of the public about the Covid-19 pandemic on social media, (Dumre et al.; 2021) approached different NLP techniques to make the dataset more reliable to build the models. All text of each word was converted into lower cases to minimize the word vector's size. In this paper, Tokenization is an NLP task that was used to split the sentence into small pieces of tokens like punctuations, symbols, words. After tokenization, the next approach was to remove the symbols and punctuations. Stemming and Lemmatization are the NLP techniques that were used to reduce the formation of the words as the base words.

As stated by (Solangi et al.; 2018), Opinion Mining and Sentiment analysis can be reviewed by the natural language processing method. Tokenization, POS tagging and parsing were the methods used to pre-process the data. "Stanford Tokenizer", "Open NLP Tokenizer" are the fundamental techniques of NLP tools that were used to remove the little helpful words like 'a', 'the' from the sentence. Text segmentation method used to find the boundaries between two different languages. This is an important factor of NLP tasks that were used in this paper for the segmentation of Chinese words in sentences. POS tagging was used to label the text and produce the lexical data and the Parsing technique was used to give the linguistic structure of the sentence that acquired the syntactic data
Several NLP tools have been introduced in this paper while performing the tokenization and Chinese segmentation. Fudan tool in Java language that supports the named entity recognition and dependency parsing. Language Technology Platform (LTP) and Niu Parser tools in C++ language that support the semantic role labelling, syntactic parsing and semantic parsing modules.

• **Review of the machine and deep learning models for sentiment classification**
Covid-19 in the US-related data on Twitter were researched and analyzed the sentiments by (Khan et al.; 2021). The text Blob method was used to label each tweet and Bag-of-Words (BoW) and TF-IDF both feature techniques were used to keep the expressive information. For US-based COVID-19 Tweets, the machine learning classifiers were trained on tweets to predict people's attitudes as positive, negative, or natural. Random Forest, Gradient Boosting Machine, Extra Tree Classifier, Logistic Regression and Support vector machine were implemented and received 92 percent, 96 percent, 95 percent, 92 percent and 94 percent accuracy respectively.

(Roy and Ojha; 2020) proposed deep learning models to predict the sentiments of the tweets related to the Covid-19 pandemic. SemEval-2016 Twitter data was collected in which sentiments were categorized. To find the efficiency of classifier models, data were preprocessed by cleaning, removing noise, and transforming. Bidirectional Encoder Representations from Transformers (BERT) is a deep learning model that primarily relies on attention to determine the contextual link between words (or sub-words) in a text.

Individual words are represented as real-valued vectors in a predetermined vector space in word embeddings, which is a class of approaches. Because each word is mapped to a single vector and the vector values are acquired in a manner that mimics that of a neural network, the approach is frequently grouped with deep learning. Bidirectional long short-term memory BiLSTM is used to receive the input in forwarding and backward directions. Convolutional Neural Networks algorithm is a neural network that consists of multiple layers of vector format data and gives output in the convolutional layer. The result of the three models was compared based on their accuracy. In findings, the BERT algorithm performed well with 64percent accuracy whereas LSTM and CNN evaluated the accuracy 60percent and 59percent respectively.

• **Review of Sentiment Analysis on Covid-19 vaccine tweets by implementing machine learning and deep learning algorithms**

There have been relatively few studies conducted so far on sentiment analysis concerning the Covid-19 vaccine.

(Cotfas et al.; 2021) has proposed the research paper to analyze the sentiments of the public towards the Covid-19 vaccine over Twitter. The first vaccine was announced on 9, November 2020 and to date of the first vaccine took place in the US, data was collected using Twitter Filtered Stream API from Twitter. By removing emails, links and undesired words and columns, data was preprocessed successfully that was followed by implementing machine learning models Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), Bidirectional Encoder Representations from Transformers (BERT), and Random Forest (RF) on it. The finding of this paper is BERT outperforms with an accuracy of 79 percent. The limitation of this work is several classification models can be used to get high accuracy. Moreover, the selected dataset was from November 9, 2020, to December 8, 2020. This work has also helped in the deep understanding of stance analysis. (Paul and Gokhale; 2020) carried out the research work to perform the sentiment classification on tweets about the Covid-19 vaccine from the day President Trump announced 'Operation warp Seed' using retweet library in R programming language. The data pre-processing was done using feature extraction and a natural language processing toolkit that was followed by the implementation of supervised machine learning models. This work has findings that the LSTM model outperforms with 82 percent accuracy.

(Adamu, Jiran, Gan and Samsudin; 2021) proposed the sentiment analysis and text analytics on tweets related to the Covid-19 vaccine using natural language processing techniques. Cross-Industry Standard Platform for Data Mining (CRISP-DM) technology was adopted in this study that followed the architecture of research by collecting data from tweets. That collected data was noisy and required cleaning. For the pre-processing data, tokenization, stemming, lemmatization, normalization of characters techniques were used. Apart from it, they used the count vector and term frequency-inverse document frequency (TF-IDF) technique to transform the text data into features and then this process was followed by text classification. According to the results of implemented machine learning models, Support Vector Machine (SVM) outperformed K-Nearest Neighbour (KNN) with an accuracy of 88percent. The findings of this work are positive sentiments are more than negative but less than neutral because it may signal that most individuals remain unsure about the vaccine's efficacy. As per the research, future work can be expanded by using a huge dataset from different web sources. Naïve Bayes classifier is the

best method of machine learning that applied on the large set of text data for analysis. Elbagir and Yang (2018) has the findings that the performance of naïve Bayes algorithm is the best for sentiment analysis of Covid-19 vaccine over Twitter. She proposed two more machine learning models Maximum entropy and SVM on the extracted and pre-processed tweet-based data along with semantic analysis in which the polarity of the sentiment was shown for the users.

In Indonesia, Nurdeni et al. (2021) has done sentiment analysis on the two types of extracted data of the Covid-19 vaccine on Twitter. Search-tweet-API and Tweepy python libraries were used to crawl tweets and filtered into two different terms based on the name of the Covid-19 vaccine - @vaksin Sinovac and @vaksin Pfizer. For sentiment analysis, both datasets were manually labelled and categorized into positive, negative, or neutral sentiment. NLTK was used to pre-process the labelled data using case folding (lower case conversion), Tokenization (chopping text into the words), Stop word removal and Stemming techniques that were followed by the feature extraction using PoS (Part of Speech) tagger. As a result, it is observed that the implementation of three classification models Naïve Bayes, Support Vector Machine, and Random Forest of machine learning performed well on both datasets and SVM performed outstandingly with accuracy 85percent and 78 percent in Sinovac and Pfizer dataset respectively. One more research has been done by Solanki and Palwe (2021) in which Bernoulli Naïve Bayes, Linear SVC and Random Forest algorithms were applied to the extracted and pre-processed Twitter data. It was observed that Bernoulli NB and Linear SVC predict 80 percent accuracy for sentiment analysis. Last but not least research paper has the findings that a huge share of the population has more neutral opinions to get vaccinated which indicates that they are still confused about whether should get a vaccination or not ? Amjad et al. (2021) approached the same method of extracting tweets, data pre-processing and sentiments analysis. The text Blob approach was used to score the data to categorical variables and label them into afraid, neutral and not afraid. After all these processes, it was observed that data was imbalanced that was balanced by using the Synthetic Minority Oversampling Technique (SMOTE) technique. To train the machine learning models, Support vector machine, random forest, AdaBoost and Multi-layer perceptron (MLP) algorithms were implemented. This study has the findings that the SVM model outperforms with an accuracy of 89 percent. These findings shed light on how sentiment analysis may be used in a broader range of research areas, including vaccine research for the Covid-19 virus.

• Conclusion of related work

There are also many kinds of research that have been done previously on the sentiment analysis of the Covid-19 vaccine. All papers used the tweets crawling method by using different Python and R libraries and different techniques of natural language processing to clean the impure text data. Several machine learning algorithms were implemented and compared their performance based upon their accuracy. There is only one paper Cotfas et al. (2021) that approached the deep learning model CNN and compared it with three other machine learning models and found that BERT methodology works better and gives good results. No major or direct study has been done on determining whether or not there is a comparison between the stacking model of machine learning and deep learning models for the sentiment analysis of the Covid-19 vaccine over Twitter. So, the present studies collect the tweets data from the Kaggle website and implement the

stacking of three machine learning models and three several deep learning models on the collected and pre-processed data and compare their accuracy to predict the accurate sentiments of the public towards the Covid-19 vaccine.

3 Research Methodology

This section explains the Scientific methodology, architectural and scientific design, and data collection procedures of this project from beginning to completion. This project, according to Goebel (2014), employs a modified KDD (Knowledge Discovery in Databases) technique. The technique, as shown in **Figure 2**, outlines the procedure and utilizes the KDD concept for sentiment analysis on Covid-19 Vaccine over Twitter.

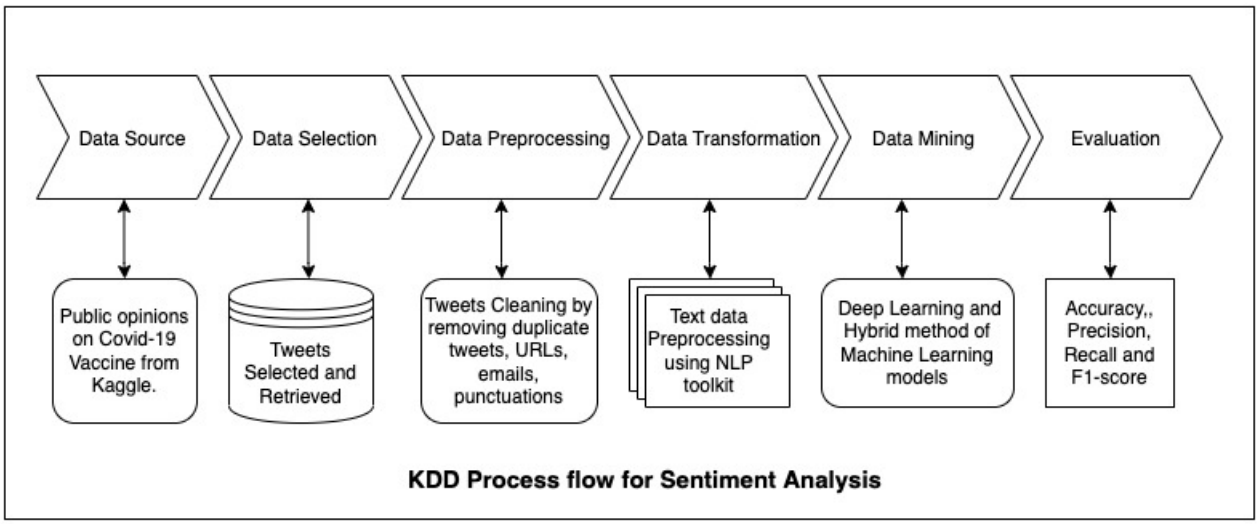

Figure 2: KDD Process flow for Sentiment Analysis

• **Data Source** – A twitter dataset related to the Covid-19 vaccine is required for sentiment analysis. The Kaggle website is the source of our data that is designed for data scientists to publish datasets, explore machine learning models.

• **Data Selection** – The dataset used in this study is freely available online, and it had been released for research purposes by a well-known author Gabriel Preda. He collected recent tweets on the Covid-19 vaccines from Twitter's official website. For this, the Tweepy package of the python library was used to get access to Twitter API. Different types of Covid-19 vaccines related tweets were collected following as - Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V. To use this data for this present study, available CSV file (size – 90.11 MB) has been downloaded. There are 16 columns and 224249 rows i.e., tweets in the dataset. https://www.kaggle.com/gpreda

• **Data Pre-processing** – Data pre-processing is the technique to clean the impure data. The available data set is in raw format that requires cleaning. Removing duplicate tweets, converting date column to date format, and sorting the data date-wise methods are used to clean the data. There is one column named 'source' that explains the platform of Twitter-like Android, Web app and more. Visualization of the tweet counts vs User Location and visualization of tweet count vs platform of Twitter as shown in **Figure 3**. From the dataset, all columns except user-location, date and text are dropped because

those are not useful for sentiment analysis.

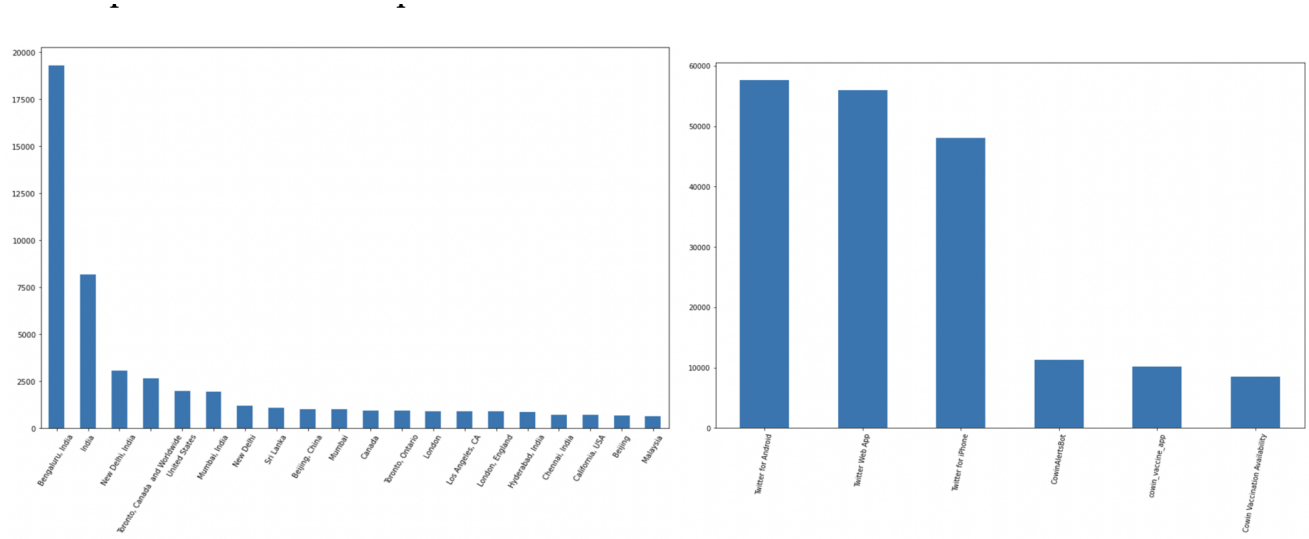

Figure 3: Tweet Count vs Location and Platform of Twitter

Now, the next step is to clean the impure text data from each tweet. Natural Language Processing toolkit is the python package used to clean the impure text into a suitable form by removing URLs, Emails, newline characters, distracting single quotes, punctuations and stop words.

• **Data Transformation** – Now next step is to transfer the data into an appropriate format before applying the machine learning and deep learning models. For sentiment analysis, the text blob approach is used to find the polarity and subjectivity of the tweets. Then, the emphasis is on labelling or classifying each tweet into one of three groups (positive, negative, and neutral) and storing it into cleaned text. Columns user-location and text are not of use now, so dropping both. Now, the dataset has cleaned text. Along with polarity, subjectivity, and sentiments, is ready for implementing the models on it.

• **Data Mining** – In this phase, the implementation of deep learning and machine learning techniques are done on the prepared and labelled data. Firstly, data is split into test and train. In machine learning models, Decision tree Classifier, XGBClassifier and Random Forest Classifier models and in deep learning, CNN, RNN and LSTM algorithms are implemented to predict the sentiments of the public related to Covid-19 vaccine on Twitter.

• **Data Interpretation / Evaluation** – During this phase, Confusion Matrix is used to find the Accuracy and Loss and depicted against positive and negative sentiments. Algorithms' implementations provide the required F1 scores, recall, accuracy, and precision for measuring the performance of the stack model of machine learning. The result is derived from Machine Learning and deep learning models have been visualized in the form of patterns and compared to which model has good accuracy.

4 Design Specification

The workflow of the research study employing the Machine Learning methods and Deep learning Methods is explained in this part. The process for this study is depicted in

Figure 4.

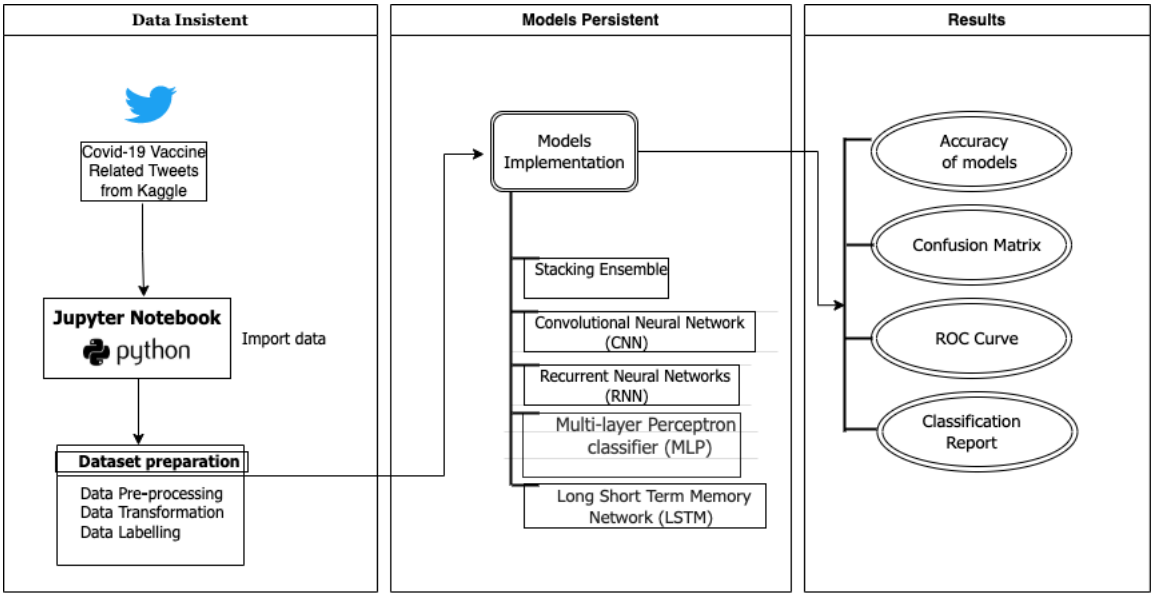

Figure 4: Customized Research Methodology

- Collect the tweets data from the Kaggle.

• Text data is the most impure data and requires a lot of processing. To extract insights from the data, cleaning and pre-processing like removing emails, URLs, stop words has been done.
• The sentiment approach is used for labelling the data set into three cases – Positive, Negative, and Neutral.
• The next stage was to use Python's Text blob package to evaluate the polarity and subjectivity of each tweet.
• Implementation of Machine Learning and Deep Learning algorithms.
• As a result, compare the accuracy of both Machine Learning models and deep learning techniques.

5 Implementation and Evaluation of Topic modelling

This section describes the outputs of transformed data and the implementation of algorithms that predicts the sentiments of people about the Covid-19 vaccine on Twitter. To determine the attitude of tweets, machine learning and deep learning strategy has been used. This approach requires a labelled dataset for training the classification models. Following that, evaluations and analyses of the findings will be carried out using performance metrics.

This research study is done by using different tools and techniques, and programming language that is followed as:
• Tools – Google Collab (execute the code through the browser, connected with Google Drive).
• Programming Language – Python with 3.6.9 version as per Google Collab

• Main Libraries – TensorFlow, NLTK, Tokenizer, WordCloud, sklearn, stop words, confusion matrix and Matplotlib.
• Techniques – Natural language processing with NLTK library, Sentiment Analysis with Text Blob approach.

Techniques used for this study and their output that comes after running the code are defined as –

• Natural Language Processing

Natural Language Processing (NLP) is a technique used to pre-process the human language data that computer programmers can understand and use for analysis. Textual data is always unstructured and impure for analysis. Text preprocessing is the way to clean the textual data that can be available for further steps. NLTK is the python package that is also known as the natural language processing toolkit. To use this toolkit, the NLTK library must be installed in a python environment and imported into the Jupyter notebook. Tokenization is a method of NLP that is used for removing the stop words and conversion of the case of the text. Stop words are common words like 'a', 'the', 'has' and more are removed from each tweet and then, all words are converted into the lower case to remove the complexity of data.

• Sentiment analysis

Sentiment analysis is a technique of natural language processing that is used to label the data. Dataset labelling is the process where text is divided into categories like 'Good' or 'Bad'. The data is selected and preprocessed, needs to be labelled. The attitude of public opinions towards the Covid-19 vaccine has been categorized into three parts – Positive, Negative, and Neutral. The text Blob approach is applied along with the sentiment analysis that returns the two properties of sentiments – Polarity and Subjectivity. The next step is to store the cleaned and labelled data into a data frame shown in **Figure 5**.

	date	cleaned_text	results	polarity	subjectivity	sentiment
0	2020-12-20	folks said daikon paste could treat cytokine s...	{'polarity': 0.0, 'subjectivity': 0.0, 'sentim...	0.0	0.000000	Neutral
1	2020-12-13	world wrong side history year hopefully bigges...	{'polarity': -0.5, 'subjectivity': 0.9, 'senti...	-0.5	0.900000	Negative
2	2020-12-12	coronavirus sputnikv astrazeneca pfizerbiontec...	{'polarity': 0.0, 'subjectivity': 0.0333333333...	0.0	0.033333	Neutral
3	2020-12-12	facts immutable senator even youre ethically s...	{'polarity': 0.1, 'subjectivity': 0.55, 'senti...	0.1	0.550000	Positive
4	2020-12-12	explain need vaccine pfizerbiontech	{'polarity': 0.0, 'subjectivity': 0.0, 'sentim...	0.0	0.000000	Neutral

Figure 5: Categorization of Dataset into sentiments

Then, classify the sentiments into a numeric form from 0 to 2 where 0 shows negative, 1 shows positive and 2 shows neutral sentiments shown in **Figure 6**.

	date	cleaned_text	results	polarity	subjectivity	sentiment	Labels
0	2020-12-20	folks said daikon paste could treat cytokine s...	{'polarity': 0.0, 'subjectivity': 0.0, 'sentim...	0.000000	0.000000	Neutral	2
1	2020-12-13	world wrong side history year hopefully bigges...	{'polarity': -0.5, 'subjectivity': 0.9, 'senti...	-0.500000	0.900000	Negative	0
2	2020-12-12	coronavirus sputnikv astrazeneca pfizerbiontec...	{'polarity': 0.0, 'subjectivity': 0.0333333333...	0.000000	0.033333	Neutral	2
3	2020-12-12	facts immutable senator even youre ethically s...	{'polarity': 0.1, 'subjectivity': 0.55, 'senti...	0.100000	0.550000	Positive	1
4	2020-12-12	explain need vaccine pfizerbiontech	{'polarity': 0.0, 'subjectivity': 0.0, 'sentim...	0.000000	0.000000	Neutral	2
5	2020-12-12	anyone useful advice guidance whether covid va...	{'polarity': 0.4, 'subjectivity': 0.25, 'senti...	0.400000	0.250000	Positive	1
6	2020-12-12	bit sad claim fame success vaccination patriot...	{'polarity': -0.1, 'subjectivity': 0.5, 'senti...	-0.100000	0.500000	Negative	0
7	2020-12-12	many bright days best bidenharris winning elec...	{'polarity': 0.675, 'subjectivity': 0.58749999...	0.675000	0.587500	Positive	1

Figure 6: Cleaned Dataset after labelling the sentiments

WordCloud is a python library that is used to represent the main words that appear more frequently, in a collage form. Positive, Negative, and neutral are three categories that occur in the dataset. To represent the most frequent words of these three categories, WordCloud plots images. Words that occur in Positive and Negative categories are shown in **Figures 7**.

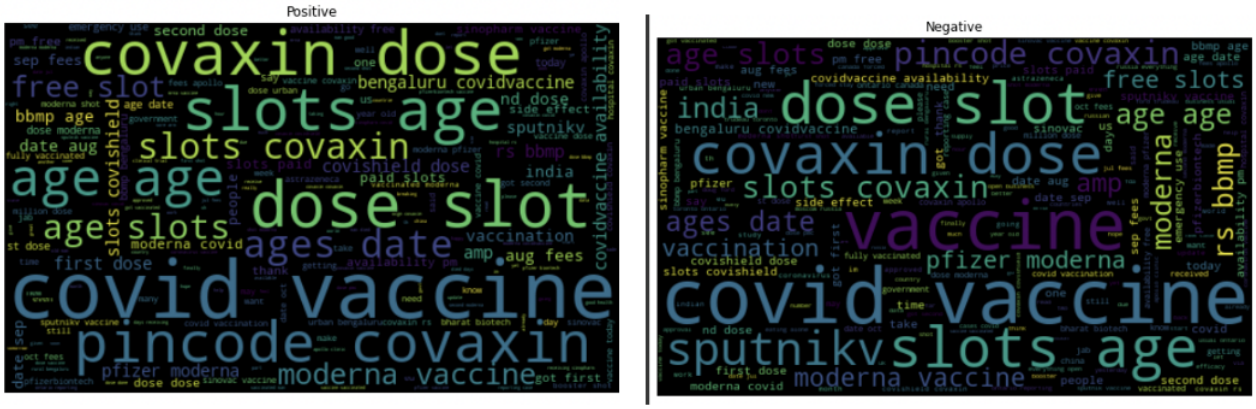

Figure 7: Positive and Negative Words

The data set is cleaned and prepared for implementing the machine learning and deep learning models. In machine learning, the stacking ensemble algorithm is the combination of the Decision tree, Random Forest Classifier, and XGBClassifier models. In Deep Learning, Multi-layer Perceptron classifier (MLP), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and long short-term memory (LSTM) are used. Then, the accuracy of each model was checked and compared.

Two terms occurred while implementing the models. Confusion Matrix and ROC curve defined as :
Confusion Matrix – In machine learning, the confusion matrix is a table (matrix) that may be used to measure the performance of the algorithms. It can be a binary or three-dimensional matrix. There are rows and columns in the confusion matrix, which represent the actual occurrences of the classes, and the expected ones.Accuracy can be defined as the sum of true positive and true negative that divide by the sum of true positive, true negative, false positive and false negative as shown in **Figure 8**.

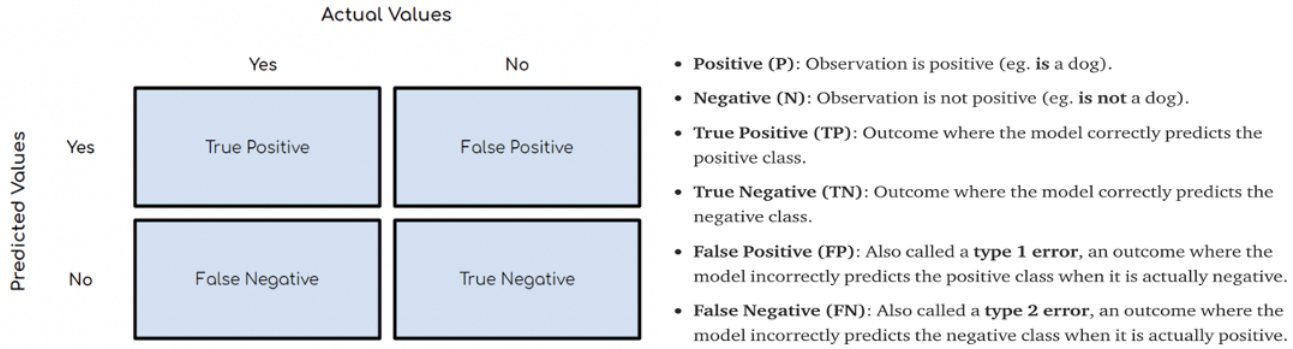

Figure 8: Confusion Matrix - Accuracy

For our study, this is helpful because neutral sentiments don't make sense to keep in the modelling and prediction. So, neutral sentiments are not considered in further analysis.

ROC curve – It is used to read the output of binary classification. Receiver Operating Characteristic is a plot of true positive rate (TPR) against false positive rate (FPR). TPR calculates as true positive divide by total positive and FPR calculates as false positive by total positive as shown in Figure 10. Confusion matrices at all threshold values are used to generate the ROC curve, which summarizes performance.

5.1 Implementation and evaluation of Machine Learning

This section explains the implementation of stacking ensemble of three different machine learning models - Random Forest Classifier, XGBClassifier, and Decision Tree Classifier to find the performance of different machine learning models.

Stacking Ensemble

Combining the predictions of numerous high-performing machine learning models is what stacking, a machine learning ensemble approach does. Classification and regression models may be assembled using a technique called stacking. Models that are used to predict outcomes on test datasets form the first layer. For example, a Meta-Classifier can take all the predictions from baseline models and construct a new prediction. In Python, the stacking ensemble is standardly implemented in the scikit-learn module. Following are the machine learning classification models in which Random Forest Classifier is a Meta-Classifier and others two XGBoost Classifier and Decision Tree Classifier are based classifiers.

• Random Forest Classifier

Random Forest is one of the most versatile and simple algorithms to implement. Multiple decision trees are combined to form a random forest, which is why the technique is called Random Forest. As the number of trees in the random forest classifier becomes up, the accuracy goes up as well.

• XGBoost Classifier

Distributed gradient boosting (XGBoost) is an efficient, adaptable, and portable library. The Gradient Boosting framework is used to develop machine learning algorithms. Ha-

doop, SGE, and MPI all execute the same code, and it can solve problems with billions of samples.

• **Decision Tree Classifier** A Decision Tree is a tree-structured classifier. Internal nodes contain attributes of a dataset; limbs depict decision rules, and leaf nodes indicate the conclusion. Leaf nodes represent the results of decisions, whereas Decision nodes are used to create decisions and include several branches.

Evaluation - Various matrices, such as Precision, recall, f1-score, and accuracy, are used to evaluate the model. The model's performance was assessed using a classification report and a confusion matrix. Using stacking ensemble, the classification report indicates an overall accuracy of 0.56. The values of precision, recall, and f1 scores are similar that is 0.56 and visualize the report in **Figure 9**. The model's accuracy is an indication of how well it worked.

	precision	recall	f1-score	support
Yes	0.56	0.56	0.56	4507
No	0.56	0.56	0.56	4545
accuracy			0.56	9052
macro avg	0.56	0.56	0.56	9052
weighted avg	0.56	0.56	0.56	9052

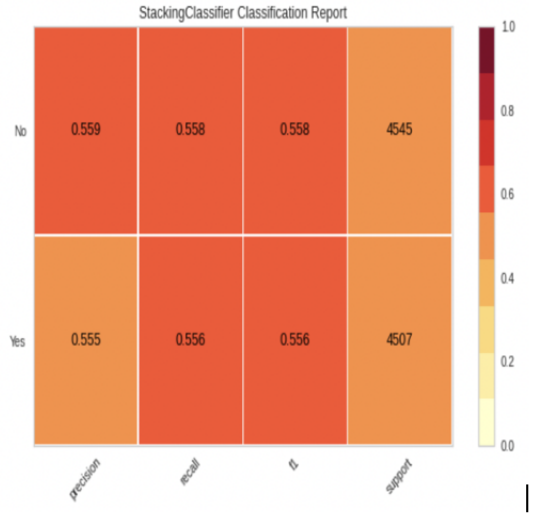

Figure 9: Classification report of Stacking Ensemble and Visualization

Accuracy = ((2535+ 2507) / (2535+ 2507+ 2000+ 2010))/100 = 0.5557 = 0.56

Figure 10 depicts the confusion matrix that shows the actual label against the predicted label. The Stacking Ensemble model correctly predicted 2507 as a true negative sentiment (TN) and 2535 as true positive sentiment (TP). It defines the total accuracy of this model as 56 percent.

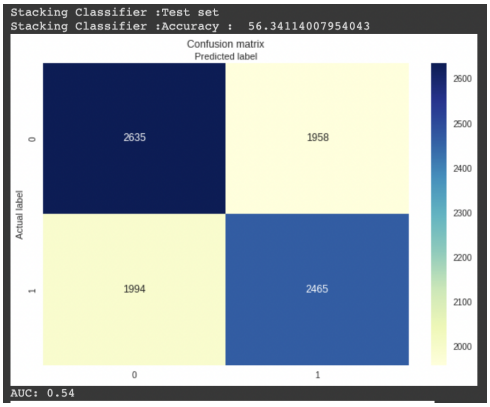

Figure 10: Confusion matrix of MLP

ROC curve as shown in **Figure 11**, depicts the true positive rate against the false-positive rate.

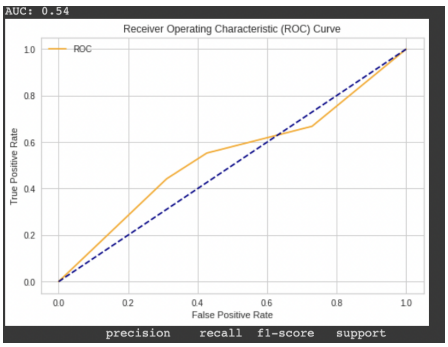

Figure 11: ROC curve of MLP

5.2 Implementation and Evaluation of Deep Learning Models

• **Recurrent Neural Networks (RNN)**
In artificial neural networks, a recurrent neural network (RNN) makes use of time series data. Many prominent apps like Siri, voice search, and Google Translate employ deep learning algorithms for ordinal or temporal issues, such as language translation and natural language processing (NLP), speech recognition, and picture captioning. Learning is accomplished similarly using training data to that used by feedforward and convolutional neural networks (CNNs). To implement this model, the confusion matrix and ROC curve are two findings that define the performance of the model.

Evaluation - ROC curve of training validation accuracy and training validation loss that depicts in **Figure 12**.

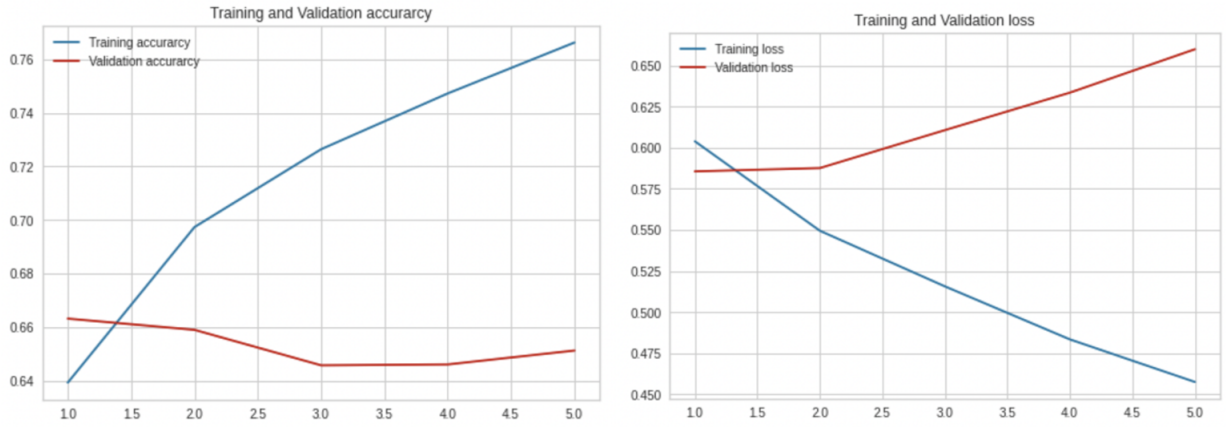

Figure 12: Training validation Accuracy and Loss

Confusion Matrix shows the accuracy of the RNN model by comparing actual labels with true labels. As per the calculations, the values of TP = 0.5418, TN = 0.7584, FP = 0.4582, and FN = 0.2416 as shown in **FIGURE 13**.
Accuracy = (0.5418 + 0.7584) / (0.5418 + 0.7584 + 0.4582 + 0.2416) = 0.6512.

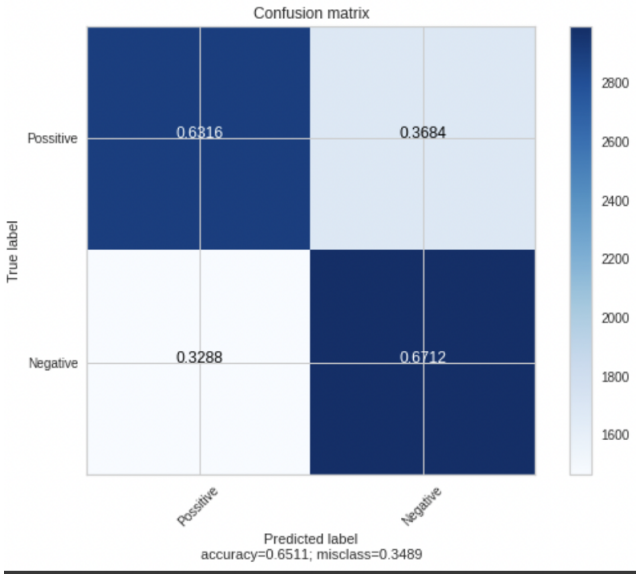

Figure 13: Confusion Matrix

It defines that this model gives the true positive sentiments 54 Percent and true negative sentiments 76 percent and the total accuracy of this model 65.12percent.

• Convolutional Neural Network (CNN)
When it comes to creating deep learning models, one of the most promising approaches is convolutional neural networks or CNN. In picture categorization and computer vision, it excels, as one illustration of how effectively it works. The convolutional layer of a CNN varies from that of other neural networks. To classify a picture, CNN scans the whole pixel matrix, inspecting every corner, vector, and pixel dimension. Using all the

matrix properties makes CNN more resilient to matrix data. Convolutional layers are a distinctive tool for CNN modelling since they include numerous features including edge detection, corner detection, and multiple texture detection. This layer can detect all the details in the picture matrix as it moves across it. Thus, the network's convolutional layers can recognize increasingly complex information as the layers become deeper. To implement this model, the confusion matrix and ROC curve are two findings that define the performance of the model.

Evaluation - ROC curve of training validation accuracy and training validation loss that depicts in **Figure 14**. The curve is not good as it is in linear form.

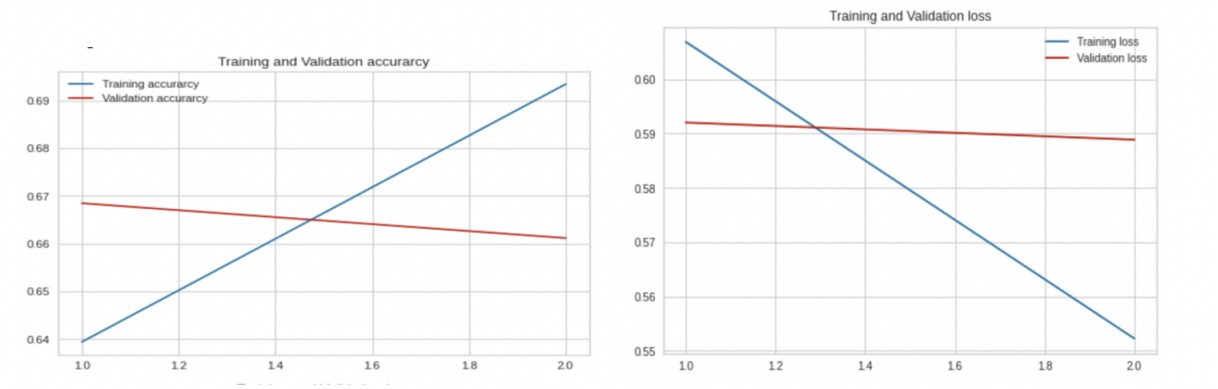

Figure 14: Training validation Accuracy and Loss

Confusion Matrix shows the accuracy of the CNN model by comparing actual labels with true labels. As per the calculations, the values of TP = 0.5418, TN = 0.7584, FP = 0.4582, and FN = 0.2416 and the average roc curve is shown in **Figure 15**.
Accuracy = (0.5360 + 0.7801) / (0.5360 + 0.7801+ 0.4640 + 0.2199) = 0.6593. It defines that this model gives the true positive sentiments 53percent and true negative sentiments 78percent and the total accuracy of this model 65.93percent.

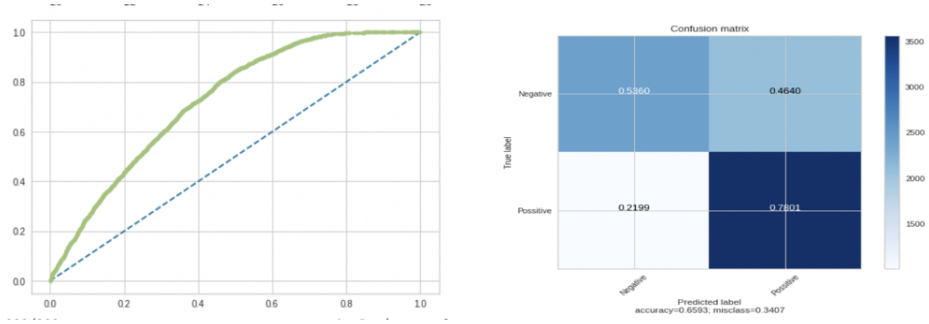

Figure 15: ROC Curve and Confusion Matrix

- **Long Short Term Memory Network (LSTM)**

For long-term memory, a recurrent neural network (also known as RNN) is utilized. LSTM architecture divides into three cells named gates with different function systems. Like a basic RNN, there is also a hidden state where H(t-1) is the previous timestamp's hidden state.

Evaluation - ROC curve of training validation accuracy and training validation loss that depicts in **Figures 16**. The curve is not good as it is in linear form.

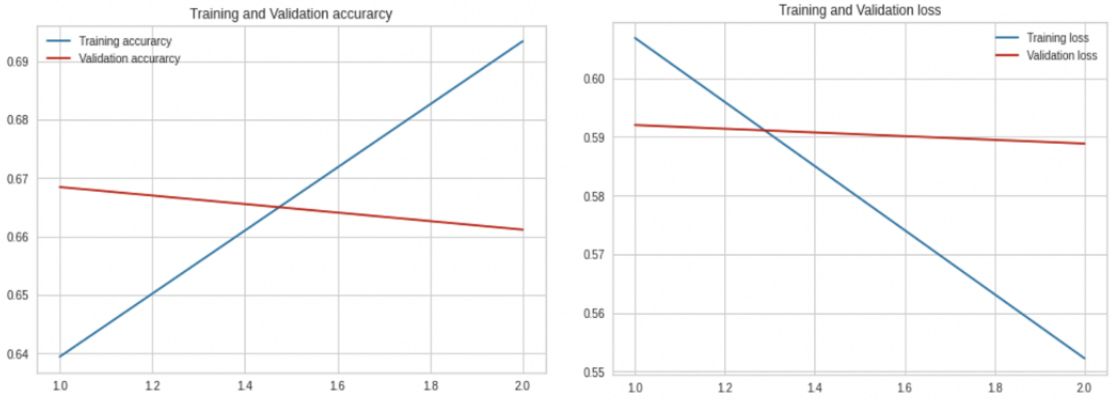

Figure 16: Training validation Accuracy and Loss

Confusion Matrix shows the accuracy of the CNN model by comparing actual labels with true labels. As per the calculations, the values of TP = 0.5418, TN = 0.7584, FP = 0.4582, and FN = 0.2416 and the average roc curve is shown in **Figure 17**.
Accuracy = (0.5968 + 0.7250) / (0.5968 + 0.7250+ 0.4032 + 0.2750) = 0.6612.

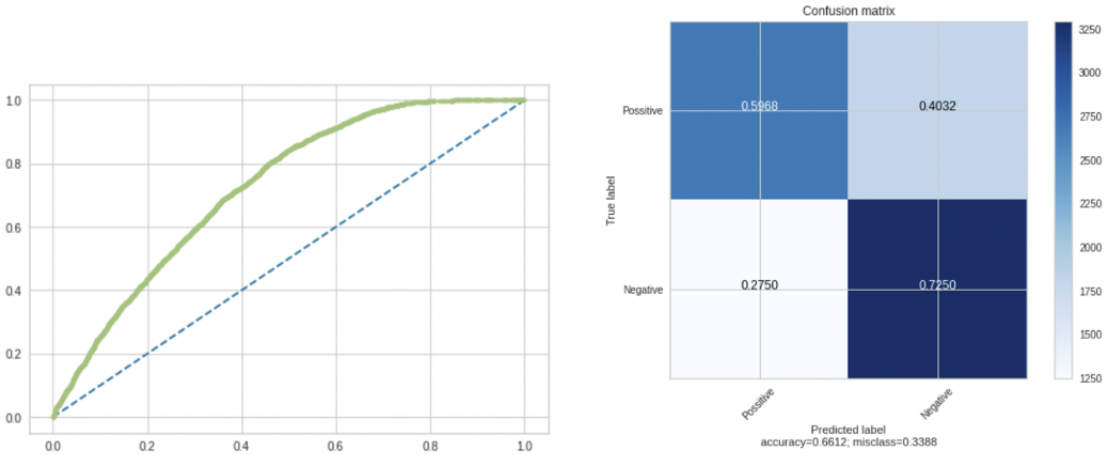

Figure 17: ROC Curve and Confusion Matrix

Figure 20 depicts the confusion matrix that shows the true label against the predicted label. The LSTM model correctly predicted 72Percent as a true negative sentiment (TN) and 59Percent as true positive sentiment (TP). The total accuracy of this model is

66.12percent.

- **Multi-layer Perceptron classifier (MLP)**

MLP Classifier uses a Neural Network to conduct classification, unlike other classification techniques like Support Vectors or Naive Bayes Classifier. MLP's implementation is no more difficult than other Scikit-Learn classifiers, such as support vectors and so on.

The model's performance was assessed using a classification report and a confusion matrix. Using MLPClassifier and Scikit-Learn classifiers, the classification report indicates an overall accuracy of 0.89. There are 0.51, 0.50, and 0.47 precision, recall, and f1 scores and visualize the report as shown in **Figure 18**. The model's accuracy is an indication of how well it worked.

	precision	recall	f1-score	support
Yes	0.51	0.25	0.33	4507
No	0.51	0.76	0.61	4545
accuracy			0.51	9052
macro avg	0.51	0.50	0.47	9052
weighted avg	0.51	0.51	0.47	9052

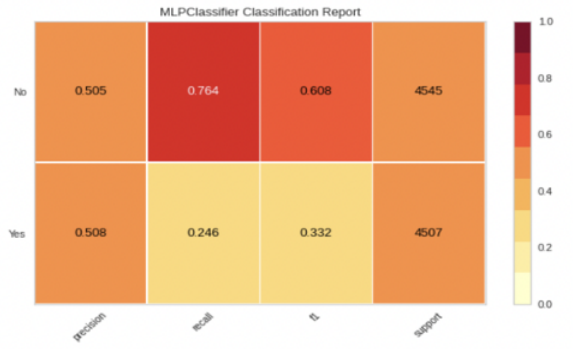

Figure 18: MLP Classification report and Visualization

Confusion Matrix shows the accuracy of the CNN model by comparing actual labels with true labels. As per the calculations, the values of TP = 0.5418, TN = 0.7584, FP = 0.4582, and FN = 0.2416 as shown in **Figure 19**.
Accuracy = ((1109 + 3471) / (1109 + 3471+ 1074 + 3398))/100 = 0.5057.
ROC curve as shown in **Figure 19**, depicts the true positive rate against the false-positive rate.

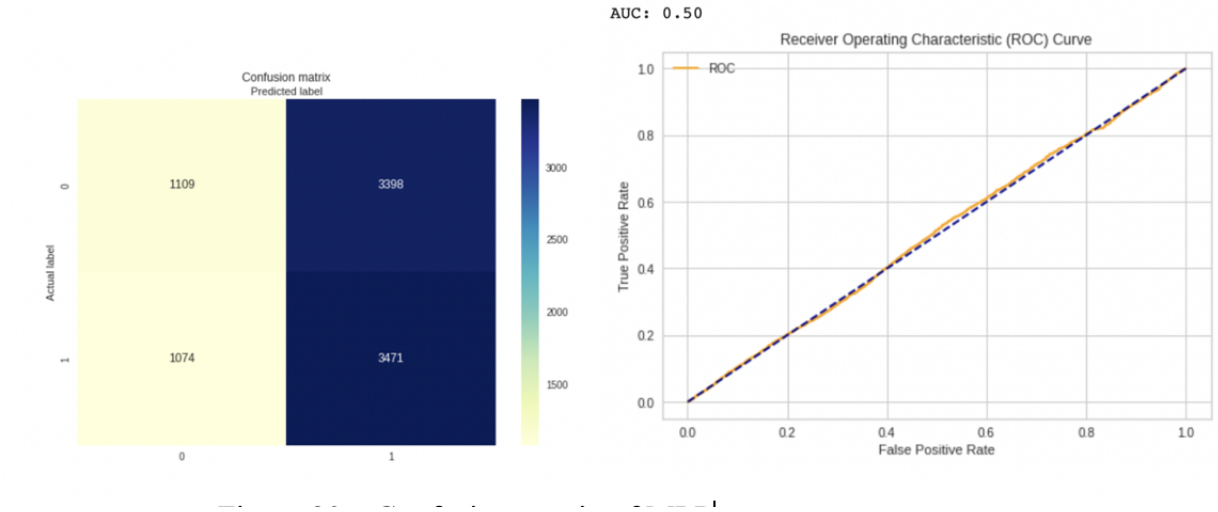

Figure 19: MLP Confusion Matrix and Visualization

So, this is all about how models are implemented and giving outputs based on the sentiments of the public on the Covid-19 vaccine topic over Twitter.

6 Discussion

The fundamental goal of the research was to develop models that accurately and efficiently execute sentiment analysis. Awareness of present limits and gaps was gained after conducting thorough literature research. A lexicon-based classification model or machine learning methods based on labelled data were utilized by most researchers. The unexplored area was the hybrid methodology of machine learning and deep learning models. In this study, the Stacking Ensemble method of three machine learning models – Decision tree classifier, random forest classifier, XGBoost classifier and RNN, CNN, MLP, and LSTM models of deep learning is used to classify the sentiments of people. In this study, the accuracy of a variety of methods is considered and compared as shown in **Figure 20**.Visualization of comparison is shown in **Figure 21**

Model	Accuracy %	Precision	Recall	F1 score	Misclass
Stacking Ensemble	56	56	56	56	-
Multi Layer Percepton (MLP)	51	51	25	33	-
Convolutional Neural Network (CNN)	66	-	-	-	34
Recurrent Neural Networks (RNN)	65	-	-	-	35
Long short-term memory (LSTM)	66	-	-	-	34

Figure 20: Comparison in table form

.

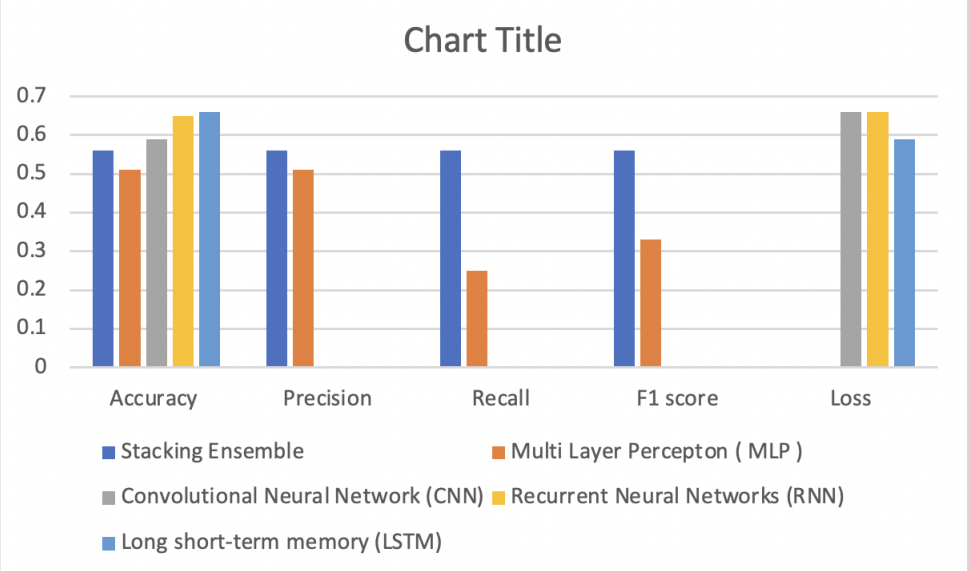

Figure 21: Comparison of Implemented Models

7 Conclusion and Future Work

The announcement of the Covid-19 vaccine made pure relaxation in the whole world. Today's era is modern and uses more social networking to show their attitudes towards everything. This research aims to analyze the sentiments of the public related to the Covid-19 vaccine over Twitter. To achieve this, data was collected from the Kaggle website and preprocessed using the natural language processing method. Sentiment Analysis is quite beneficial in research and methods. The Stacking Ensemble method of three machine learning classification models and Deep learning models were implemented. Convolutional Neural Network and Long short-term memory models produced the most accurate findings, with a rate of 66Percent. Compared to other models, two deep learning models have fared rather well. This experiment is on when the vaccine comes on the market. Now, everyone is getting vaccination in the whole world.

This study can be analyzed further to predict which vaccine is more effective on human bodies. This can be analyzed with the same methods but with a different dataset. In trends, people show their attitudes after getting vaccinated like how they feel, what problems they have faced, which vaccine one should prefer and so many. Collecting the related tweets and analyzing them makes one of future work. Moreover, Scientists and public health authorities are concerned about the latest version of the coronavirus Omicron because it has an extremely high number of mutations that make it more transmissible and less responsive to existing vaccinations. So, the experiments of the implemented methods can be used to predict the suspected cases of Omicron variant.

8 Acknowledgement

I'd want to express my gratitude to Prof. Vladimir Milosavljevic for clearing up all my questions and ensuring that I stayed on the right path during the supervision. It would have been impossible to carry out this study without their assistance. I also would like to

acknowledge the Department of M.Sc. Data Analytics to provide me with report writing-related information and the National College of Ireland for allowing me to show my skills through research.

References

Adamu, H., Bin Mat Jiran, M. J., Gan, K. H. and Samsudin, N.-H. (2021). Text analytics on twitter text-based public sentiment for covid-19 vaccine: A machine learning approach, *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pp. 1–6.

Adamu, H., Jiran, M. J. B. M., Gan, K. H. and Samsudin, N.-H. (2021). Text analytics on twitter text-based public sentiment for covid-19 vaccine: A machine learning approach, *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, IEEE, pp. 1–6.

Amjad, A., Qaiser, S., Anwar, A., Ali, R. et al. (2021). Analysing public sentiments regarding covid-19 vaccines: A sentiment analysis approach, *2021 IEEE International Smart Cities Conference (ISC2)*, pp. 1–7.

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R. and Hassanien, A. E. (2020). Sentiment analysis of covid-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media, *Applied Soft Computing* **97**: 106754.

Cotfas, L.-A., Delcea, C., Roxin, I., Ioanăş, C., Gherai, D. S. and Tajariol, F. (2021). The longest month: Analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement, *IEEE Access* **9**: 33203–33223.

Dumre, R., Mishra, S. and Dave, A. (2021). Twitter sentiment analysis on covid-19, *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, pp. 1–5.

Elbagir, S. and Yang, J. (2018). Sentiment analysis of twitter data using machine learning techniques and scikit-learn, *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, pp. 1–5.

Khan, R., Rustam, F., Kanwal, K., Mehmood, A. and Choi, G. S. (2021). Us based covid-19 tweets sentiment analysis using textblob and supervised machine learning algorithms, *2021 International Conference on Artificial Intelligence (ICAI)*, IEEE, pp. 1–8.

Monitoring the public opinion about the vaccination topic from tweets analysis (2019). *Expert Systems with Applications* **116**: 209–226.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417418305803*

Nurdeni, D. A., Budi, I. and Santoso, A. B. (2021). Sentiment analysis on covid19 vaccines in indonesia: From the perspective of sinovac and pfizer, *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, IEEE, pp. 122–127.

Paul, N. and Gokhale, S. S. (2020). Analysis and classification of vaccine dialogue in the coronavirus era, *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 3220–3227.

Roy, A. and Ojha, M. (2020). Twitter sentiment analysis using deep learning models, *2020 IEEE 17th India Council International Conference (INDICON)*, IEEE, pp. 1–6.

Saad, S. and Saberi, B. (2017). Sentiment analysis or opinion mining: A review, *International Journal on Advanced Science, Engineering and Information Technology* **7**: 1660.

Solangi, Y. A., Solangi, Z. A., Aarain, S., Abro, A., Mallah, G. A. and Shah, A. (2018). Review on natural language processing (nlp) and its toolkits for opinion mining and sentiment analysis, *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, IEEE, pp. 1–4.

Solanki, P. and Palwe, S. (2021). Understanding sentiments on corona vaccine using social media analysis, *2021 Smart Technologies, Communication and Robotics (STCR)*, IEEE, pp. 1–6.

Trupthi, M., Pabboju, S. and Narasimha, G. (2017). Sentiment analysis on twitter using streaming api, *2017 IEEE 7th International Advance Computing Conference (IACC)*, IEEE, pp. 915–919.

Wagh, R. and Punde, P. (2018). Survey on sentiment analysis using twitter dataset, *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, pp. 208–211.

Xiong, S., Lv, H., Zhao, W. and Ji, D. (2018). Towards twitter sentiment classification by multi-level sentiment-enriched word embeddings, *Neurocomputing* **275**: 2459–2466.

Yang, X. and Sornlertlamvanich, V. (2021). Public perception of covid-19 vaccine by tweet sentiment analysis, *2021 International Electronics Symposium (IES)*, pp. 151–155.