Analysing COVID-19 Vaccine-Related Tweets: Insights into Public Sentiments on Social Media Platforms

SARAVANAN KANMANI

Final Thesis Report

MARCH 2024

ACKNOWLEDGEMENTS

I am extremely grateful to my thesis supervisor, Dr. J V Ramana Reddy, for his helpful guidance and support throughout my research. His expertise, patience, and encouragement have been instrumental in shaping my work. I would also like to thank upGrad and LJMU academic excellence, for their insightful comments and suggestions. Their feedback has helped me to refine my ideas and improve the quality of my research. Lastly, I am grateful to my family and friends for their solid support and encouragement. Their love and praise have been a constant source of inspiration throughout my academic journey.

ABSTRACT

The COVID-19 epidemic has been a worldwide health disaster that has affected millions of individuals leading to many deaths. While vaccines have been developed to prevent the spread of virus, there has been some hesitancy among people to accept them. In this research, the aim is to analyze COVID-19 vaccine-related tweets to gain insights into the common sentiments shared by people on social media platforms like X (formerly Twitter). Openly available dataset that contains tweets related to COVID-19 vaccine used for this research. Lexicon-based and deep learning methods, BERT and Bi-LSTM with BERT designed to analyze the tweets. A lexicon approach, VADER sentiment tool used to assign sentiments (i.e., positive, negative, or neutral) for the cleaned tweets. The common keywords used in the tweets in each sentiment are visualized using WordClouds. Bootstrap resampling with sample replacement has been performed to solve the class imbalance and resource requirements. The prediction of sentiments was performed using a deep learning methods BERT and Bi-LSTM with BERT word embeddings. The performance of the models evaluated using accuracy, recall, precision and F1-score. BERT model achieved higher accuracy of 95% and Bi-LSTM with BERT have decent loss trend. The findings of this research could be helpful for public organizations and health authorities to be better prepared for future outbreaks similar to COVID-19.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI…….. Artificial Intelligence
API…..... Application Programming Interface
BBC……... British Broadcasting Corporation
BERT……..... Bidirectional Encoder Representations from Transformers
Bi-LSTM……. Bidirectional Long Short-Term Memory
BiGRU……. Bidirectional Gated Recurrent Unit
CNN……... Convolutional Neural Network
COVID-19…… Corona Virus Disease of 2019
DL…….... Deep Learning
DT…….... Decision Tree
RF…….... Random Forest
KNN……... K-Nearest Neighbor
GPU……. Graphics Processing Unit
LR……. Logistic Regression
LSTM……... Long Short-Term Memory
ML…… Machine Learning
NB…… Naïve Bayes
NLP……. Natural Language Processing
NLTK……... Natural Language Toolkit
RAM……. Random Access Memory
RMDL……... Random Multimodal Deep Learning
SARS-CoV2… Severe Acute Respiratory Syndrome–related Coronavirus
SVM……. Support Vector Machine
TF-IDF……. Term Frequency–Inverse Document Frequency
TSA…… Twitter Sentiment Analysis
VADER…… Valence Aware Dictionary and sEntiment Reasoner
WHO…… World Health Organization
X…….. Twitter platform

CHAPTER 1

INTRODUCTION

1.1 Background of the study

An infectious disease caused by the SARS-CoV2 (Severe Acute Respiratory Syndrome related Coronavirus) virus is COVID-19 (Coronavirus 2019). The exact origin of the virus is still unknown but it was first identified in Wuhan, China, in December 2019 and spread to the entire globe (Kumar et al., 2021). Transmission and Life-Cycle of COVID-19 is shown in *Figure 1.1* (Funk et al., 2020). The number of people infected with and dying from COVID-19 was increasing rapidly every day. On March 11, 2020, the World Health Organization (WHO) declared the COVID-19 outbreak a pandemic (WHO media briefing on COVID-19, 2024). The COVID-19 pandemic has led to a dramatic loss of human life worldwide and presents a unique challenge to public health, food systems and the world of work. This pandemic is one of the most significant crises of modern times, and its impact on the world is unprecedented (Kaur et al., 2021). The virus can mainly spread through an infected persons when they cough, sneeze, talk, sing or breathe. As of March 13, 2024, over 774 million of people affected by COVID-19 and more than 7 million deaths caused as mentioned in the WHO Dashboard (WHO - COVID-19 cases reported, 2024). The pandemic has had devastating medical, economic, and social consequences. Therefore, it is crucial to develop and distribute safe and effective preventive vaccines as soon as possible. While mask wearing and social distancing have been proven effective in reducing the spread of COVID-19, the development and widespread adoption of a preventive vaccine will be crucial for long-term control of the pandemic (Chou and Budenz, 2020). Though different countries followed various measures, the best way to get rid of COVID-19 disease is vaccination (Umair et al., 2021; Umair and Masciari, 2023).

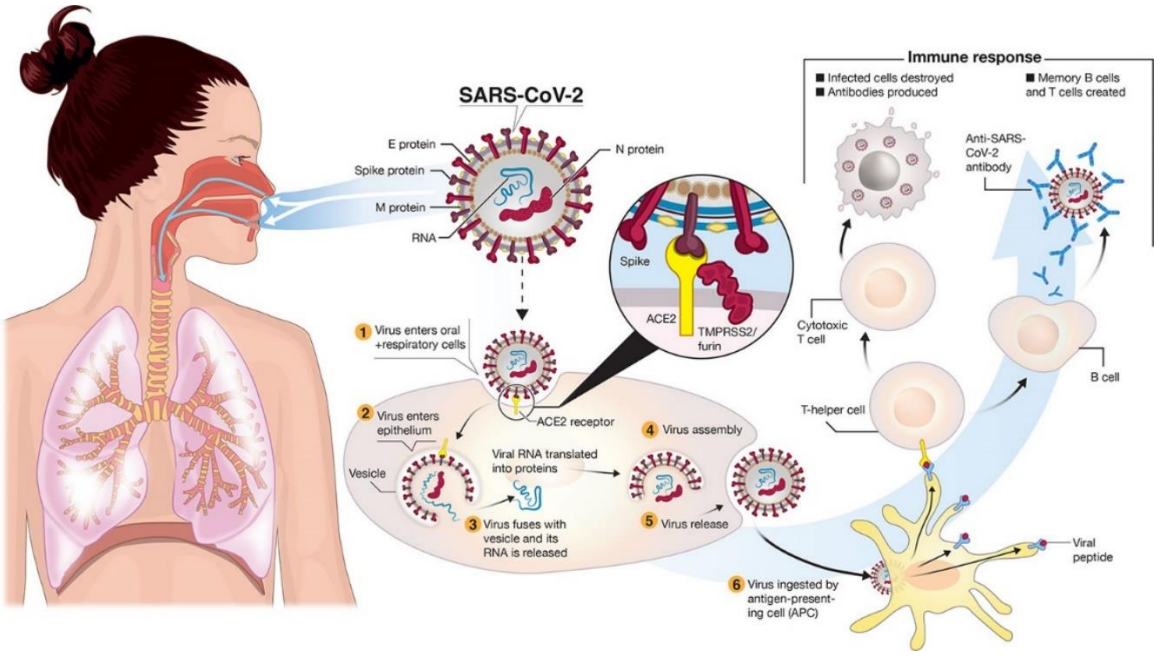

Figure 1.1. Transmission and Life-Cycle of COVID-19

Many pharmaceutical, research institutes, and government put their efforts to find COVID-19 vaccines. The two-dose regimen of the BNT162b2 mRNA vaccine (Pfizer/BioNTech) was 95% efficacious at preventing COVID-19 at 2 months with no increased risk of serious adverse events. The vaccine had a favorable safety profile and was highly efficacious in preventing COVID-19 with up to 6 months of follow-up (Polack et al., 2020). After the introduction of vaccine, people hesitated to accept the vaccine because of the fear of its rumor and/or side-effects (Green et al., 2021; Umair and Masciari, 2023). The success of vaccine campaigns to control COVID-19 is not solely dependent on vaccine efficacy and safety but also willingness of the general public and healthcare workers to receive the vaccine plays a crucial role. Low rates of COVID-19 vaccine acceptance were reported in the Middle East, Russia, Africa, and several European countries. This could pose a significant challenge to global efforts aimed at controlling the COVID-19 pandemic (Sallam, 2021). There are several reasons why people may be hesitant to accept the vaccine, and some of these reasons for vaccine hesitancy among medical and health science students attending Wolkite University in Ethiopia are listed in *Figure 1.2* (Mose et al., 2022). Therefore, it is essential for government and public health authorities to understand the people thoughts about COVID-19 vaccines to plan it better now and in similar pandemic situation in future.

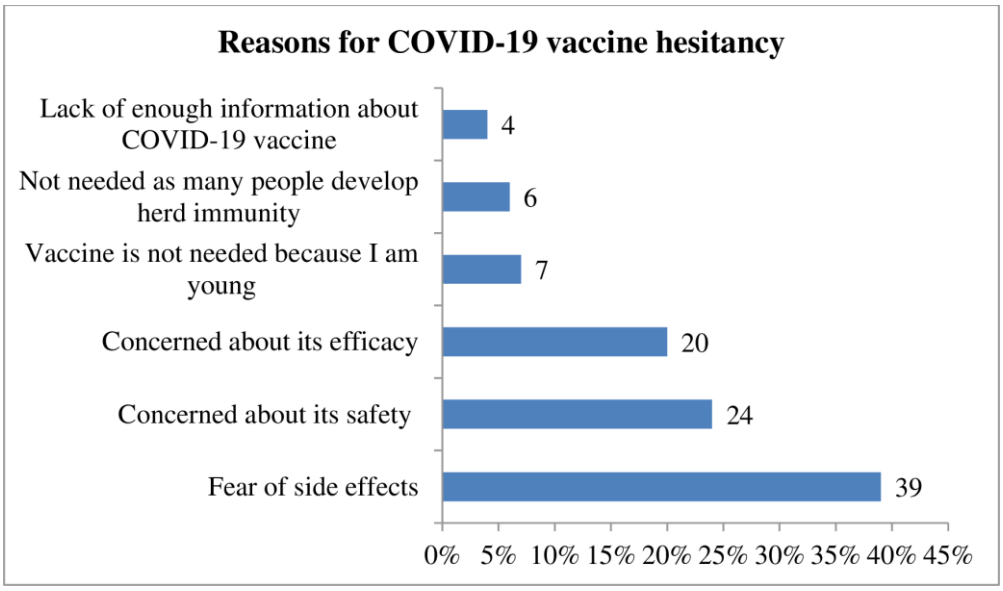

Figure 1.2. Vaccine hesitance reason among students in Ethiopia

Social media analysis is a valuable resource for public health experts and government officials. It enables them to track the population's emotional fluctuations and provide real-time, up-to-date information on public awareness and response to crisis situations, thus enhancing the operational response to the crisis (Power et al., 2014). People express their thoughts in social media platform like X (formerly known as Twitter) as it is publicly accessible by everyone. The tweets in social media can be analyzed to understand the people's reaction on any recent topic (Luo and Xu, 2021). X platform is a social networking platform that allows users to post their opinions or comments as tweets. During the COVID-19 pandemic era, lot of people shared their sentiments about the virus and its vaccines in X platform. *Table 1.1* displays a few examples of tweets related to COVID-19 vaccines. Sentiment analysis is a popular method to analyze the text and to find the polarization (i.e., positive, negative or neutral) of the tweets (Raheja and Asthana, 2021).

Table 1.1. Examples of COVID-19 vaccine tweets

Covid-19 vaccine-related tweets
I pucking do not intend to get vaccinated. Wht if my employer force to get vaccinated?
IT HAPPENED!!!!!!!! I got my first dose of the #PfizerBioNTech vaccine today – THE BEST HOLIDAY PRESENT EVER!!!
#BREAKING #PfizerBioNTech #COVID19 #vaccine provokes serious allergic reaction in #USA

#PfizerBioNTech Just wondering, has every Pfizer employee all the way to the boardroom been vaccinated yet?
Happy Cold Chain Day! The only way to get a vaccine distributed across the county #coldchain #PfizerBioNTech #industrial #cre

This study is to analyze public feelings and sentiment about COVID-19 vaccines which they have expressed in social media platform X. The purpose is to help the government organization and health authorities to consider the sentiments of the public and consider while they design vaccine policy in such pandemic situations. A dataset from Kaggle website will be used for this research and text will be pre-processed for a suitable analysis. Two sentiment analysis approaches Lexicon-based and deep learning are expected in this research. Sentiments of the tweets will be classified using Lexicon based VADER tool whereas the sentiment prediction will be performed using two deep learning models BERT and Bi-LSTM with BERT. The performance of the model will be evaluated using appropriate metrics. A visualization to get the common keywords used on each sentiment will be plotted using WordCloud.

1.2 Problem Statement

COVID-19 has been severely affected the life of people around the world as more and more people were getting infected everyday with COVID-19 and dying from it at an alarming rate. A record of the WHO's global-scale actions and response to COVID-19 disease over time shown in *Figure 1.3* (Voidarou et al., 2023). Vaccination, wearing masks, social distancing, and personal hygiene have been crucial in controlling the spread of COVID-19. After the roll out of vaccines for the virus COVID-19, public shared their opinions regarding the vaccines in social media like X platform. Anti-vaccination movements have leveraged social media to influence people and reduce vaccine acceptance rates, which has delayed efforts to prevent or reduce the spread of the coronavirus pandemic (Chiou and Tucker, 2018). Many researchers analyzed the sentiments of tweets regarding COVID-19 vaccines to understand the public thoughts. (Albahli and Nawaz, 2023) presented a novel DL approach "TSM-CV" for sentiment analysis of tweets regarding coronavirus vaccines to know the human behavior. The authors used both past and real-time data from X platform for their analysis. They proposed a RMDL classifier for their sentiment prediction and concluded that analyzing COVID-19 vaccine sentiments can support the health authorities to take actions to get rid of associations that are against vaccination. People's opinion about COVID-19 vaccination process also

important research (Said et al., 2023) that indicates the individuals have neutral sentiments about vaccines. In this study, the authors collected the tweets from X platform for a month and applied an ensemble deep learning model LSTM-2BiGRU that outperformed other classical models like LR, NB, DT, RF and KNN. Another study from (Umair and Masciari, 2023) to understand the public feelings and sentiments showed that sentiment and spatial analysis helps in identifying the people's attitude towards vaccines. The authors also visualized COVID-19 vaccines tweet data geo-graphically and implemented several geo-spatial methods like hotspot analysis, kernel density estimation. The researchers missed the opportunity to compare the BERT model used in this study with other sentiment analysis techniques.

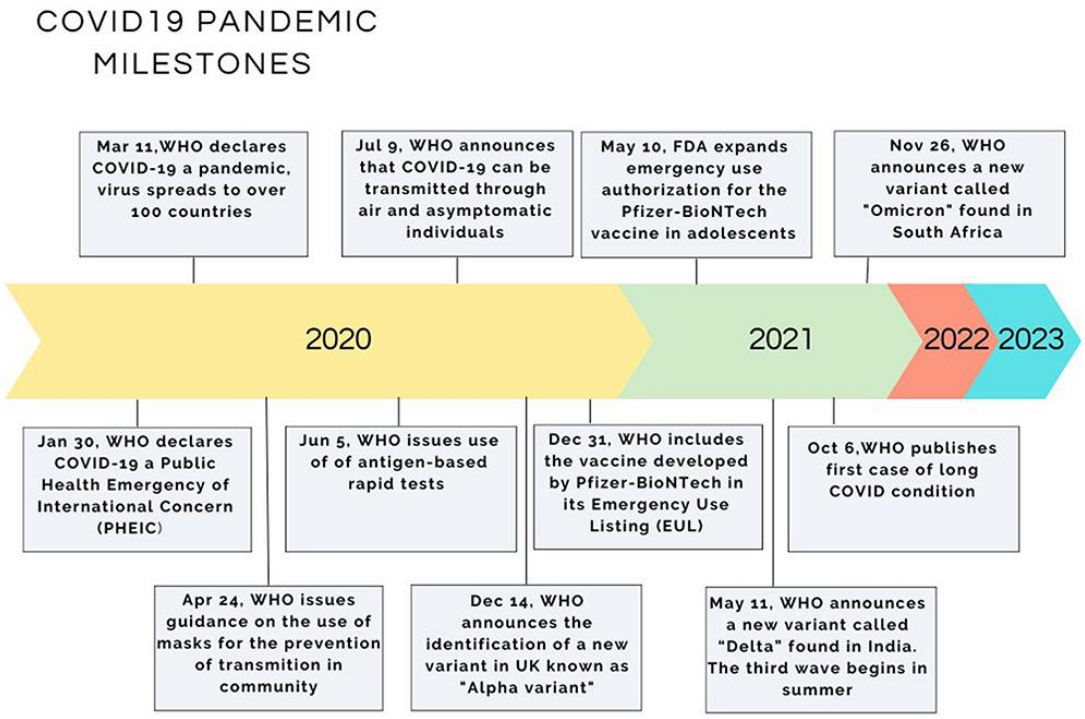

Figure 1.3. Timeline of COVID-19 from WHO

Public feelings and sentiments of COVID-19 vaccination changed over time in response to different BBC news reports was researched in the study of (Amujo et al., 2023). This research is based on three UK based vaccines that are Pfizer-BioNTech, Moderna, and Oxford AstraZeneca. The authors used six period of data based on the BBC news reports of development of COVID-19 vaccines and implemented BERT model to analyze the public views. In this research the authors found that Moderna vaccine had more positive sentiments

compared to other two vaccines and AstraZeneca had more negative sentiment. The paper also found that BBC news report had a significant impact on the public sentiment. The study didn't look to address the impact of misinformation on vaccine sentiment. Vaccine hesitancy is the key concern and various researchers used ML techniques to examine the vaccine uncertainty based on tweets from X platform. To understand the people views before and after vaccination, the authors (Qorib et al., 2023a) used NRCLexicon technique to label tweets into ten different classes and the significance of the associations among the basic emotions were checked using t-test. This study showed that public feelings turned gradually to positive about COVID-19 vaccination. Same set of authors (Qorib et al., 2023b) conducted another research to investigate COVID-19 vaccine hesitation through examining three sentiment calculation methods (Azure Machine Learning, VADER, and TextBlob). They found TextBlob with TF-IDF and LinearSVC classification model performed well and the combination of CountVectorizer and TF-IDF decreases the model accuracy. This study as well concluded that hesitancy in Covid-19 vaccine progressively decreased over a period. Despite the vaccine drives, there was an increasing in vaccine hesitancy due to misinformation regarding the vaccines were spread on social media either via humans or bots. Therefore, the authors (Hayawi et al., 2022) researched about detecting misinformation in tweets related to vaccination. The authors collected the tweets from X and misinformation was manually read and labelled with the help of consistent sources and health specialists. The proposed BERT model performed well on the test set and the model is effective in finding misrepresentation regarding COVID -19 vaccines on public network.

During COVID-19 period, people shared lot of posts and reviews in X about the disease and its vaccines. These posts and tweets were examined in the study of (Kathiravan et al., 2023) to understand the psychological and emotional impacts of people. In this paper, the authors researched the state-of-art automatic extraction of emotions from public tweets related to coronavirus and it provides public mental health insights about COVID-19 for the health authorities. Deep Learning techniques that competently work on unstructured data was not considered in this study. TSA (Twitter Sentiment Analysis) being a popular topic, the authors (Aslan et al., 2023) proposed a novel approach using CNN optimized via arithmetic optimization algorithm (TSA-CNN-AOA) to understand people's views on COVID-19 epidemic. This study concluded that the approach they proposed is successful for TSA that

can help to minimize and eliminate the impact of disease. Tweets by Indian citizens related to COVID-19 pandemic and vaccine drive was analyzed to help the policy makers and health workers to obtain the insights and make right decisions for similar pandemic outbreaks (Ainapure et al., 2023).

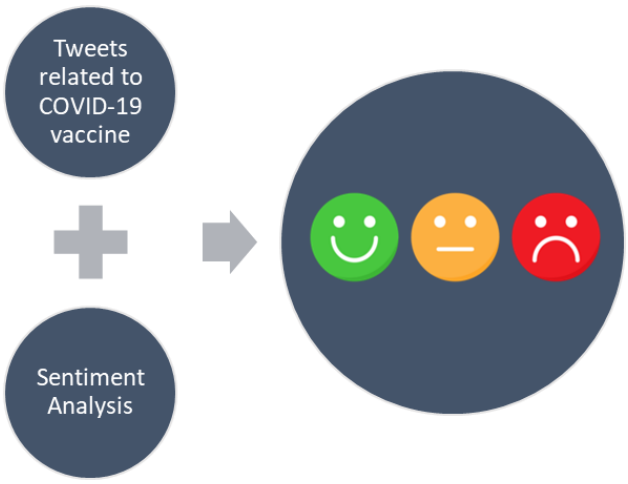

Figure 1.4. Sentiments of COVID-19 vaccine-related tweets

Based on the research analysis conducted, there are few gaps as most of the researchers studied the sentiment of COVID-19 tweets using either lexicon or deep learning and with limited number of tweets. The objective of this study is to identify the sentiments (i.e. positive, negative or neutral) from the COVID-19 vaccine related tweets as shown in *Figure 1.4*. This research use lexicon-based, deep learning and hybrid models for sentiment prediction that might improve the model's performance. The study can be useful for various stakeholders, including government policymakers, non-profit organizations, and profit organizations. The results of the study can help policymakers understand how people are reacting to the COVID-19 vaccines and frame their rules. Profit organizations can use the information to analyze various sentiments and start the production of essential items, thereby making profits. NGOs can use the related facts and information to decide their strategy for training people. The study will follow the standard operating procedure for sentiment analysis, which includes exploratory data analysis, data preprocessing, classification and prediction models, and evaluation.

1.3 Aim and Objectives

The primary aim of this research is to study public sentiments related to COVID-19 vaccines using tweets from X platform. The research use lexicon, deep learning and hybrid models to classify and predict the sentiment that find the common sentiment across the tweets. The purpose of this research is to benefit the public organization, health authorities and policymakers to get insights and to consider the public opinion while introducing their vaccination policy for similar upcoming pandemic outbreak. Profit organizations can analyze various sentiments and use the information to produce essential items, thereby generating revenue. NGOs can leverage the related facts and information to strategize their efforts towards educating people.

The research objectives are framed based on the aim of this study which are as follows:

- To analyze the COVID-19 vaccine-related tweets posted by public and find the most common sentiments expressed.
- To find the recurrent keywords for each sentiment discussed in COVID-19 vaccine related tweets.
- To classify the sentiments using suitable lexicon, deep learning and hybrid model.
- To evaluate the performance of the classification models developed.

1.4 Scope of the Study

This research explores public sentiments towards COVID-19 vaccine related tweets using lexicon based, deep learning and hybrid techniques. The study can provide insights to policymakers about COVID-19 vaccines. It also delivers understanding about people's sentiments so the healthcare companies can create vaccines and essential items that cater to the needs of the public.

Sentiment analysis will be performed only on historical COVID-19 vaccine associated tweets data which is publicly available and no real time data will be used in this study. Also, this study will be conducted only on tweets text data that were posted in English language. The study will be conducted in open sources software and models.

To make this research achievable within given timeframe the scopes are defined. This process also supports in identifying the intended audience and the methods employed for sentiment analysis.

1.5 Significance of the Study

Sentiment analysis of the tweets about coronavirus vaccine can provide the public thoughts about COVID-19 disease and its vaccines. By understanding the public opinions and attitudes, the government organization and health authorities can be prepared for similar pandemic situation on how to approach the people for vaccination. Also, the study will be helpful for future researchers who can fine tune the designed model for better performance.

1.6 Structure of the Study

This study focuses to analyze public sentiment towards COVID-19 vaccines by analyzing tweets from a specific platform. The existing literatures will be studied to understand the problems, research gaps and the methods to overcome the problem. An appropriate methodology will be used to tackle the problem. The study uses publicly available tweets related to COVID-19 vaccines to gather sentiments using lexical, deep learning and hybrid methods. Lexicon method VADER will be used to find the sentiment based on the intensity scores of the tweets. The study also presents a deep learning model BERT and a hybrid model Bi-LSTM with BERT for sentiment classification. The models will be evaluated using common metrics like accuracy, precision, recall, and F1-score. The results of the models will be discussed and conclusion will be made on the study results. The study is motivated by the fact that people around the world are still hesitant to get vaccinated, which is impacting the vaccination program and causing the virus to continue to spread.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The emergence of COVID-19 pandemic in the end of 2019 demanded immediate responses around the globe including the development and distribution of vaccines. The acceptance of vaccines among the public was another challenge after their expansion. Despite the increasing vaccination rate, a significant number of people have raised doubts about the safety of vaccines (Wang and Chen, 2023). Several groups that are against vaccines often dispute the effectiveness of vaccines in fighting this infectious disease by spreading conspiracy theories and myths (Qorib et al., 2023b). So, understanding the public sentiment became crucial as the vaccination campaigns accelerated. Social media platforms, especially X platform, emerged as valuable sources of opinions shared as tweets or comments. Sentiment analysis of tweets provides a valuable research opportunity, as millions of X platform users share their messages, ideas, opinions, feelings, understanding, and beliefs through X platform posts (Samuel et al., 2020). Sentiment analysis involves extracting emotional tones (positive, negative, or neutral) from textual data and many researchers measured sentiments of public using sentiment analysis techniques on various topics. Sentiment analysis of COVID-19 vaccine-related tweets provides insights into public perceptions, concerns, and attitudes toward the vaccines. Several studies have demonstrated the significance and usefulness of scrutinizing social media to comprehend the public's attitude and discourse regarding COVID-19 vaccination. Although previous research has identified gaps, sentiment analysis can be conducted using both lexicon-based and machine learning approaches to determine the most operative model for predicting public view around COVID-19 vaccine.

This chapter, Literature Review, aims to explore existing studies on sentiment analysis of COVID-19 vaccine-related tweets. This review helps to understand the public sentiment around COVID-19 vaccines and informs communication strategies for vaccine promotion in the pandemic situation. This also demonstrate the different methods used to classify and predict the sentiments.

2.2 COVID-19 pandemic and the vaccines

The COVID-19 pandemic emerged in December 2019 in Wuhan, China, caused by the novel virus SARS-CoV-2. Since then, it has rapidly spread worldwide, infected over 112.20 million people and resulting in approximately 2.49 million deaths as of February 26, 2021. The progression of COVID-19 and its management can be summarized by examining four key aspects as shown in *Figure 2.1* (Kumar et al., 2021). Across the globe, lives were lost due to the COVID-19 pandemic as the infection and mortality rates caused by the coronavirus were steadily increased (Kaur et al., 2021). To curb the transmission and impact of the disease, it was advised to use face masks, practice hand sanitation, and maintain social distancing. The effectiveness of wearing masks in reducing COVID-19 transmission has been supported by the finding (Kwon et al., 2021), even in situations where social distancing is inadequate. Though the preventive measures helped in controlling the disease spread but the long-term control relied on the development and widespread adoption of preventive vaccine (Chou and Budenz, 2020). The COVID-19 pandemic has had terrible medical, economic, and social consequences. Therefore, it is imperative to develop safe and effective precautionary vaccines to contain the pandemic.

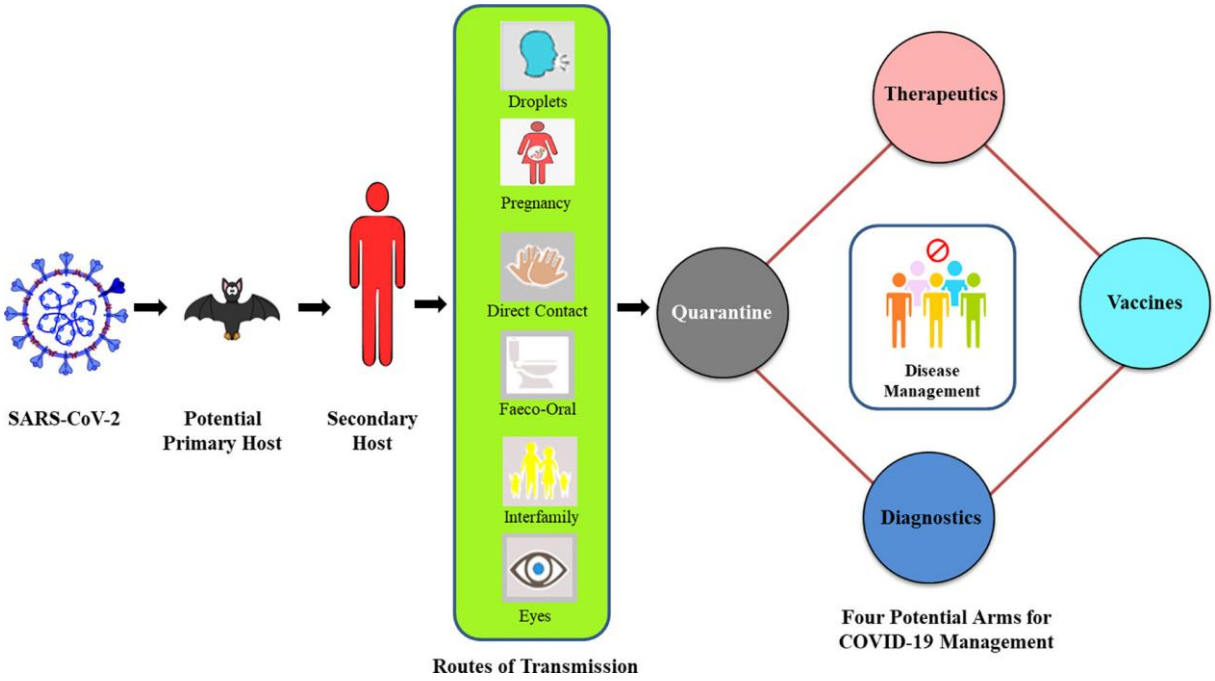

Figure 2.1. COVID-19 progression and four key aspects

The rise of safe and effective vaccines has emerged as a promising solution to enhance population immunity and effectively manage disease outbreaks (Polack et al., 2020). Vaccination stands out as one of the most potent and economically efficient public health measures (Sallam, 2021). Consequently, several SARS-CoV-2 vaccines have been developed and globally approved, including Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin and Sputnik V (Rani and Jain, 2023). According to a study (Levin et al., 2021), the humoral response to the BNT162b2 COVID-19 vaccine was found to be substantially decreased six months after the second dose. The decrease was more pronounced among men, persons aged 65 years or older, and persons with immunosuppression. Though there are several controls and introduction of new vaccines to limit the spread of COVID-19, the pandemic is still ongoing due to the emergence of new variants of the virus that are more contagious and can evade the immune system (Gong et al., 2023). Thus, encouraging vaccination remains crucial in safeguarding individuals from severe illness. As of March 13, 2024, the World Health Organization has administrated a total of 13.59 billion vaccine doses worldwide (WHO - COVID-19 vaccination, World data, 2024).

2.3 Sentiment Analysis

The expression of emotions is an essential component of human communication. Sentiment analysis (SA), also referred to as opinion mining, is an NLP technique used to find out the emotional tone conveyed by a piece of text. According to research, the opinions of stakeholders have a greater impact on decision-making than facts, for both individuals and communities such as governments and organizations (Dhanalakshmi et al., 2016). Nasukawa was the first to propose the idea of sentiment analysis (Nasukawa and Yi, 2003). Sentiment Analysis is the computational examination of individual's emotions and viewpoints related to a specific study. It spans across various domains, drawing insights from psychology, sociology, natural language processing and machine learning. In recent times, the surge in data volume and computational capabilities has facilitated advanced analytical approaches. Machine learning has emerged as the preeminent tool for sentiment analysis (Ligthart et al., 2021). The text’s sentiments or opinions are classified into positive, negative, or neutral categories. The fundamental tasks in sentiment analysis encompass classifying sentiment at different levels (Yadav and Vishwakarma, 2020) as mentioned shown in *Figure 2.2* (Birjali et al., 2021) and the details below.

1. Document level: Determining the overall sentiment of an entire document.
2. Sentence-level: Analyzing sentiment within individual sentences.
3. Aspect-level: Focusing on sentiment related to specific aspects or features.

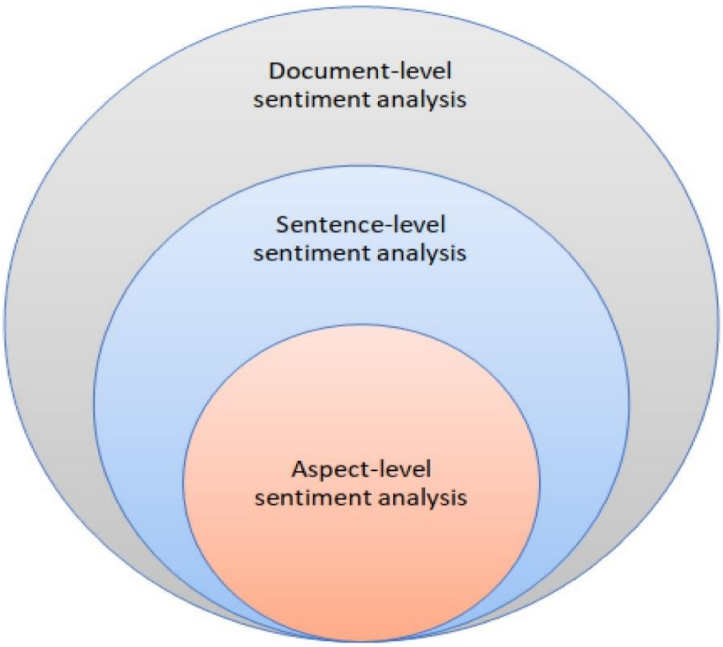

Figure 2.2. Levels of sentiment

The importance of sentiment analysis is increasing to study the growing opinions on social media and other sites at a faster pace. There are three main techniques in sentiment analysis as shown in *Figure 2.3* and they are Lexicon-based methods, Machine learning based and hybrid methods (Aqlan et al., 2019).

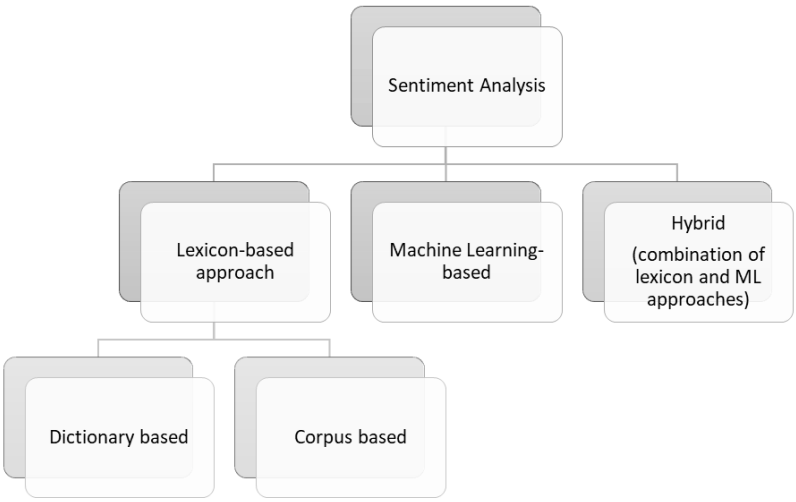

Figure 2.3. Techniques in Sentiment Analysis

2.3.1 Lexicon-based approach

The lexicon-based methods rely on sentiment lexicons, which serve as crucial indicators of sentiment and are often referred to as opinion words. Lexicons can be categorized into dictionary-based: sentiment words are manually collected using specific guidelines and corpus-based: specific presentations of sentiment are identified by finding sensitive words within a large corpus (Thangavel and Lourdusamy, 2023).

2.3.2 Deep Learning approach

Deep Learning draws inspiration from the architecture and operation of the human brain to enhance machine learning techniques. Handcrafted features are used in traditional approaches, such as lexicon-based approaches, which can be a time-consuming and tedious process. Additionally, these approaches may not generalize well to other domains or areas. Deep learning reduces the burden of feature design by automatically creating the required features for the classification process as the network learns (Ligthart et al., 2021).

2.3.3 Hybrid approach

Hybrid model combines rule-based and ML techniques or combining multiple deep learning models that helps to overcome limitations of individual methods. Hybrid techniques have demonstrated their potential as models for mitigating sentiment errors on training data that is becoming increasingly complex (Dang et al., 2021).

2.4 COVID-19 vaccine sentiment on X platform

The rise of social networking platforms has become an essential aspect of daily existence for people worldwide. Social media plays a crucial role in ensuring successful risk communication during pandemics, especially with the rapid spread of mobile phones and internet technology in recent times (Rahmanti et al., 2021). Through social media, individuals openly share their thoughts revealing their emotional reactions and broadcasting aspects of their lives. Social media can have a significant impact on shaping public perception during health events, such as global pandemics (Malecki et al., 2021). Over recent years, analyzing X platform data has gained popularity as a method for assessing people's perceptions and understanding their social network connections (De Rosis et al., 2021). During the widespread

COVID-19 pandemic, people used X platform a lot to share their feelings and thoughts so it was filled with covid related sentiments. Thus, analyzing tweets collected from a specific group of people can help identify patterns in their behavior (Kaur et al., 2020).

As part of a randomized control trial carried out in the United Kingdom and United States, study participants were exposed to instances of misinformation circulating on X platform. These included a post that falsely asserted that a COVID-19 vaccine would modify human DNA, and another post that falsely alleged that the vaccine would lead to 97% of recipients becoming infertile. Likewise, approximately 20% of Americans believe that the government is using COVID-19 vaccines to microchip the public, which highlights the prevalence of concerns regarding digital surveillance and the commercialization of personal data (Pertwee et al., 2022). Hence scrutinizing people's sentiments regarding COVID-19 vaccination becomes crucial for governments and health ministries to understand how people feel about getting the COVID-19 vaccine. This helps them know what the public thinks about the vaccination process against the virus (Said et al., 2023).

2.5 Related Research

In recent times, the COVID-19 disease has deeply squeezed people's lifestyle. Given these extraordinary circumstances, numerous researchers have embarked on analyzing people's sentiments regarding COVID-19 from various perspectives (Singh et al., 2021). Researchers have employed various methods to analyze public sentiments related to COVID-19, based on their opinions using X platform data. (Ainapure et al., 2023) investigated the impact of integrating deep learning models with lexicon-based approaches to assess the sentiments expressed in COVID-19 related tweets. Lexicon-based methods utilize predefined sentiment dictionaries, whereas deep learning models acquire knowledge from data, enabling a thorough and holistic analysis. The polarity of tweets was classified using a lexicon-based approach with the help of VADER and NRCLex tools. Additionally, a recurrent neural network was trained using Bi-LSTM and GRU methods to classify vaccination-related tweets. An optimized CNN-based approach, utilizing an arithmetic optimization algorithm, seeks to recognize nuanced sentiments by analyzing COVID-19-related tweets. The collaboration of convolutional neural networks (CNNs) and optimization techniques significantly improves sentiment classification accuracy (Aslan et al., 2023).

In another study, researchers combined two methods, TextBlob and machine learning classifiers to analyze sentiments in COVID-19 related tweets. This approach sheds light on public opinions, emotions, and prevailing sentiments. Notably, TextBlob enables sentiment analysis without requiring extensive training data (Kathiravan et al., 2023). By combining BERT with deep CNNs, researchers achieved state-of-the-art results in sentiment analysis of COVID-19 tweets. This approach effectively captures contextual information and complex patterns. The BERT model captures the complex semantics of tweets related to COVID-19 by acquiring contextual representations of words. On the other hand, the deep CNN captures the hierarchical organization of the tweet embeddings (Joloudari et al., 2023). Another research (Albahli and Nawaz, 2023) employed deep learning techniques to analyze sentiments expressed on X platform regarding COVID-19 vaccines. The use of deep learning models allows for capturing nuanced sentiments beyond simple positive or negative labels. This work is to enhance public awareness of coronavirus vaccines, which can aid health departments in halting anti-vaccination movements and encouraging people to receive booster doses of the coronavirus vaccine. *Table 2.1* shows studies related to sentiment analysis on COVID-19 and vaccines.

Table 2.1. Related SA studies on COVID-19 and vaccine-related tweets

No.	Title	Author(s)	Dataset	Sentiment Classifications	Methods	Evaluation
1	TSM-CV: Twitter Sentiment Analysis for COVID-19 Vaccines Using Deep Learning	Saleh Albahli, Marriam Nawaz	Twitter (both historical and real time tweets)	FastText approach for computing semantically aware features VADER to assign the labels as Positive, Negative, or Neutral.	NMF used for feature reduction Random Multimodal Deep Learning classifier (RMDL) for sentiment prediction	Accuracy -94.81% Precision - 97.63% Recall - 97.28% F1-score - 97.5% AUC - 92.59% Confusion matrix
2	Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset	Miftahul Qorib, Timothy Oladunni, Max Denis, Esther Ososanya, Paul Cotae	English tweets related to keywords #vaccine or #covid19 from Twitter	Sentiment score computation using Azure ML, VADER and TextBlob. Vectorization of the dataset using Doc2Vec, TF-IDF, and CountVectorizers	Random Forest, Logistics Regression, Decision Tree, LinearSVC, and Naïve Bayes	Accuracy, Precision, Recall and F1-score are 96.75%, 96.92%, 92.81% and 94.7% respectively
3	Sentiment Analysis of COVID-19 Tweets Using TextBlob and Machine Learning Classifiers	P. Kathiravan, R. Saranya, and Sridurga Sekar	Tweets related to keywords "Pandemic", "Coronavirus", "COVID-19", "SARS-CoV-2" and "Omicron" from	TextBlob text processing	Support Vector Machine (SVM), K Nearest Neighbor (KNN), Logistic Regression (LR), and Decision Tree (DT)	Accuracy of SVM model is 99.4%

			Twitter			
4	TSA-CNN-AOA: Twitter sentiment analysis using CNN optimized via arithmetic optimization algorithm	Serpil Aslan, Soner Kızıloluk, Eser Sert	Tweets related to COVID-19 from Twitter	FastText word embeddings	CNN-AOA based feature extraction CNN, KNN, SVM, DT	Accuracy of TSA-CNN-AOA (Decision tree) is 92.53%, TSA-CNN-AOA (SVM) is 95.01% and TSA-CNN-AOA (KNN) is 95.1%
5	Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches	Bharati Sanjay Ainapure, Reshma Nitin Pise, Prathiba Reddy, Bhargav Appasani, Avireni Srinivasulu, Mohammad S. Khan and Nicu Bizon	Tweets about COVID-19 and vaccination from Twitter	Lexicon-based approach: VADER and NRCLex Deep Learning approach: Bi-LSTM and GRU	RNN trained using Bi-LSTM and GRU techniques, achieving 92.70% and 91.24% accuracy on the COVID-19 dataset	Accuracy of Bi-LSTM model is 92.7% and GRU model is 91.24% on COVID-19 dataset Accuracy of Bi-LSTM model is 92.8% and GRU model is 93.03% on vaccination dataset

6	Sentimental and spatial analysis of COVID-19 vaccines tweets	Areeba Umair & Elio Masciari	Tweets related to COVID -19 vaccine from Twitter	TextBlob to assign sentiment polarity	BERT for sentiment classification	Positive class: Precision - 55%, Recall - 69%, F1-score - 58% Negative class: Precision - 54%, Recall - 85%, F1-score - 64%
7	ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection	K. Hayawi, S. Shahriar, M.A. Serhani, I. Taleb, S.S. Mathew	Tweets related to COVID -19 vaccine from Twitter	Word embedding using TF-IDF and Glove. Labels using LabelEncoder	XGBoost with TF-IDF, LSTM with Glove, BERT	Maximum accuracy in BERT model in test set was 98%. Precision - 97%, recall - 98% and F1-score - 98%
8	Sentiment Computation of UK-Originated COVID-19 Vaccine Tweets: A Chronological Analysis and News Effect	Olasoji Amujo, Ebuka Ibeke, Richard Fuzi, Ugochukwu Ogara and Celestine Iwendi	Tweets related to COVID -19 vaccine by BBC news from Twitter	NA	BERT	NA
9	COVID-19 Vaccine Hesitancy: A Global Public Health and Risk Modelling Framework	Miftahul Qorib, Timothy Oladunni, Max Denis, Esther Ososanya	Tweets related to COVID -19 and vaccine from Twitter	NRCLexicon	1DCNN, LSTM, BERT	BERT model with ten classes produced highest accuracy 96.71%.

	Using an Environmental Deep Neural Network, Sentiment Classification with Text Mining and Emotional Reactions from COVID-19 Vaccination Tweets	and Paul Cotae				LSTM model achieved the accuracy of 89.93%.
10	A Deep Learning Approach for Sentiment Classification of COVID-19 Vaccination Tweets	Haidi Said, BenBella S. Tawfik, Mohamed A. Makhlouf	COVID-19 vaccine-related tweets from Twitter	VADER	ML models LR, NB, DT, RF, and KNN. DL models LSTM, BiGRU. Ensemble model LSTM-2BiGRU	Accuracy, Precision, Recall, and F1-Score. LSTM-2BiGRU model performed well with the accuracy of 92.46%.

Both sentiment and spatial patterns in tweets related to COVID-19 vaccines analyzed in a study (Umair and Masciari, 2023) that provides insights into how different regions perceive vaccination efforts. The authors used TextBlob() function to determine the polarity of the tweets and classified the sentiment using BERT model. By analyzing sentiments over time, a study explores how news events impact public perceptions and emotions regarding vaccination. This research (Amujo et al., 2023) computes sentiments from COVID-19 vaccine-related tweets originating in the UK. In this study, BERT model applied for sentiment classification of tweets related to COVID-19 vaccinations and the authors found that the negative sentiment as common in the period considered for the study. The authors (Qorib et al., 2023b) contribute to understanding public attitudes, concerns, and overall sentiment dynamics by a study that identifies key sentiments expressed by users, by applying machine

learning algorithms such as Naive Bayes, Support Vector Machine, and Logistic Regression. It sheds light on vaccine hesitancy, concerns, and overall public sentiment.

The study (Qorib et al., 2023a) proposes a comprehensive framework that combines environmental factors, sentiment classification, and emotional reactions from tweets. By modeling vaccine hesitancy globally, it provides a holistic understanding of vaccination sentiments. Tweets categorized into ten sentiments using NRCLexicon method and multi-classification model using neural networks. The authors introduced a new dataset from X platform specifically focused on vaccine misinformation (Hayawi et al., 2022). By identifying false narratives related to COVID-19 vaccines, this dataset contributes to combating misinformation and promoting accurate information dissemination. Research found that trust, anticipation, and fear were the three most predominant emotions. The authors (Saleh et al., 2023) analyzed 2.4 million English tweets from nearly 1 million user accounts matching keywords related to covid-19 and vaccines.

The paper (Rani and Jain, 2023) proposes a deep fusion model aiming to understand public opinion around COVID-19 vaccine and omicron variant. The proposed model aims to accurately classify the sentiment polarity of COVID-19 vaccine and omicron variant tweets by overcoming the challenge of contextual polarity ambiguity in user-generated data. Emergency use authorization of COVID-19 vaccines in India was analyzed using Lexicon-based and machine learning method in the study (Paliwal et al., 2022). Another research (Ferdous et al., 2022) proposes a neural network model that uses text vectorization and compares it with long short-term memory (LSTM) and bidirectional long short-term memory (Bi-LSTM) to analyze the sentiment of COVID-19 vaccine-related tweets.

The reviews on the above studies collectively contribute to our knowledge of public sentiments, vaccine hesitancy, and the impact of COVID-19 vaccines on society. Researchers and policymakers can leverage these insights to inform public health strategies, communication efforts, and targeted interventions.

2.6 Discussion

Based on the reviews, it is evident that various research has been performed to analyze the sentiment of COVID-19 and its vaccine from public posts. The majority of people expressed neutral or positive sentiments towards COVID-19 vaccines. Positive sentiment was more prevalent than negative sentiment (Huangfu et al., 2022). In the study (Jabalameli et al., 2022), only lexicon-based approach was used to identify the sentiments from the tweets whereas using both lexicon-based and deep learning approach together provide a better model. Supervised learning and ensemble techniques are the primary machine learning approaches used for sentiment analysis. The machine learning approach for sentiment analysis relies on a large labeled training dataset, as the model's performance is reliant on the quality of the dataset. Deep learning models like Bi-LSTM and GRU better in capturing sequential dependencies, while transformer-based models like BERT achieve state-of-the-art performance. However, considering computational resources and interpretability, hybrid approaches and ensemble methods remain valuable alternatives (Ainapure et al., 2023).

A deep learning model, BERT compared with other three ML models such as LR, SVM and LSTM in the research (Chintalapudi et al., 2021). However, BERT model outperformed other three models with better accuracy. This study extracted the emotions using single country tweets which may not be applicable across the globe. Public emotions related to COVID-19 vaccines and vaccination campaigns were analyzed using a Bi-LSTM model trained on tweets from Bangladesh (Zulfiker et al., 2022). The model's limitation is that it can only predict two classes of emotions such as positive and negative, and it does not have the capability to handle tweets that are ambiguous in nature. Enhancing CNN-based sentiment analysis involves optimizing hyperparameters and selecting relevant features. Also, comparing the developed model against existing methods is crucial (Aslan et al., 2023).

The study (Kathiravan et al., 2023) offers valuable insights at the same time additional research could improve robustness and generalizability of sentiment analysis models for COVID-19-related tweets. BERT and Deep CNN models offer promising avenues for sentiment analysis but addressing the research gap in lightweight BERT models and fine-tuning strategies can further advance our understanding of COVID-19 sentiments on social media (Joloudari et al., 2023). Also, the study does not explicitly examine the impact of

misinformation on vaccine sentiment. However, incorporating techniques for detecting misinformation could yield valuable insights into how false narratives shape public opinion. Potential implications and recommendations have not been discussed in the study (Albahli and Nawaz, 2023) so comprehensive discussion on the results could benefit. Understanding the basic factors that contribute to this unusual occurrence could ease approaches for effective communication during critical vaccine-related situations (Umair and Masciari, 2023).

The paper (Qorib et al., 2023a) provides valuable insights on COVID-19 vaccine hesitancy. However, analyzing user engagement (likes, retweets, replies) with vaccine-related tweets could uncover influential voices and community dynamics. Understanding how to mitigate the impact of misinformation on vaccine uptake is crucial. The main limitation from the studies is that it only analyzed tweets from X platform, which may not representative of the global public perceptions. Another gap in the research of understanding public attitudes towards the COVID-19 vaccine is that the studies did not analyze tweets in language other than English (Hayawi et al., 2022; Saleh et al., 2023).

2.7 Summary

The analysis conducted clearly indicates that various methods has been proposed to accurately assess human sentiment regarding the COVID-19 vaccine. Researchers have used lexicon, traditional ML, deep learning and hybrid methods to analyze the sentiment. Also, some of the studies analyzed the tweets' location wise and analyzed the tweets accordingly. Few researchers classified tweets only into binary class (i.e. positive or negative). Still, there remains a gap that requires performance enhancement. In previous studies, researchers typically employed only one sentiment classification method, such as lexicon-based analysis or machine learning. However, in this study, we utilize both lexicon-based and deep learning approaches for sentiment analysis. This comprehensive approach allows us to evaluate and identify the most operative model for predicting public view regarding COVID-19 vaccination.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Amidst the COVID-19 pandemic, a large number of individuals took to X platform to express their thoughts and opinions on the disease and its vaccine. X platform has become a valuable resource for current news and commentary due to its broad demand and ability to scale public opinion. This makes X as an excellent platform to research a variety of subjects (Arora et al., 2019). X platform is a social media platform that allows users to engage in real-time conversations and stay updated with news and events through a newsfeed, using a micro-blogging format. X platform allows users to create tweets that can include text content up to 280 characters in length, as well as pictures, hashtags, and videos. Tweets sent by users are publicly available, but the sender can control the distribution of the communication by making it accessible only to their followers. X platform has 666 million active users worldwide and it is expected that these figures will continue to grow as mobile device usage and mobile social networks increasing as per (Statista - Popular social networks worldwide as of October 2023, 2024). This study aimed to analyze the sentiments of COVID-19 vaccine-related tweets following lexicon-based, deep learning and hybrid methods. Lexicon based approach will be used to classify the sentiments and deep learning models will able to predict the sentiment. Combination of lexicon, deep learning and hybrid models provides a robust framework for sentiment analysis. The dataset is available to the public on Kaggle website. The data will be cleaned and improved through a sequence of preprocessing steps. Sentiment classification will be performed using Valence Aware Dictionary and Sentiment Reasoner (VADER), a rule and lexicon-based technique. Finally, sentiment prediction will be done using two deep learning models, transformed-based ML-model called Bidirectional Encoder Representation from Transformers (BERT) and Bidirectional Long Short-Term Memory (Bi-LSTM) with BERT word embeddings. The models performance will be evaluated using Accuracy, Precision, Recall, and F1-score. Also, the prediction of both the models will be compared using classification report.

3.2 Research Approach

In this section, the approach as shown in *Figure 3.1* to conduct the research is described. The subsequent sections define each action in detail.

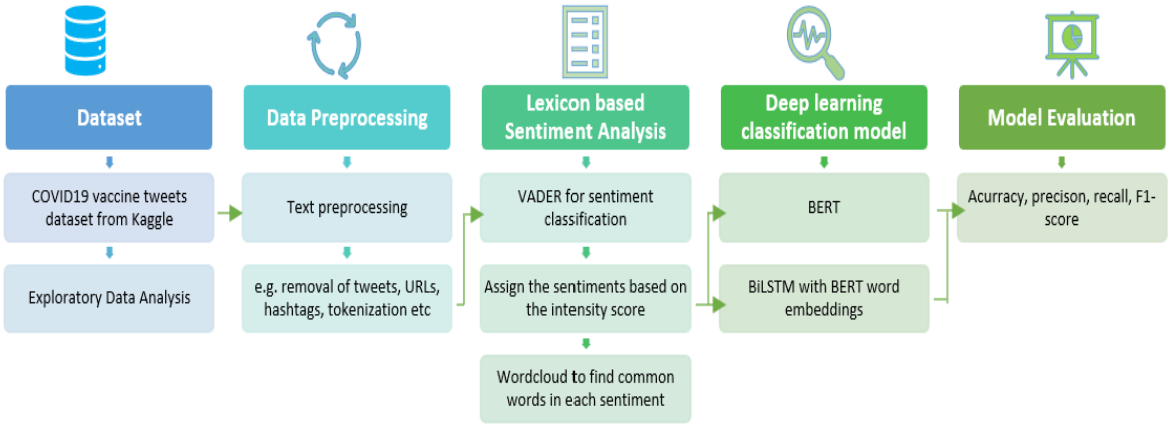

Figure 3.1. Overview of research methodology

3.2.1 Understanding the dataset

This study utilizes a dataset named 'COVID-19 All Vaccines Tweets' which is publicly available on Kaggle website (Preda, 2021). The author of the dataset collected recent tweets about COVID-19 vaccines that are being used on a large scale across the world. Python package “Tweepy” was used to collect data from Twitter API. The dataset includes 228,207 tweets related to COVID-19 vaccines such as Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V, spanning from 12 December 2020 to 23 November 2021. In addition to the text, the dataset includes 16 features such as the user’s name, location, number of followers, number of friends, date, hashtags, retweets, etc. The tweets in the dataset are in English language. *Table 3.1* shows few sample tweets from the dataset.

Table 3.1. Sample tweets from the dataset

Text
The first COVID-19 vaccine doses have arrived in Oregon. Watch #MorningsontheDove from 6am -9am tomorrow for detâ€¦

https://t.co/eNy64xR7KX
#AnitaQuidangen, a personal support worker in #Toronto, became one of the first people in #Canada to receive aâ€¦ https://t.co/bgbOIL12Hp
#BreakingNews A nurse in New York City on Monday became the first person in the United States to receive the coronaâ€¦ https://t.co/02Mu5HKYs5
President Trump says the first #PfizerBioNTech #COVID19 vaccine in the United States has been administered. https://t.co/c4AIbXNRMz
Two people in the United Kingdom have experienced an allergic reaction to the Pfizer/BioNTech COVID-19 vaccine. Reaâ€¦ https://t.co/2LNAcAa80R
The US public will start receiving the Pfizer/BioNTech coronavirus vaccine from Monday after it was authorised forâ€¦ https://t.co/tyUU4ZSjy1

Of the records in the dataset, 70% contain user location data. Based on this information, 25% of the tweets in the dataset are from India, 3% are from the USA, and the remaining tweets are from other countries. According to the data presented in *Figure 3.2*, there was a noticeable increase in the number of tweets in the middle of 2021, indicating that people were discussing the vaccine more often during that time.

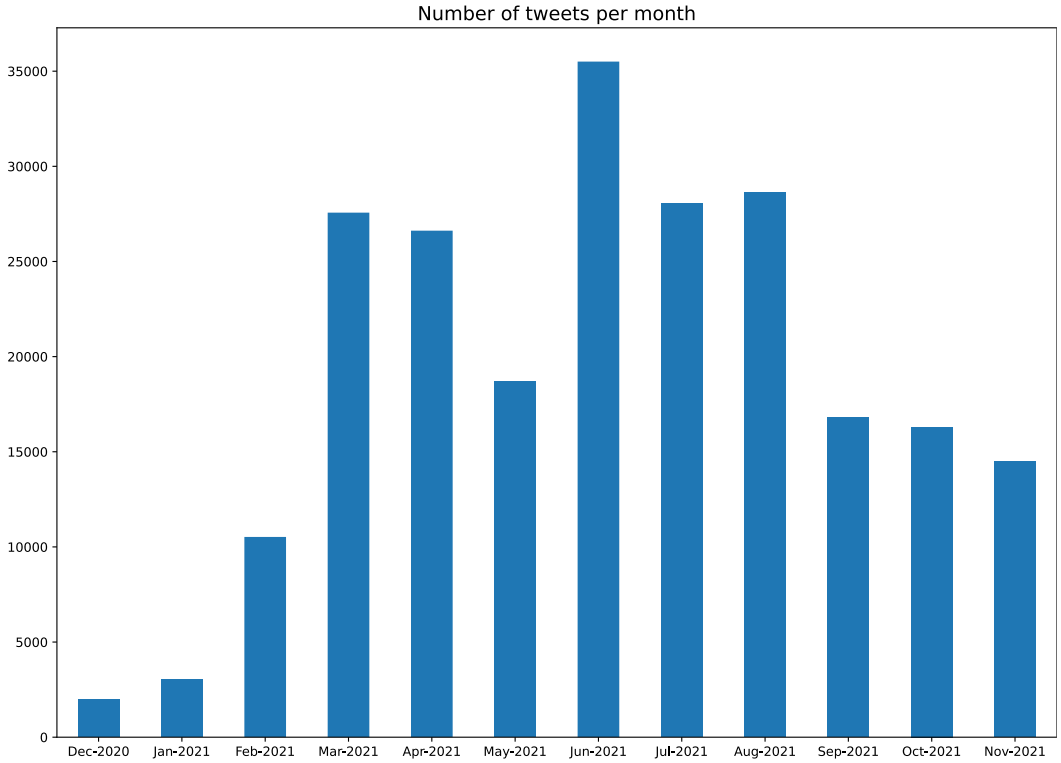

Figure 3.2. Number of tweets per month from dataset

3.2.2 Data preprocessing

Data cleansing is a crucial step in text mining that involves extracting meaning from text by removing non-analyzable words and other irrelevant components. Due to the presence of grammatical errors and typos, tweet text from Twitter is not suitable for direct utilization in model preparation. Twitter data often includes irrelevant special characters, expressions, links, tags, emojis, signs, etc., that can have a negative impact on experimental studies during the analysis process. The subsequent sections provide detailed information and a visual summary of the all the preprocessing steps shown in *Figure 3.3*. This study focuses solely on analyzing the sentiments of tweets. Therefore, removed all columns except the text column are removed. The dataset does not contain the retweets and no null value found in the text column.

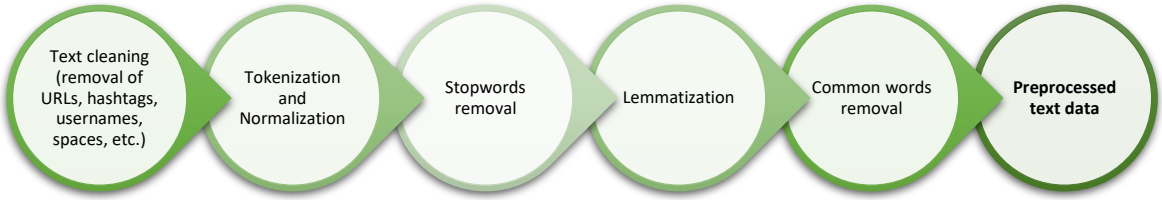

Figure 3.3. Steps of text data preprocessing

3.2.2.1 Text Cleaning

The tweet text may contain hyperlinks, punctuations, signs, hashtags, and usernames that are not relevant for the sentiment analysis of the text. Also, the frequent words such as covid19, vaccine and corona will be removed as their frequency is higher than other words which may cause skewness. Removing these unnecessary items can enhance the model's efficacy. Therefore, they are excluded from the text.

3.2.2.2 Emojis

Emojis are frequently used in the Tweets by people to express their feelings. Converting these emojis into their meaningful words might improve sentiment analysis of the tweets.

3.2.2.3 Tokenization and Normalization

Tokenization is the process of splitting the tweets into separate units such as words, symbols, keywords, and meaningful phrases known as tokens. This can be helpful to analyze the words used in a sentence. Normalization is the process of converting all the words in a collection to lowercase, so that they are normalized.

3.2.2.4 Stop words

Words such as "a", "the", "is", "are", etc. are called stop words. As these words carry very little useful information, removing them ease the text analysis.

3.2.2.5 Lemmatization

Lemmatization is used to reduce the words to its base form but it returns base or dictionary form of a word, which is known as lemma. For example, the lemma of the words "improve", "improving", and "improvements" is "improve". According to researches (Albahli and Nawaz, 2023; Amujo et al., 2023; Qorib et al., 2023b), this method has the potential to improve sentiment analysis of text.

3.2.3 Sentiment analysis using Lexicon-based approach

Text sentiment analysis is a popular topic in NLP that involves computing sentiment scores for text. Lexicon-based strategies are simple and efficient methods that rely on a sentiment dictionary. This dictionary contains a list of lexical features such as words and phrases, each labelled with a semantic orientation of positive, negative, or neutral (Thangavel and Lourdusamy, 2023). The initial step involves pre-processing the input text followed by its transformation into a bag of words. The sentiment values for each word are extracted by comparing them with a dictionary. Finally, an aggregation function such as sum or average is applied to the individual sentiment scores to predict the overall sentiment of the text. Lexicon-based models can achieve good classification performance without requiring a large amount of labeled training data, making them superior to classifier models. Some popular lexicon-based models used for sentiment analysis include AFINN, SentiWordNet, VADER, and Bing Liu’s Lexicons. In this study, VADER tool will be used for sentiment classification.

3.2.3.1 VADER for sentiment analysis

VADER is a sentiment analysis tool that is designed to analyze sentiments expressed in social media platform. It is a rule-based model that uses a sentiment lexicon and rules to determine the sentiment score of a piece of text. The sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative. For example, words like good, appreciate are positive lexicons whereas bad, awful are negative lexicons. VADER not only talks about the positivity and negativity score but also tells us about how positive or negative a sentiment is. The VADER sentiment lexicon is specifically designed to detect both the polarity and intensity of sentiments expressed in social media contexts (Hutto and Gilbert, 2014).. The compound score is a metric used to calculate the sentiment of a given text or sentence. It is calculated by summing the valence scores of each word in the lexicon, adjusting them according to certain rules, and then normalizing the result to be between -1 (most extreme negative) and +1 (most extreme positive). In this research, the preprocessed text will be fed into VADER, which performs sentiment classification by taking into account both the terms used and the score of the added emotional polarity. Based on the compound score, each sample is assigned with one of three labels: positive (>=0.05), negative (<=-0.05), or neutral (other score) as shown in Figure *3.4*.

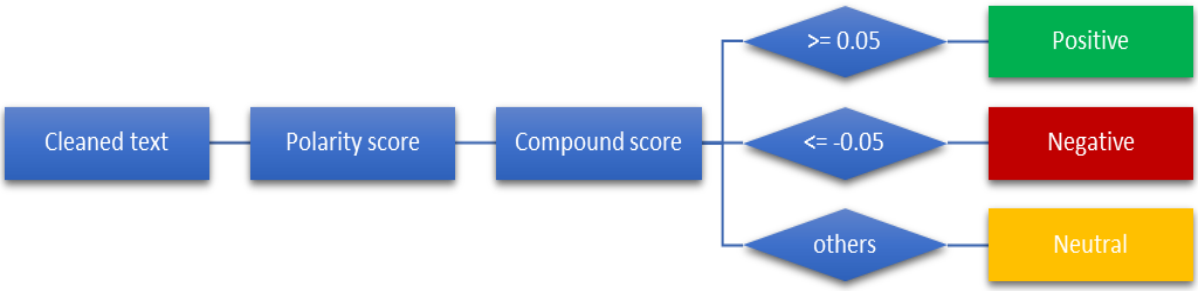

Figure 3.4. VADER - Sentiment classification

VADER is a lightweight and functional tool for sentiment analysis and it doesn't require extensive training data. However, it relies on pre-defined lexicon so it might miss some important sentiments if a word misspelled, used incorrectly, short or abbreviated. Also, it

cannot detect sarcasm or real truth behind the words. So, deep learning and hybrid approaches are also considered in this research.

3.2.4 Sentiment analysis using Deep learning approach

Deep learning is a subset of machine learning that involves the use of artificial neural networks with multiple layers to model and solve complex problems. Deep learning algorithms can automatically learn and improve from data without the need for manual feature engineering. Deep learning has been used extensively for sentiment analysis due to its ability to capture the context of words in a sentence, which is important for understanding the sentiment of a text. The use of neural networks, such as RNNs, CNNs, and transformer models, has led to significant progress in sentiment analysis tasks (Sahoo et al., 2023). The paper "Attention is All You Need" (Vaswani et al., 2017) introduced Transformers, which have transformed the field of natural language processing, including sentiment analysis. Transformers are a type of neural network that use self-attention mechanisms to capture global dependencies between words in a sentence without relying on sequential processing. The self-attention mechanism helps the model focus on relevant words and weigh their importance when forming representations. To improve sentiment analysis performance, Transformers are often combined with pre-trained language models, such as BERT or GPT, to leverage their contextualized word embeddings.

3.2.4.1 BERT

BERT uses a bidirectional encoder stack of transformers and ELMo word embeddings to achieve state-of-the-art accuracy on many NLP and NLU tasks. The architecture of BERT consists of two sizes: BERT BASE model has 12 encoders with a hidden layer’s size of 768 and 110 million parameters and the BERT LARGE model has 24 encoders with a hidden layer size of 1024 and 340 million parameters. The BASE model is used to measure the performance of the architecture comparable to another architecture, while the LARGE model produces state-of-the-art results (Sahoo et al., 2023). Compared to traditional models like word2vec, BERT has an advantage. In word2vec, each word has a fixed representation that is independent of the context in which it appears. On the other hand, BERT generates a representation that is dynamically informed by the words surrounding it. For different natural language processing tasks, the model input can be fine-tuned. The pre-training process of BERT model involves gradually adjusting the model parameters. This helps the semantic

representation of text output by the model to describe the essence of the language and facilitate subsequent fine-tuning for specific NLP tasks (Cai et al., 2020). Masked Language Modeling (Masked LM) and Next Sentence are two pre-training tasks proposed by BERT to achieve the goal of generating a semantic representation of text that can describe the essence of the language and facilitate subsequent fine-tuning. Masked LM is a pre-training task that involves masking some of the input tokens and training the model to predict the masked tokens based on the context provided by the other tokens. Next Sentence is a pre-training task that involves predicting whether two sentences are consecutive in the original text.

Due to its bidirectional nature, BERT can also handle sarcasm by considering preceding and following words. There are some challenges with BERT model that they require significant computational resources to train and test the model. Also, the model is more effective only when trained on more amount of data.

3.2.4.2 Bi-LSTM

LSTM is type of recurrent neural network (RNN) architecture that is designed to model chronological sequences and their long-range dependencies more precisely than conventional RNNs. Bi-LSTM is a sequence model that contains two LSTM layers, one for processing input in the forward direction and the other for processing in the backward direction. So, they are capable of capturing the context of a word by considering both the past and future words in a sentence (Gou and Li, 2023).

3.2.4.3 Bi-LSTM with BERT word embeddings

The Bi-LSTM with BERT method is distinct from the traditional method of combining weights between BERT and Bi-LSTM. Rather than using a weight-combination approach, Bi-LSTM with BERT employs BERT as its primary component and Bi-LSTM as its secondary component. BERT is effective in comprehending statistical patterns among words with similar meanings, while Bi-LSTM excels in capturing contextual details. This is similar to how human language functions, as grammar is rooted in statistical patterns (Gou and Li, 2023). To predict the sentiments, BERT word embeddings will be generated and passed on to the Bi-LSTM model. This combination allows the model to benefit from both BERT's contextual understanding and Bi-LSTM's bidirectional processing. The structure of the model is shown in Figure 3.5 (Cai et al., 2020).

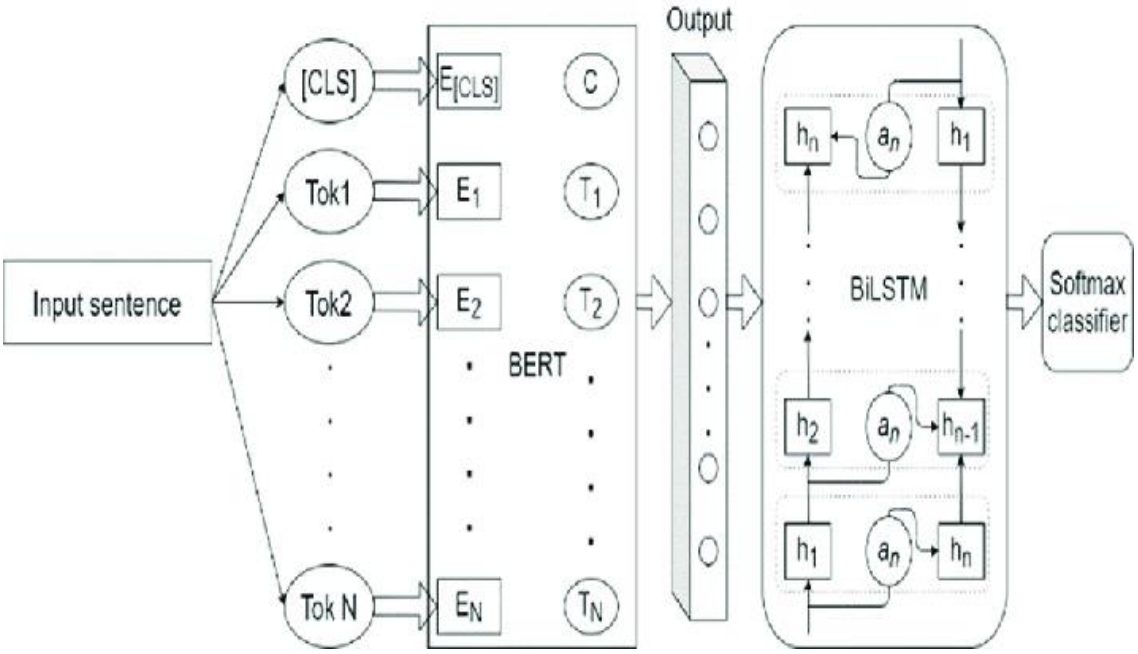

Figure 3.5. Structure of Bi-LSTM with BERT

Combining the models Bi-LSTM with BERT can improve the model performance as BERT has the capability to learn statistics features of the neighboring words and Bi-LSTM is capable in learning the context info. The BERT model generates a word vector that reflects the input sequence's bidirectional Transformer structure when a textual sequence is inputted. The SoftMax activation function is used to generate the output of the sentiment classification. Integrating the Bi-LSTM with BERT for sentiment classification contents efficiently improves the final feature vector of the sentiment classification with accuracy (Zhou, 2023).

3.2.5 Model Evaluation

There are various types of evaluation indicators to evaluate the performance of the models. Therefore, selecting the appropriate evaluation indicators based on the problem is an essential approach to measure the effectiveness of the model. Accuracy, precision, recall and F1-measure are the commonly used for sentiment analysis. The following components are required to compute these metrics:

1. True Positive (TP): the number of positive observations that were correctly classified.
2. False Positive (FP): the number of positive observations that were incorrectly classified.

3. True Negative (TN): the number of negative observations that were correctly classified.
4. False Negative (FN): the number of negative observations that were incorrectly classified.

Accuracy is a commonly used evaluation metric in classification problems. It is calculated as the ratio of the number of correct predictions to the total number of predictions made by the model. However, it has some limitations when the training set has unbalanced samples. For instance, if the negative sample ratio is 99%, a model that predicts all samples as negative would have an accuracy rate of 99%. This approach is not the most reasonable way to evaluate the model's performance. The formula of the accuracy is stated in 3.1.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{3.1}$$

Precision is a metric that measures the proportion of true positives among all positive predictions made by the model. It is commonly used to evaluate the predictive power of a model and it is important measure when the cost of false positive is high. The formula of the precision is stated in 3.2.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.2}$$

Recall measures the proportion of true positives out of the total number of actual positive cases in the data. Recall is useful when the cost of false negatives is high, and the number of false negative to be minimized. It is also known as sensitivity or true positive rate. The formula of the recall is shown in 3.3.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.3}$$

The F1-score is a way of combining the precision and recall of the model, and it is useful when the classes are imbalanced. It is the harmonic mean of precision and recall. The F1-score ranges from 0 to 1, with a higher score indicating better performance.

$$\text{F1 score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3.3)$$

3.2.6 Resampling using bootstrapping

Bootstrapping is an arithmetic technique that come under the wider class of resampling methods. It provides random sampling with replacement from an original sample. (Efron and Tibshirani, 1994) introduced the concept of bootstrapping, to create assurance without needing to understand the exact distribution. The idea of bootstrap has been commonly incorporated and adapted in multiple ways (Ma et al., 2024). Due to the limitation of computational resources and time, bootstrap resampling with sample replacement will be used in this study. The models will be executed with the multiple set of bootstrapped samples with sample replacement. An example of bootstrapping is illustrated in the Figure *3.6*, 100 samples from each sentiment will be taken to form 300 equally distributed dataset. The bootstrapped data will be further split into train and test which will be used for training and testing the models.

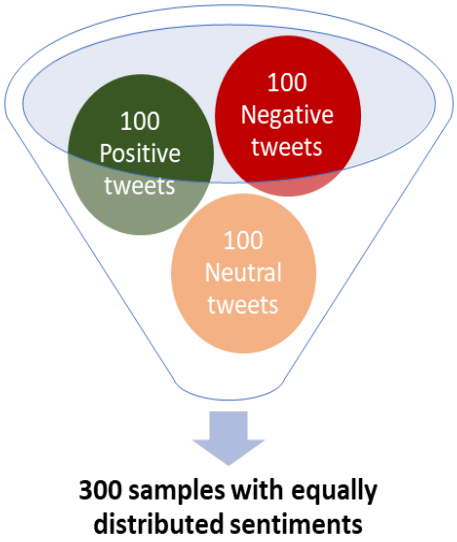

Figure 3.6. Bootstrap resampling

3.3 Resource requirements

The primary hardware and software required to run the models are discussed in the following sections. The execution time required to run the models solely depends on machine used and

the number samples used for training and testing. For instance, with the configuration mentioned in section 3.3.1, the estimated time to execute deep learning models on 5000 samples falls within the range of two to three hours.

3.3.1 Hardware Requirements

A computer having at least 24 GPU, 12GB of RAM, 8 CPU and 40 GB of storage required to run the deep learning and hybrid models. The computer should also connect to internet and able to compile and execute the python codes. Otherwise, cloud computing GPU platforms such Google Colab, Paperspace, etc., can also be considered to run the deep learning models. Paperspace is a cloud platform that provides GPU enabled Jupyter Notebooks and this platform will be used to execute the developed code of this study.

3.3.2 Software Requirements

The study required the following software are required, but this list is not limited.

- Python (3.6 or higher).
- Jupyter notebook (6.1 or higher) for python coding.
- Pandas (1.4.2 or higher) for data processing.
- Numpy (1.21.5 or higher) for array computation, mathematical functions, etc.
- Matplotlib (3.5.1 or higher), Seaborn (0.11.2 or higher) for visualization.
- VADER (3.3.2 or higher). vaderSentiment package, NLTK, NumPy and SciPy required for VADER.
- BERT (1 or higher). It requires transformers package, PyTorch and TensorFlow.
- WordCloud (1.9.2 or higher). wordcloud package, NumPy, Pillow, matplotlib required to use WordCloud.
- Libraries such as contractions, emoji, squarify, transformers, tensflow and keras are required.

3.4 Summary

This thesis presents an approach to analyze COVID-19 vaccine-associated tweets data using a combination of lexicon, deep learning and hybrid techniques. The dataset used in this study is publicly available on Kaggle website. Sequence of data preprocessing steps performed to improve and clean the data. A rule and lexicon-based sentiment technique called VADER will

be used for sentiment classification. Bootstrap resampling with sample replacement method will be used to resample the original population. Finally, prediction of sentiments has been performed using a deep learning, BERT and a hybrid model, Bi-LSTM with BERT word embeddings. The performance of the model will be assessed with Accuracy, Precision, Recall and F1-score. Decent computational resources are required to run the deep learning and hybrid models. While the study utilizes publicly available tweet datasets, no user identities will be revealed in any of the sections. Also, the sentiments assigned to the tweets are purely based on the developed methods with no biases involved.

CHAPTER 4

ANALYSIS AND DESIGN

4.1 Introduction

This chapter provides detailed explanations of the model development and describes the exploratory data analysis conducted on the dataset. The first few subsections cover the dataset description, data import, basic checks, and data visualization to gain insights into the tweet texts and the relevant features. In the subsequent sections, data cleaning steps executed to address the noisy, redundant and unnecessary details in the tweet text are elaborated. Plotted WordCloud to visualize the key words before and after text cleaning. Then the polarity scores for the cleaned text computed using VADER sentiment analysis and assigned three sentiment labels (i.e. Positive, Negative and Neutral) based on these scores. WordCloud generated for each sentiment that shows the major common words in the tweets text. Sample of each sentiment is taken and a new data frame created to address the class imbalance problem. This resampling approach helped to reduce the computational requirement while running the deep learning models. Later, two deep learning models BERT and Bi-LSTM with BERT word embeddings are trained and evaluated. Finally, the model performance is evaluated using accuracy, precision, recall and F1-score. The experimented analysis and design steps are discussed in detail in the following sections.

4.2 Dataset Description

The dataset "COVID-19 All Vaccines Tweets" from Kaggle website is used for this research. The dataset includes 228,207 records and 16 features related to tweets discussing COVID-19 vaccines worldwide during the period from December 2020 to November 2021. The information of dataset such as name of columns, number of records, missing value in percentage and its data type are shown in *Table 4.1*. There are five fields with missing records, user location and hashtags columns are the top two columns that missing values. The text column that contains the actual tweets posted by public do not have any null values. There are six numerical, two Boolean and eight categorical columns. The challenge in the dataset is that the tweets are in short form and not the extended or full tweet text.

Table 4.1. COVID-19 All Vaccine Tweets dataset information

Column Name	Number of records	Missing value (%)	Data Type
id	228207	0	int64
user_name	228205	0.001	object
user_location	161296	29.32	object
user_description	211189	7.457	object
user_created	228207	0	object
user_followers	228207	0	int64
user_friends	228207	0	int64
user_favourites	228207	0	int64
user_verified	228207	0	bool
date	228207	0	object
text	228207	0	object
hashtags	178504	21.78	object
source	228088	0.052	object
retweets	228207	0	int64
favorites	228207	0	int64
is_retweet	228207	0	bool

4.3 Exploratory Data Analysis

In this section, the findings from exploratory data analysis are documented. The tweets data in the dataset are posted in Twitter by 85,549 unique users and out of that only 3,017 are verified users. *Figure 4.1* shows that only 3.5% of unique users are verified and remaining are not verified users.

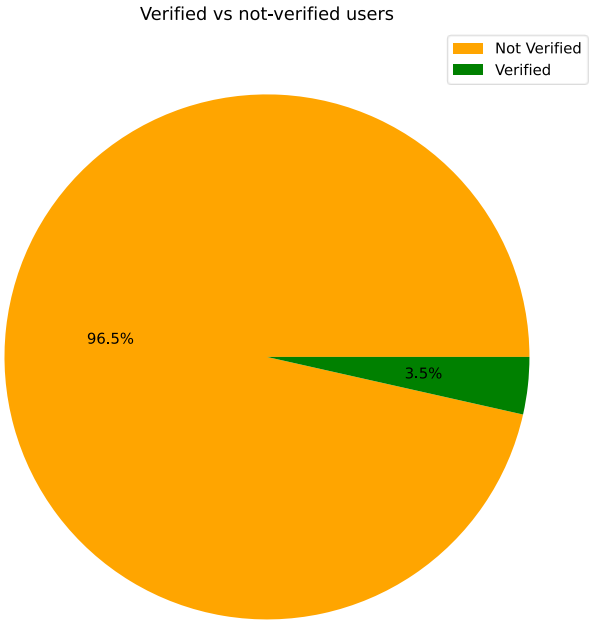

Figure 4.1. Verified vs Not-Verified users

The text field contains 226,373 distinct tweets and there are no retweets included in the dataset. The top 5 retweeted tweets with the retweet count are listed in the *Table 4.2*.

Table 4.2. Top 5 retweets

Text	No. of retweets
This video fits the last almost 2 years into 2 minutes. At #SputnikV we strongly believe that it is only through Va… https://t.co/Ggi7X5qO8x	12294
RDIF, Laboratorios Richmond launched production of #SputnikV in Argentina, the first country in Latin America to ma… https://t.co/oEMaUwVR92	11288
Why we need Two Doses of mRNA Vaccine 💉 #vaccines #COVID19 #Pfizer #moderna #VaccinesSaveLives #vaccinated https://t.co/RFRmPAyubD	7695
We completely reject the false and malicious reporting by @CNBCTV18News on COVAXIN® supplies to international marke… https://t.co/OXgKYg2YLL	6018

ICMR study shows #COVAXIN neutralises against multiple variants of SARS-CoV-2 and effectively neutralises the doubl… https://t.co/0IYwr0KymJ	4851

In the period covered by the dataset, maximum number of tweets posted in the day of Wednesday as shown in *Figure 4.2*. The number of tweets posted remains comparatively consistent on all weekdays except Sunday as there is noticeable decrease in tweet activity.

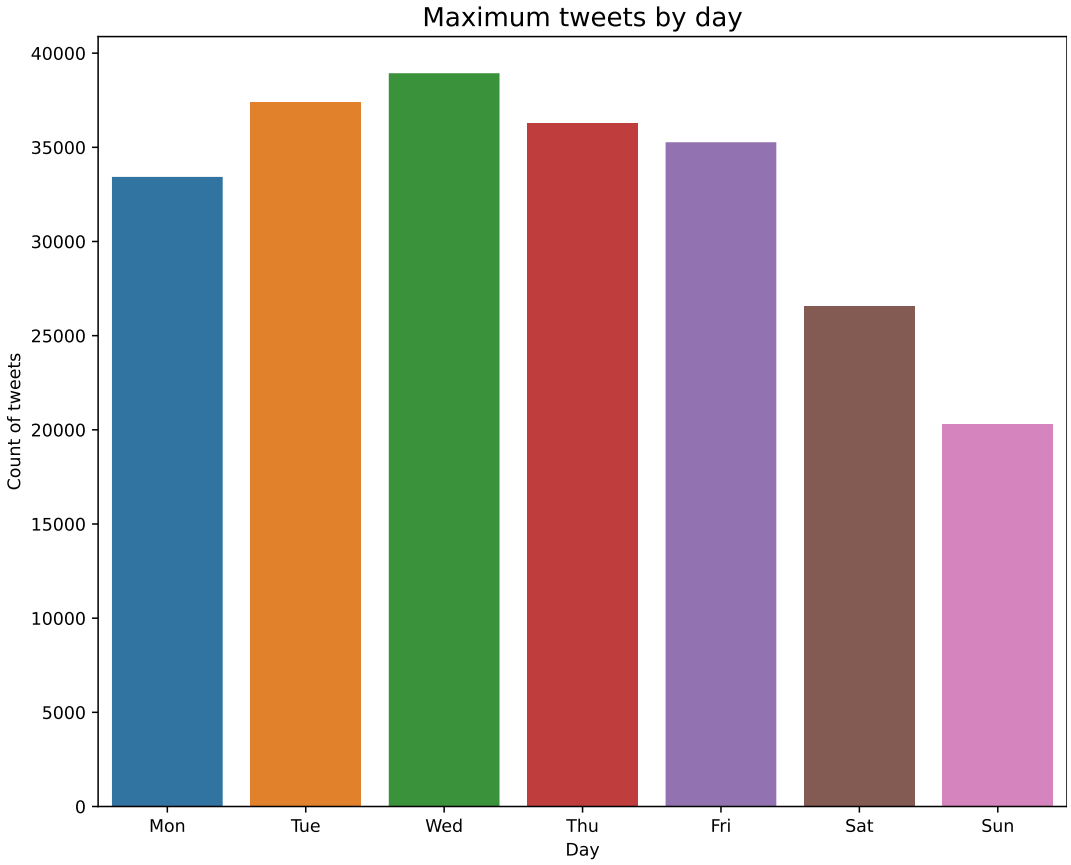

Figure 4.2. Day of tweets

There are two users who posted more than 5% of total tweets in the dataset and they are "CoWin Blore 18-44" and "CowinBangalore". The top ten users who posted a greater number of tweets are displayed in the *Figure 4.3*.

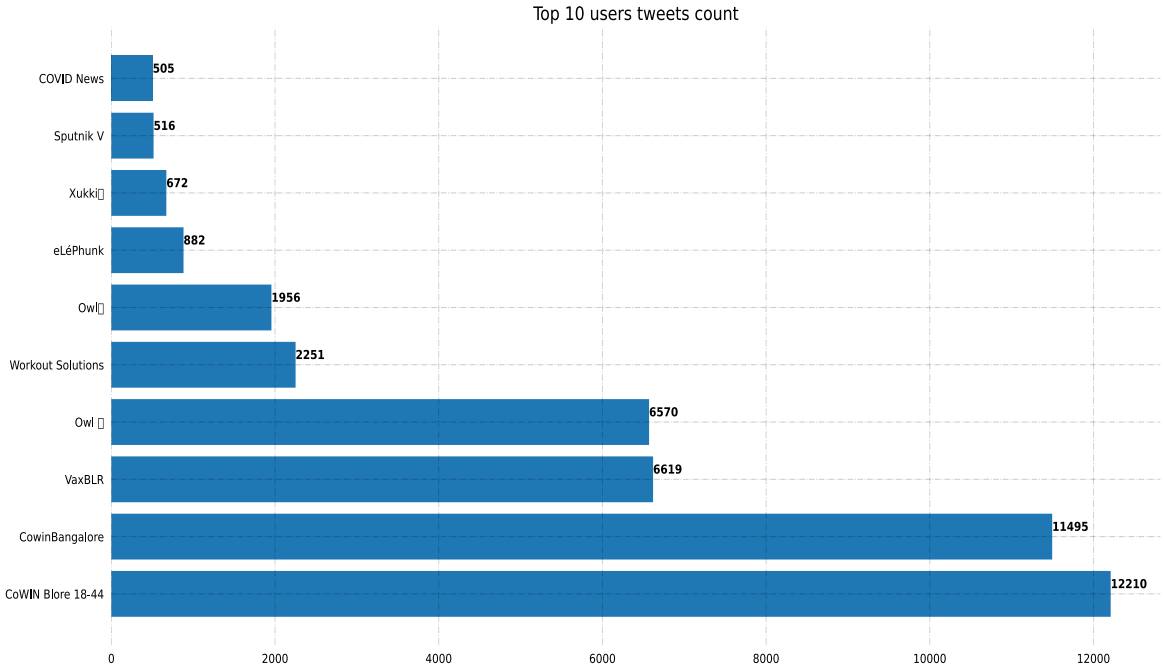

Figure 4.3. Top ten users by tweets

The length of the tweet text is plotted using histogram as in ***Figure 4.4*** and found that most of the tweets are around 140 to 145 words. This implies that the users exposed their thoughts about vaccines in detail.

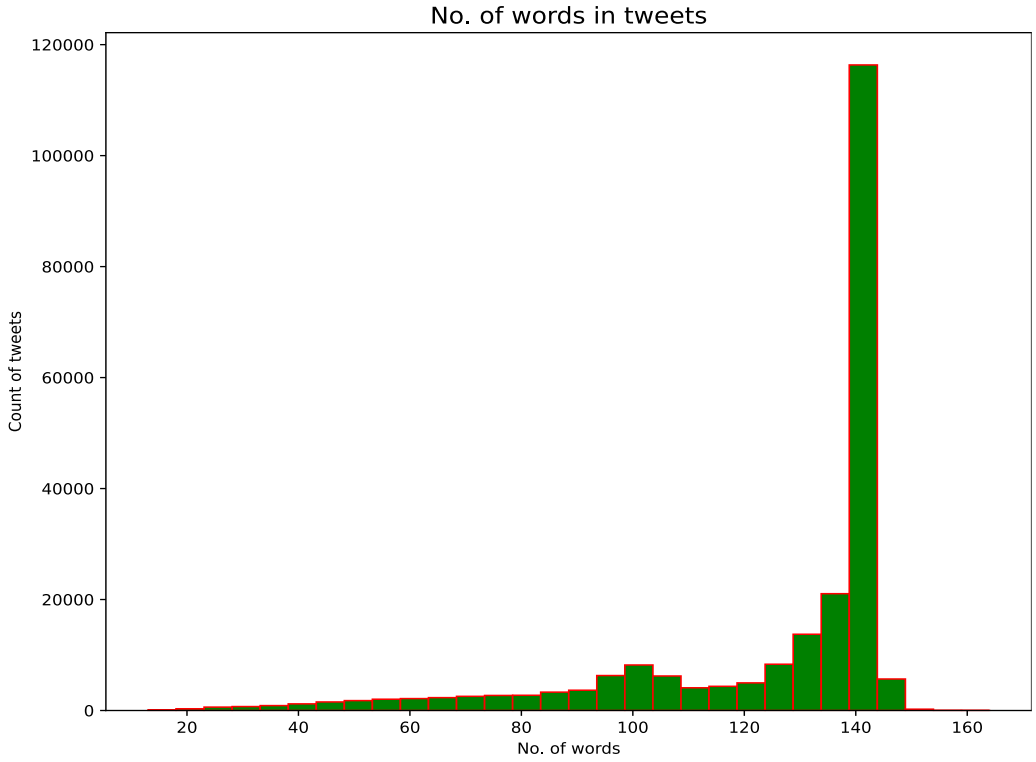

Figure 4.4. Number of words in Tweets

Hashtags starts with # symbol in tweets and it makes easier for people to find similar content in the social media platforms. For example, adding hashtag #covaxin to the tweets makes that tweet related to the larger topic about covaxin. The top ten hashtag used in the tweets are visualized and shown in *Figure 4.5*. The hashtag 'covaxin' mentioned in 22,740 tweets followed by 'moderna' in 13,142 tweets and so on.

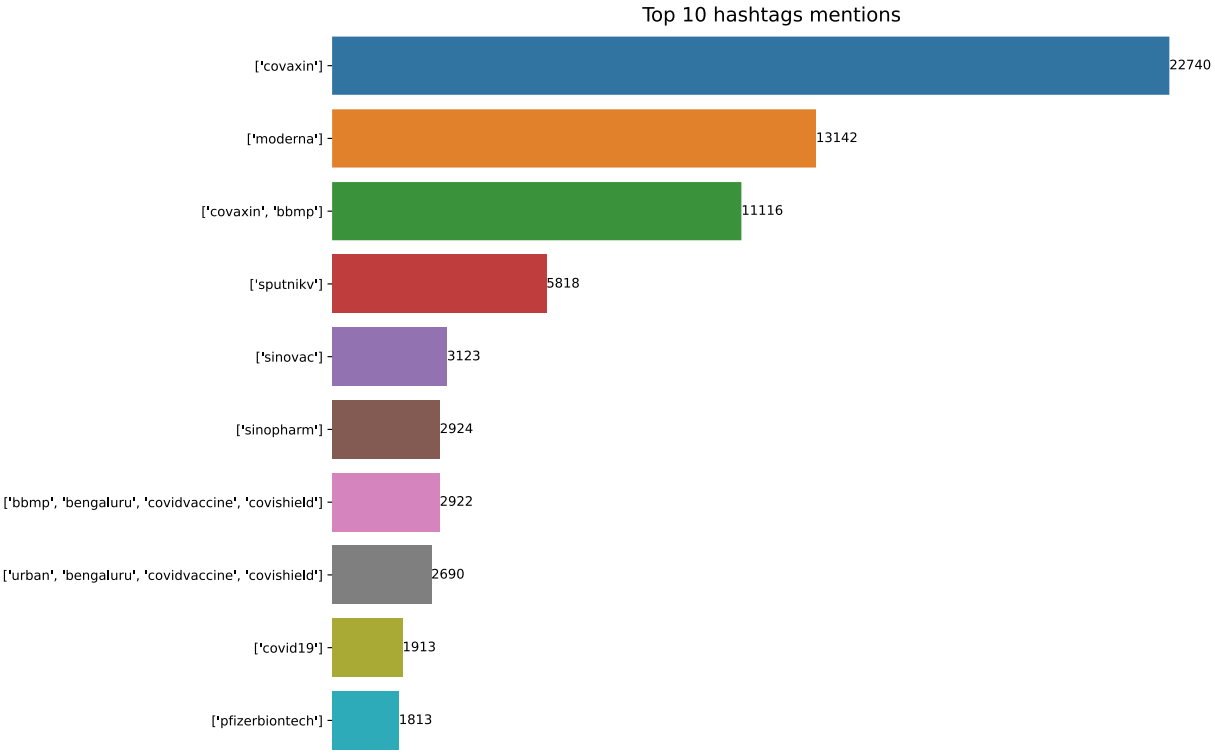

Figure 4.5. Top 10 hashtags mention

4.4 Data preparation

Tweets posted in the social media platform contains various junk characters, emojis, URL, etc. so a series of data preprocessing were followed as mentioned in section 3.2.2. As there are some duplicate records in the text field, retained only the first occurrence and the subsequent duplicates are discarded. Sentiments will be predicted on the text column so considered only the text field into a new data frame. The new data frame contains 226,373 records and one feature i.e. text.

4.4.1 Text preprocessing

The tweets text is converted to lowercase at first so that the identical words are considered as single. When the tweets are in multiple lines, they are imported in the text column with newline character (\n) representing a new line. So new line character is removed from the texts. Contractions found in the tweets are expanded to their base forms using the python package "contractions". Tweets contains the user mentions with the tag @ are eliminated from the text. Hashtags that start with # symbol are not relevant for our analysis so they are also discarded. URLs in the tweets are removed. Sometimes people express their opinion as emoji's as well so the emoji are converted to their meaningful words with the help of "demojize" function in "emoji" package. Then, special characters other than the alphabets and numbers are removed from the text field. Single characters, more than one whitespace, digits, leading and ending spaces are also removed from the text.

Once the noise in the data is removed, tokenization applied using "TweetTokenizer" to split the text. Stopwords are removed to focus on more relevant content. Stemming can lead to loss of meaning but lemmatization considers the meaning of the words. So, lemmatization performed using "WordNetLemmatizer" to reduce the words to their dictionary form. The text processing steps performed are listed with an example in *Table 4.3*.

Table 4.3. Example of text preprocessing

Text processing	**Example**
Lowercase conversion	The word 'NEWS' converted to 'news'.
Remove new line string	New line character (\n) removed.
Extend contractions	"It's" is expanded to "it is".
Removal of user tags	User mention "@ZubyMusic" removed.
Removal of hashtags	Hashtag "#PfizerBioNTech" removed.
Removal of URLs	URL "https://t.co/HkGTDM5J3f" removed.
Conversion of emoji	Emoji 🤔 converted to "thinking_face".
Removal of special characters	Characters like "-", "…" are removed.
Removal of single characters	Removed single characters "a", "w", etc.
Removal of spaces more than one	"fake news" replaced with "fake news".
Removal of digits	Numbers like "6" are removed.
Removal of leading and ending spaces	" due to process " replaced as "due to process".
Tokenization	"explain need vaccine" splatted into "explain", "need", "vaccine".

Removal of stopwords	Words like “to”, not” are removed.
Lemmatization	“facts” restated as “fact”

Finally, top ten common words found in the text are removed as they can skew the results. The top ten common words are shown in *Figure 4.6* and they are vaccine, dose, slot, age, covaxin, covid, got, first, hospital and pincode.

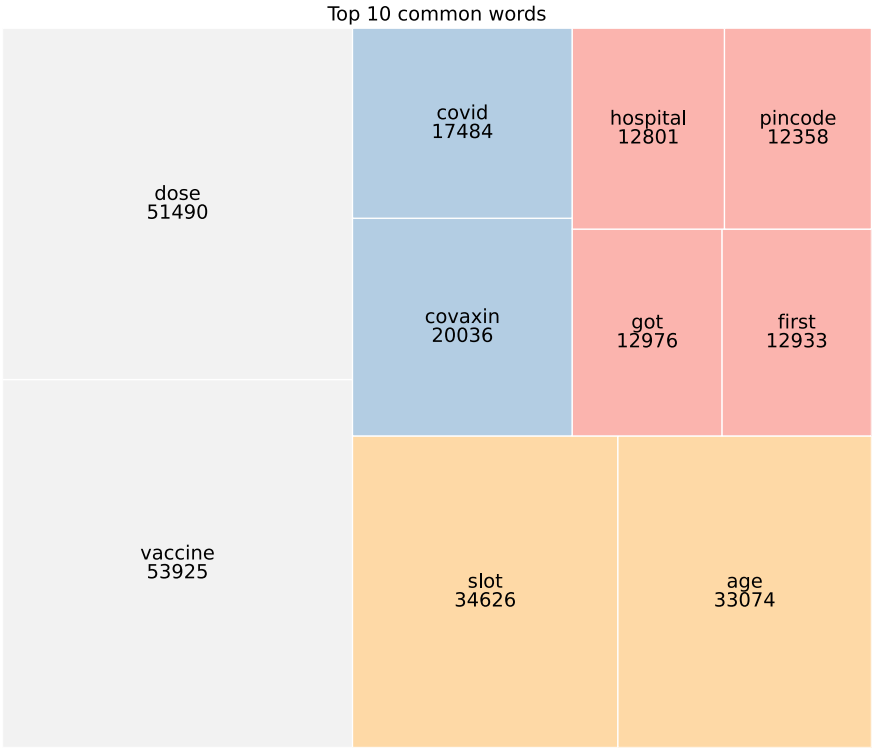

Figure 4.6. Top 10 common words

Few examples of the tweets before and after preprocessing is show in *Table 4.4.*

Table 4.4. Tweets before and after preprocessing

Tweets before processing	**Tweets after processing**
Same folks said daikon paste could treat a cytokine storm #PfizerBioNTech https://t.co/xeHhIMg1kF	folk said daikon paste could treat cytokine storm

#coronavirus #SputnikV #AstraZeneca #PfizerBioNTech #Moderna #Covid_19 Russian vaccine is created to last 2-4 years… https://t.co/ieYlCKBr8P	russian created last year
#ICYMI The #FDA Authorized the #PfizerBioNTech #COVID19 #Vaccine for the United States last night\n\nhttps://t.co/CtYGB3fNnE	authorized united states last night

4.5 Sentiment classification using VADER

The next step after the text cleaning is to assign a sentiment based on the tweets. Python package "SentimentIntensityAnalyzer" from "vaderSentiment" library was used to assign the sentiment intensity scores for each tweets text. Sentiments (positive, negative or neutral) assigned based on the compound score in the intensity score as described in section 3.2.3. An example of sentiments classified using VADER is shown in *Figure 4.7*. The first tweet has the compound score of 0.4019 (>=0.05) so 'Positive' sentiment assigned to it. Similarly, the second and third tweets have compound scores as -0.3182 (<=0.05) and 0 (>-0.05 and <0.05) so they are assigned with 'Negative' and 'Neutral' sentiments respectively.

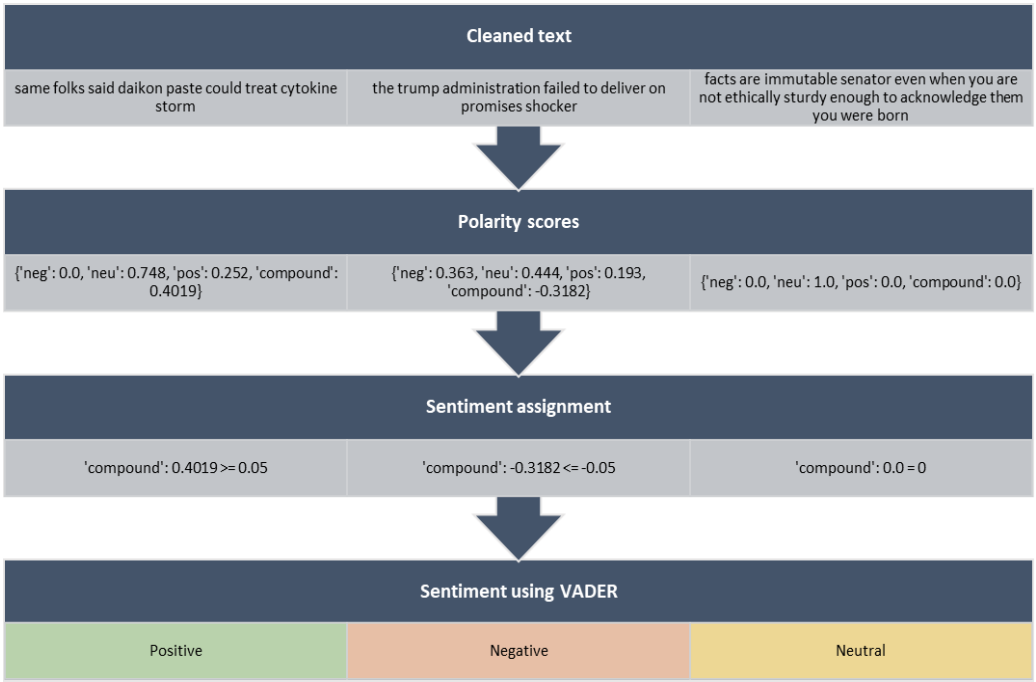

Figure 4.7. Sentiment classification using VADER

The distribution of sentiments predicted by VADER is visualized in *Figure 4.8*. Among the tweets in the dataset, approximately 46% are classified as Neutral sentiment, 40% are as Positive sentiment and 15% of tweets as Negative sentiment.

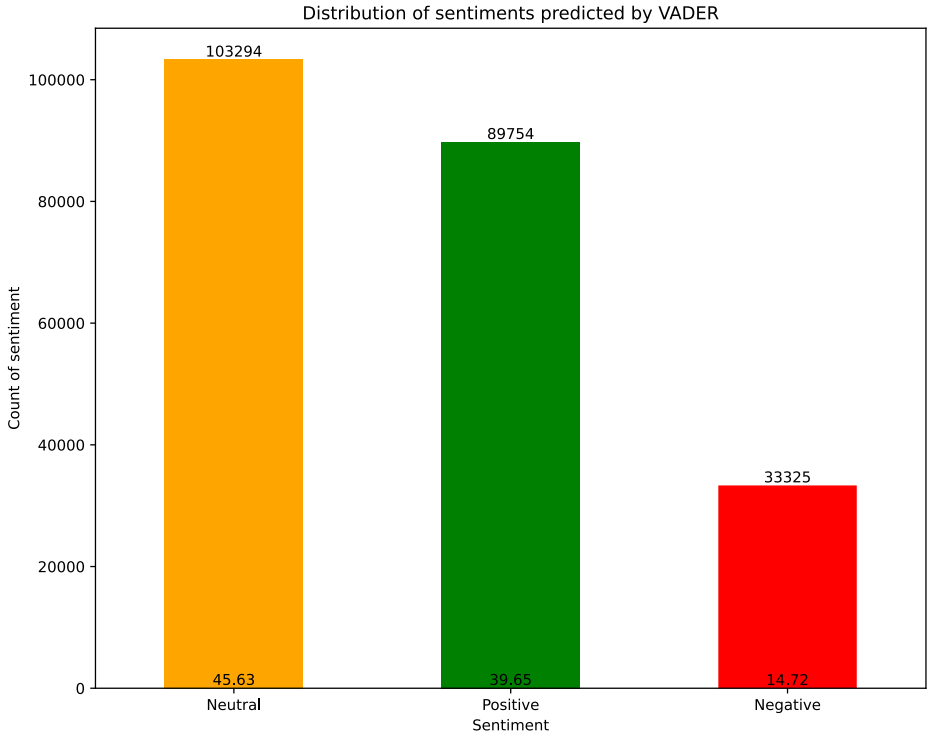

Figure 4.8. Distribution of sentiments by VADER

4.6 Word clouds

Wordclouds are plotted to identify the key topics discussed in the tweets. *Figure 4.9* shows the Wordcloud before and after text cleaning. The words “https”, “COVAXIN”, “vaccine”, etc. are the top highlighted words in the tweets text before preprocessing and the words “fee”, “free”, “shot”, etc. are the top words after text cleaning.

Figure 4.9. Wordclouds of text

Figure 4.10 shows the wordcloud depicted for each sentiment. The words like “free”, “availability”, “face”, etc. are most used in Positive tweets. Likewise, “date”, “fee”, “apollo”, etc. are top highlighted in Neutral sentiment and “shot”, “day”, “emergency”, etc. are frequent words in Negative sentiment.

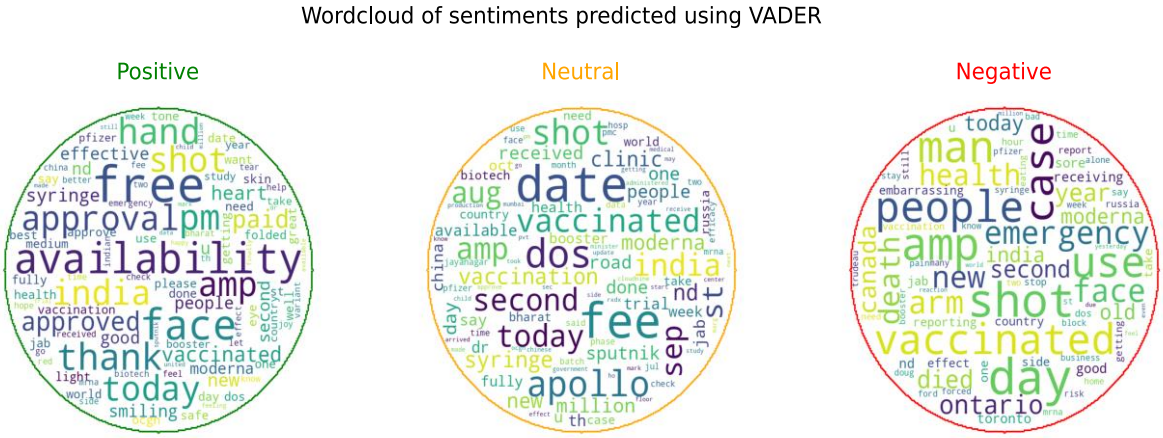

Figure 4.10. Wordclouds of sentiments by VADER

4.7 Resampling technique and Data split

In this section, the resampling technique performed as mentioned in 3.2.6 are discussed. The models require numerical label, so the sentiments are mapped as 0 for Positive, 1 for Negative and 2 for Neutral. As shown in *Figure 4.8*, the dataset shows imbalance as the negative sentiment being less compared to the other two sentiments. So, to handle the dataset

imbalance and hardware resources limitation, bootstrapping was performed before splitting it for train and test.

Bootstrap resampling with random replacement of samples is used to get specified number of samples for each sentiment from the dataset. This approach ensures that the training and testing of the model is performed on the balanced data. Then, the cleaned text variable is assigned to X, and the sentiment label variable to y. Later, the X and y data are splatted into train and test, with 30% kept for test. Finally, the train data is further divided into train and validation subsets, with 25% assigned for validation.

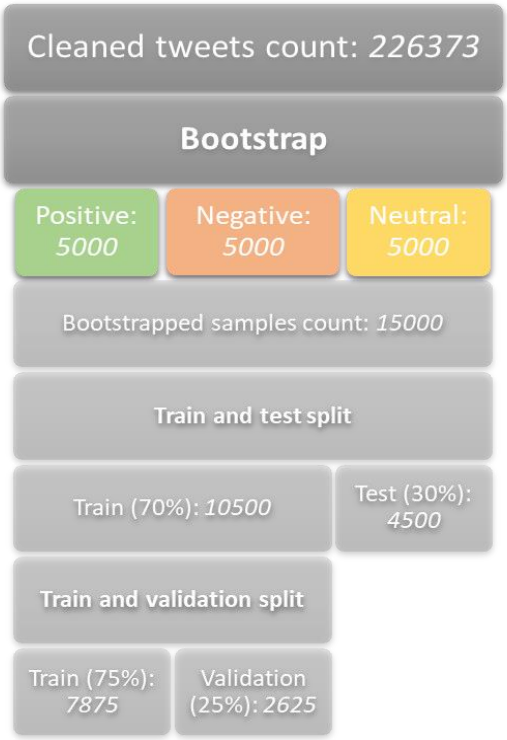

Figure 4.11. Resampling and data split flow

The *Figure 4.11* illustrates the steps performed for resampling and data split. The cleaned tweets dataset consists of 226,373 records. From each sentiment category, random samples of 5,000 records were selected, resulting in a new data frame with 15,000 records. This resampled data frame was then split into 10,500 training records and 4,500 test records. Finally, initial training set of 10,500 records was further divided into 7,875 for training and 2,625 for validation.

4.8 Sentiment Analysis using Deep learning models

The resampled data is then passed to deep learning models to identify the sentiments. In this research, two deep learning models BERT and Bi-LSTM with BERT word embeddings are designed for the sentiment prediction task.

4.8.1 BERT

The "BertTokenizer" from Hugging Face Transformers library (Hugging Face, 2024), using a pretrained BERT base uncased model is used to tokenize the given tweets into integer sequences. In this tokenizer, padding enabled to add special padding tokens and truncation is used to limit the maximum length of the sentence as specified. The tokenized inputs are returned from this tokenizer as tensors. As shown in *Table 4.5* the input text is passed to the BERT tokenizer and it returns text's input IDs, decoded IDs with special tokens and the attention mask.

Table 4.5. An example of BERTTokenizer output

Cleaned text	wait watch look like bullish
Input IDs of the text	[101 3524 3422 2298 2066 7087 4509 102 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Decode IDs of the text	[CLS] wait watch look like bullish [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
Attention mask of the text	[1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

Another powerful model "TFBertForSequenceClassification" from Hugging Face Transformers library (Hugging Face, 2024) is used to perform the tweets classification. It is a fine-tuned BERT model transformer with a sequence classification head on top. This model is initialized with a pre-trained BERT base uncased model and passed three labels as there are three sentiments i.e. positive (0), negative (1) and neutral (2). During the training, the model fitted using train and validation data with 5 epochs and 32 batch size. The model is compiled with hyperparameters such as Adam optimizer with learning rate of 2e-5 and loss function "SparseCategoricalCrossentropy" to compute the cross-entropy loss between the predicted

logits and the actual class labels. Metric “SparseCategoricalAccuracy” used to evaluate the performance of the model.

Input IDs contains tokenized representation of the text, Token type IDs indicates the segmented token to specify different sentences and attention mask identifies the tokens that the model should pay attention to. These three components from the BERT tokenizer are passed to the BERT sequence classification model for training. Finally, the model is evaluated and predicted on the test data and the test accuracy and loss are captured. The output of the model is a tensor that contains probabilities or logits for each sentiment label. The argmax from NumPy library used to find the index of the highest confidence value for each input. The index links to the predicted sentiment label and they are mapped to the appropriate sentiments (i.e. 0 for Positive, 1 for Negative and 2 for Neutral). The accuracies and losses of train and validation dataset are plotted for each epoch. Also, the precision, recall and F1-score of each sentiment is visualized. The models can be further fine tuned with different batch size or epochs to even get a better accuracy. Also, the model can be tested for more sample sets or the complete dataset after oversampling method.

4.8.2 Bi-LSTM with BERT word embeddings

As discussed in 3.2.4.3, the tweets text is converted to word embeddings and then passed to Bi-LSTM layers to perform the sentiment classification. The model “AutoTokenizer” from Hugging Face Transformers library (Hugging Face, 2024) used along with pretrained BERT base uncased model to tokenize and encode the tweets. The tokenizer is configured with parameters such as padding, truncation, and adding special tokens are set to true. This ensures that input texts are padded, truncated to their maximum length, and the special tokens like [CLS] and [SEP] are included accordingly. The tokenizer is also set to return PyTorch tensors. An example of encoded text using the tokenizer is shown in *Table 4.6*.

Table 4.6. An example of AutoTokenizer output

Cleaned text	news eua family waiting
Tokenized text	['news', 'eu', '##a', 'family', 'waiting']
Encoded text	[101, 2739, 7327, 2050, 2155, 3403, 102]

The model “AutoModel” from Hugging Face Transformers library (Hugging Face, 2024) used along with pretrained BERT base uncased model to generate word embeddings. The Input IDs and attention mask from the tokenizer are passed to the model and it returns the word embeddings. Then the word embeddings are converted to NumPy arrays and target variable sentiment labels are converted to arrays using “OneHotEncoder”.

A sequence model with a bidirectional LSTM layer has been created. This bidirectional model creates two LSTM, one for processing the input sequence in the forward direction and the other in backward direction. The number of neurons and input shape of the LSTM layer is matches the shape of the word embeddings. To avoid overfitting, L2 regularization applied to the kernel and the recurrent weights. A dropout layer is added after the Bi-LSTM layer to randomly set a fraction of input units to 0 during training. A fully connected dense layer with three output units is added. The SoftMax activation function is used to convert the output into probabilities for each class. The model compiled with the parameters Optimizer set to Adam, loss function to categorical cross-entropy and accuracy for evaluation during training.

Once the model was created, it was trained using corresponding word embeddings and encoded target labels. The model iterates for five epochs with 16 batch size. Also, callback function used for early stopping to avoid over fitting. The hyperparameters of the model can be fine-tuned to get better model metrics and also more training data can be used instead of limited records.

4.9 Model Execution

Due to computational cost and time constraints, Paperspace’s GPU infrastructure was used to execute both the BERT and Bi-LSTM models. These two models are exclusively trained and tested using the sample data as specified in section 4.7. These models are executed five times using resampling with replacement of samples to get a better model metrics. The BERT model approximately took 40 seconds for training and validation in all the five iterations. The Bi-LSTM with BERT model took more time for generating the word embeddings but it takes less than ten seconds for the model training and validation. For each iteration, metrics such as accuracy and loss are recorded. Finally, the average of accuracy is calculated to arrive at the final accuracy of the models.

4.10 Summary

To summarize, the dataset contains 228,207 records with 16 features including tweets. The tweets were cleaned and preprocessed. Visualizations such as top ten users by tweet count, comparison between verified and non-verified users, maximum tweets per day and the distribution of word count in the tweets were created to understand the dataset. Text cleaning, tokenization, stop words removal and lemmatization were applied to the tweets. The top ten most frequent words were also removed from the tweets. Sentiment analysis was conducted using the VADER tool that provides polarity score and sentiments Positive, Negative and Neutral assigned based on the score. Bootstrap resampling was used to address the issue of an unbalanced dataset. Finally, sentiment classification was performed using two deep learning models BERT and Bi-LSTM with BERT word embeddings. The models were executed on five sets of resampled samples and the average accuracy was calculated for each model.

CHAPTER 5

RESULSTS AND DISCUSSIONS

5.1 Introduction

This chapter describes the sentiment analysis results of the approaches lexicon and deep learning. In the lexicon approach, the classified sentiments are visualized in WordClouds with the common words used. The results from the two models BERT and Bi-LSTM with BERT are presented in the deep learning approach. The models train and test executions for each iteration are visualized. The performance of these two models is compared and discussed the better model.

5.2 Lexicon sentiment analysis

VADER sentiment analysis was used to identify the intensity scores of each preprocessed tweet. Figure *4.8* shows the distribution of sentiments predicted by VADER. The majority of the tweets related to COVID-19 vaccine are categorized as Neutral sentiment followed by Positive. Only 15% of the tweets in the dataset are Negative sentiment, which implies very less people have negative thoughts about the vaccine and most of the users have mixed feeling between neutral and positive. An example of tweet and their sentiment classified by VADER tool is listed in *Table 5.1*.

Table 5.1. Sentiment by VADER

Tweets	**Sentiment by VADER**
#coronavirus #SputnikV #AstraZeneca #PfizerBioNTech #Moderna #Covid_19 Russian vaccine is created to last 2-4 years… https://t.co/ieYlCKBr8P	Positive
Coronavirus: Iran reports 8,201 new cases, 221 deaths in the last 24 hours #Iran #coronavirus #PfizerBioNTech… https://t.co/mwDNAdmb7F	Negative
Facts are immutable, Senator, even when	Neutral

you're not ethically sturdy enough to acknowledge them. (1) You were born i… https://t.co/jqgV18kch4	

WordClouds has been plotted using the tweets from each sentiment that shows the common words used in the tweets. Some of the important words in the sentiments highlighted by wordclouds in *Figure 4.10* are given below:

- The words such as “free”, “availability”, “approval”, “thank”, “effective” denotes positive attitude of the public about the vaccine.
- The words like “emergency”, “embarrassing”, “death”, “died”, “forced” are considered as negative and they denote the negative feedback of the public about the vaccine.
- The words such as “fee”, “date”, “received”, “syringe”, “booster” are marked as neutral that implies the sentiment related with these words are neither positive nor negative.

5.3 Deep learning sentiment analysis

Two deep learning models BERT and Bi-LSTM with BERT were developed for sentiment prediction and classification. These models were fine-tuned and executed for five iterations using different set of samples with same size. The outcome of these models is discussed in the below sub sections.

5.3.1 BERT

The initial training run of the BERT model is shown in *Figure 5.1*. Throughout the epochs, the model’s training accuracy steadily improved from 73% to 98% while the validation accuracy varied between 86% and 92%. Likewise, the training loss gradually decreased from 0.63 to 0.06 but the validation loss continued moderately in the range of 0.38 to 0.29. The test accuracy model for the first set of samples achieved 93% with a loss of 0.28. The model performance for the first set of samples was reasonable as the accuracy increased and the loss decreased.

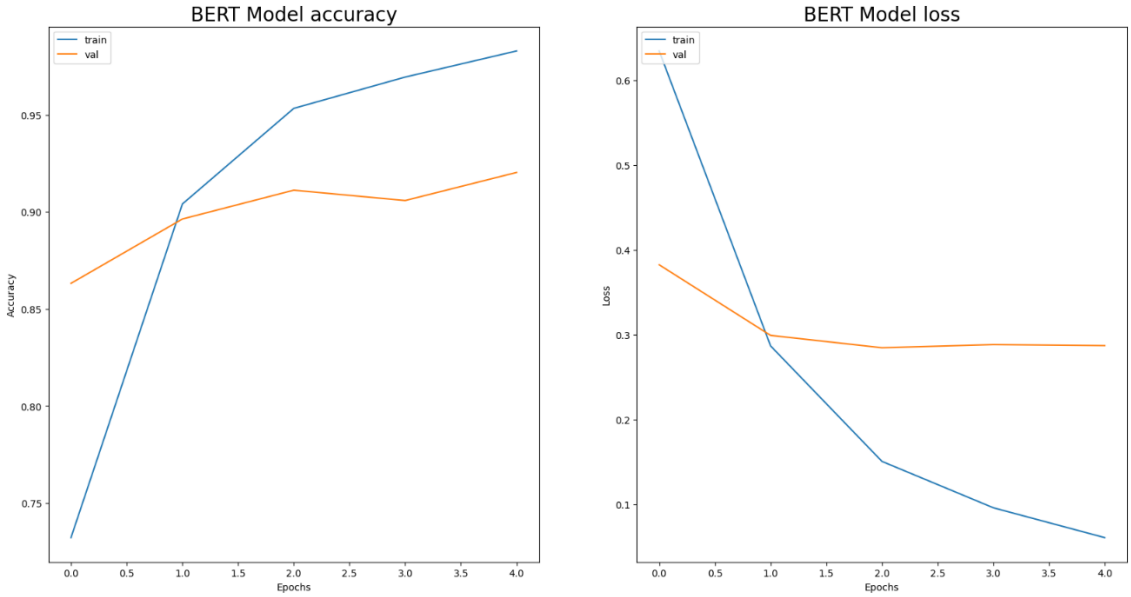

Figure 5.1. BERT first sample set training metrics

The BERT model ran again for second set of samples and the trend of accuracy and loss is shown in *Figure 5.2*. The training accuracy consequently increased from 93% to 99% and the validation accuracy was in the range of 94% to 95%. The training loss was started with 0.22 and ended in 0.03 while the validation loss was considerably increased from 0.17 to 0.25. The test accuracy and loss of the second sample set was 95% and 0.19 individually.

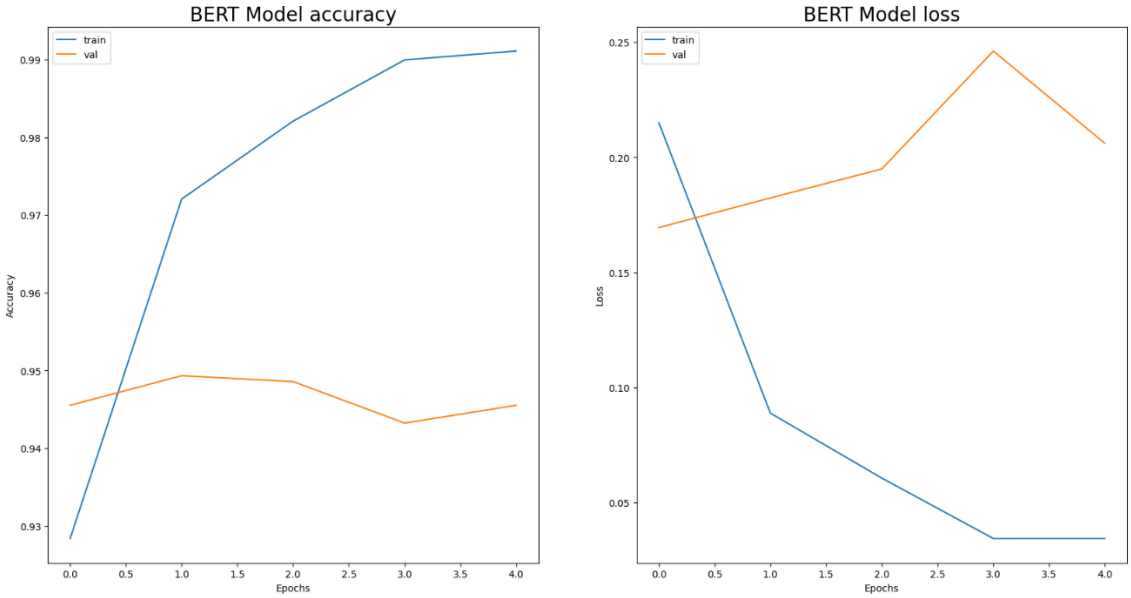

Figure 5.2. BERT second sample set training metrics

The training accuracy and loss of the next set of samples ranged from 95% to 99% and from 0.14 to 0.03 respectively. Validation accuracy was around 96% across the epochs and loss was in increase trend. *Figure 5.3* illustrates the model's training and validation accuracy and loss of third set of samples.

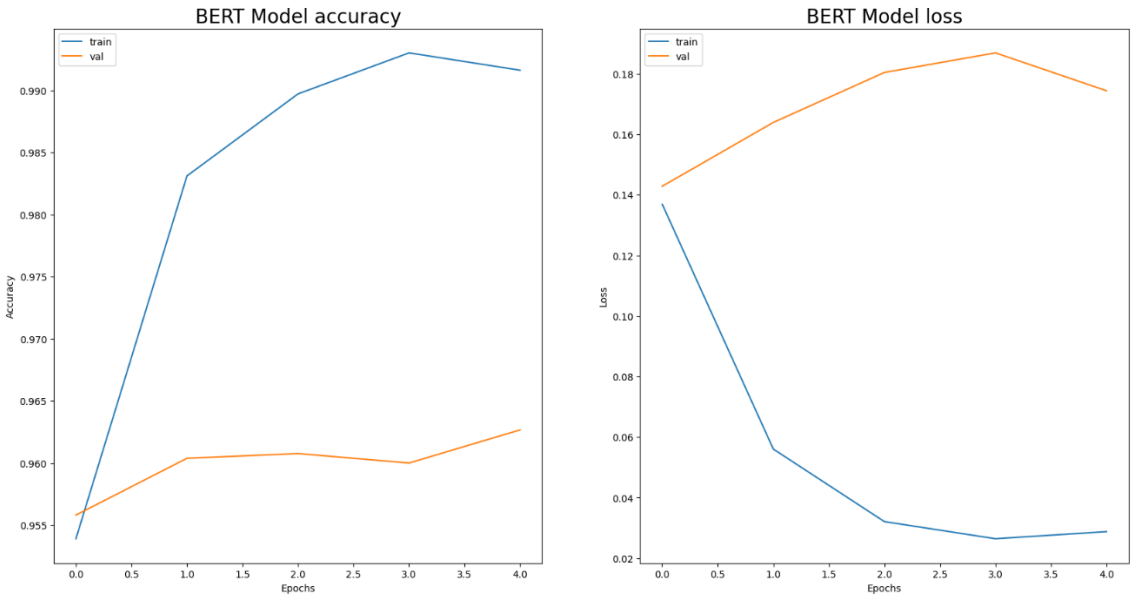

Figure 5.3. BERT third sample set training metrics

Similarly, *Figure 5.4* and *Figure 5.5* represent the model's training and validation accuracies and losses of the fourth and fifth set of samples. In the fourth run, the training accuracy ranged from 96% to 99.5% and validation accuracy was around 96%. The loss of training decreased to 0.02 from 0.1 whereas validation loss was increased to 0.14 from 0.17. In the final run, the accuracy of training was in upward direction from 96% to 99.4% and validation accuracies laid around 97%. The loss of training decreased to 0.02 and validation loss was fluctuated from 0.09 to 0.13. The test accuracies of the third, fourth and fifth sample sets are 96%, 96% and 97% respectively whereas the loss of tests are 0.15, 0.16, 0.15.

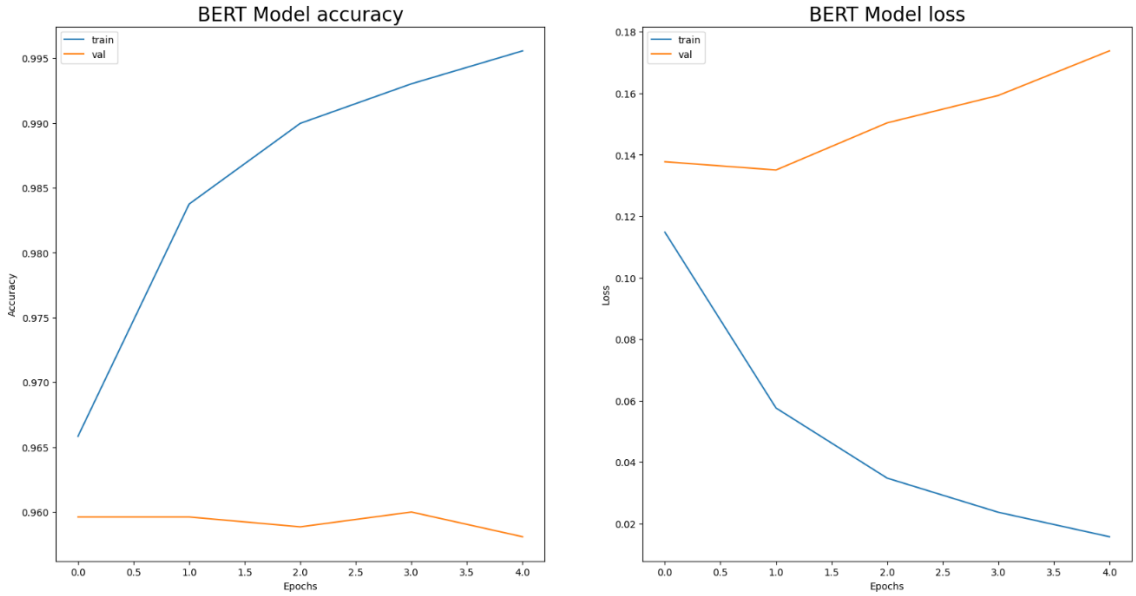

Figure 5.4. BERT fourth sample set training metrics

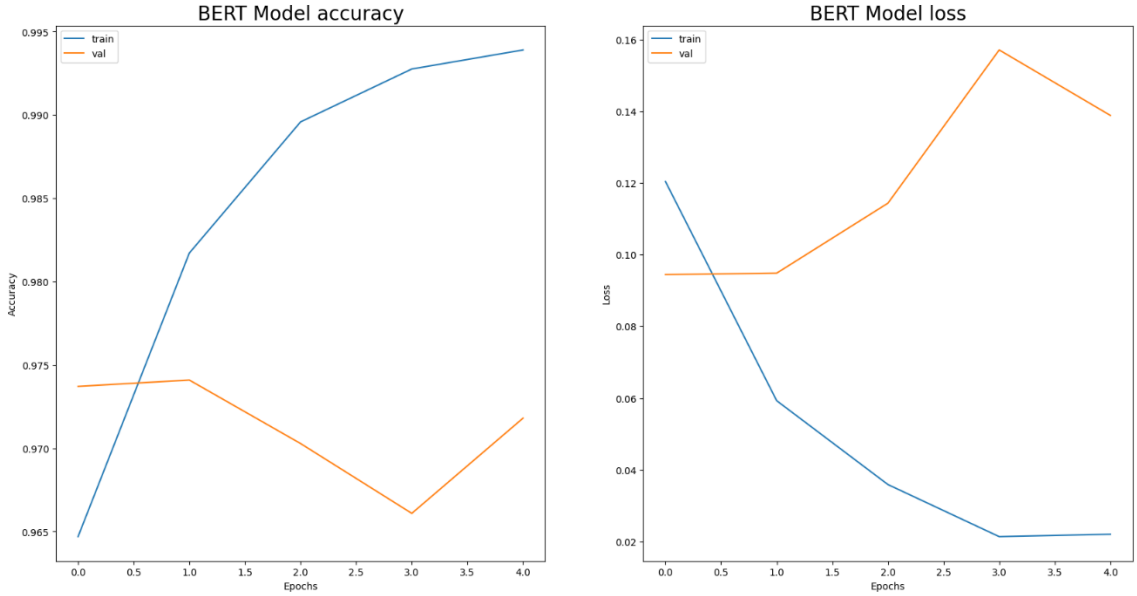

Figure 5.5. BERT fifth sample set training metrics

5.3.2 Bi-LSTM with BERT

The Bi-LSTM with BERT model for first set of samples resulted a decent accuracy range from 65% to 71% for training data and for validation data the accuracy fluctuated between 67% and 70%. The training and validation losses are similar as both reached a minimum of 0.85. *Figure 5.6* illustrates the trend of the model's training and validation accuracy and loss. The model scored 68% test accuracy with a loss of 0.86.

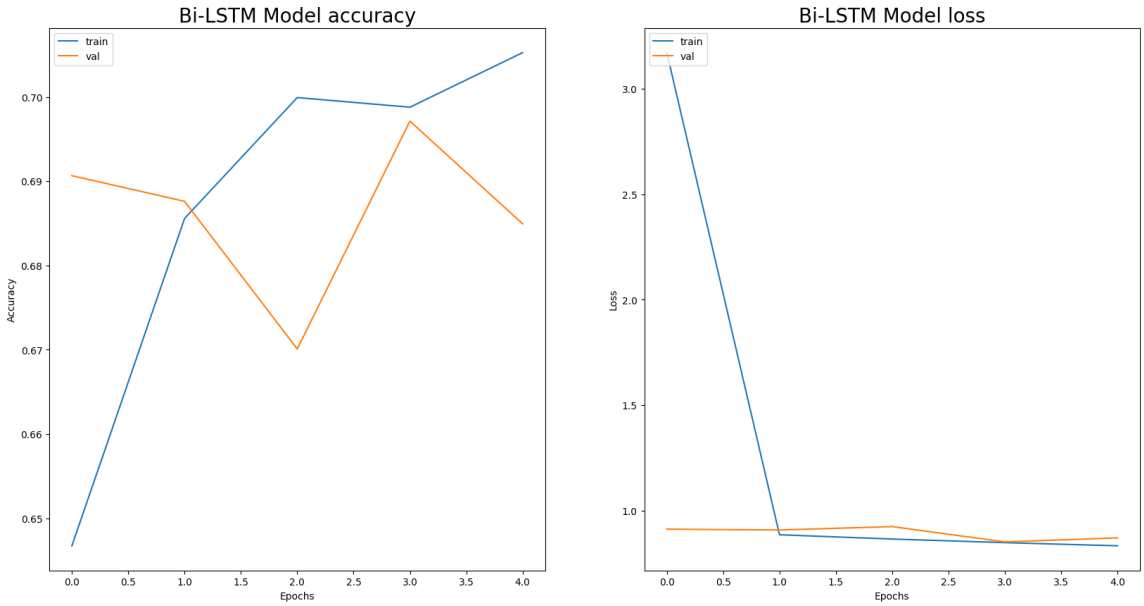

Figure 5.6. Bi-LSTM first sample set training metrics

Similarly, the accuracy of the training and validation data for the second set of samples are varied from 64% to 70% and 68% to 72% respectively. The loss of the training reduced from 3.75 to 0.86 while the validation loss was from 0.92 to 0.82. The trend of accuracy and loss is shown in *Figure 5.7*. The test accuracy and loss of the second iteration was 70% and 0.83.

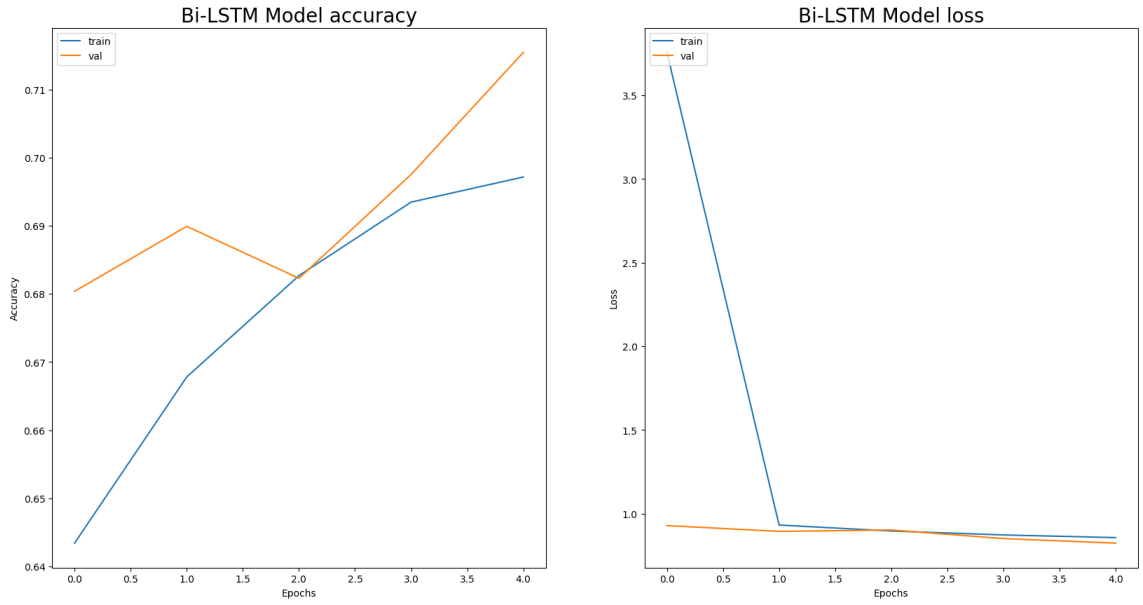

Figure 5.7. Bi-LSTM second sample set training metrics

The third, fourth and fifth iterations resulted with training accuracies from 65% to 71%, 63% to 69% and 64% to 70% respectively. Likewise, the losses of the three iterations for training and validation data are decreased consequently from 0.94 to 0.83, 0.87 to 0.82 and 0.91 to 0.88. The trends of these iterations are shown in *Figure 5.8*, *Figure 5.9* and *Figure 5.10*. The test accuracy achieved by third iteration was 71%, fourth iteration was 66% and fifth iteration was 67% with loss of 0.83, 0.9 and 0.88 respectively.

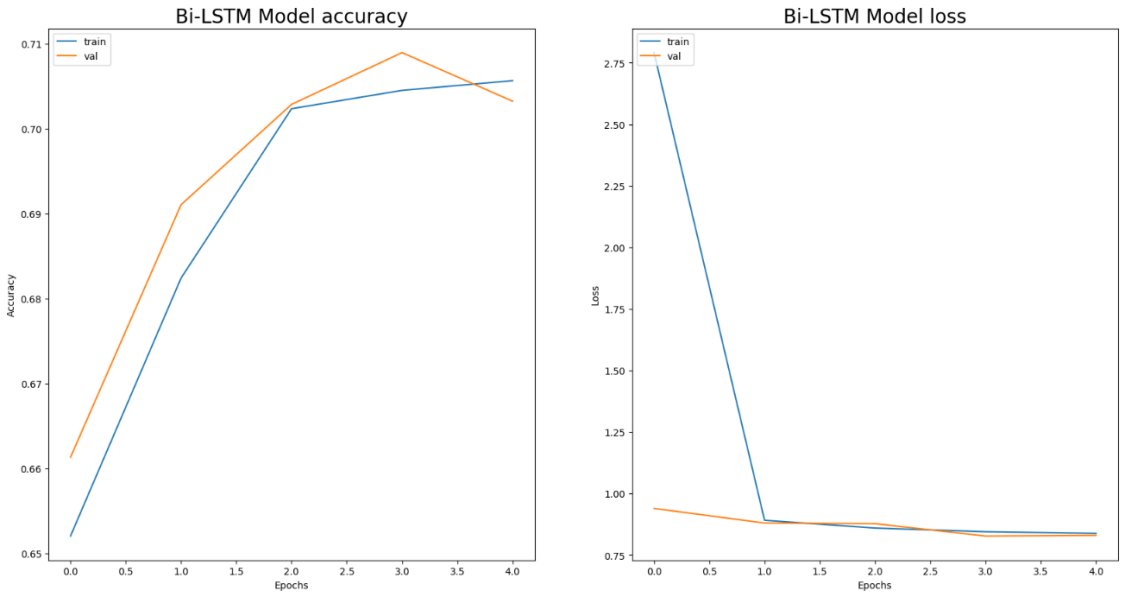

Figure 5.8. Bi-LSTM third sample set training metrics

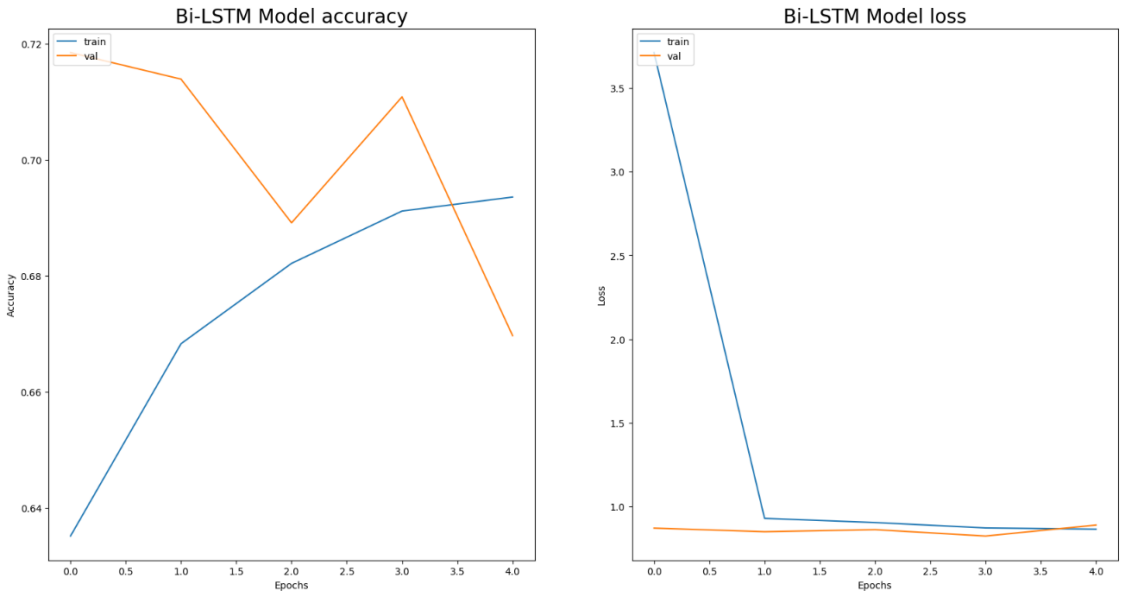

Figure 5.9. Bi-LSTM fourth sample set training metrics

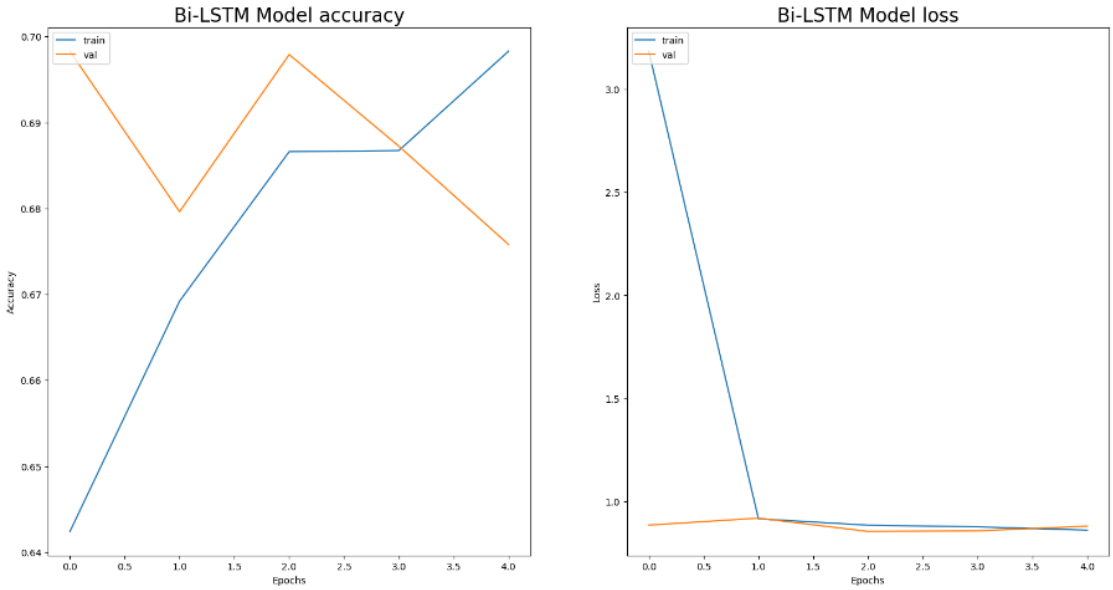

Figure 5.10. Bi-LSTM first sample set training metrics

5.4 Comparison of the models BERT and Bi-LSTM

The predicted sentiment of the models BERT and Bi-LSTM with BERT are evaluated using classification report. The precision, recall and f1-score of these models were plotted for each sentiment for all the five executions. The BERT model has precision, recall and f1-score above 92% across all the three sentiments. On the other side, Bi-LSTM with BERT model has some fluctuations. The recall (81%) was high for neutral sentiment and the precision (79%) was high for positive sentiments. The model captured most of the neutral sentiment correctly for most of the samples but missed many correct predictions for positive sentiment. The negative sentiment predicted well as the three metrics are close to each other. The classification report of first sample set is shown in *Figure 5.11*.

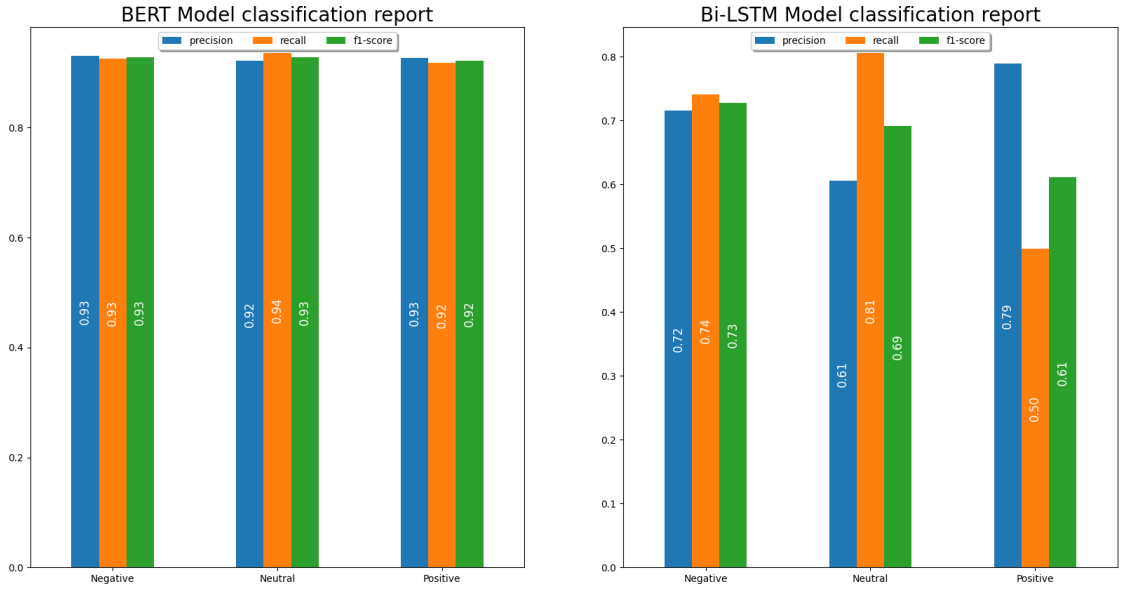

Figure 5.11. Classification report for first iteration

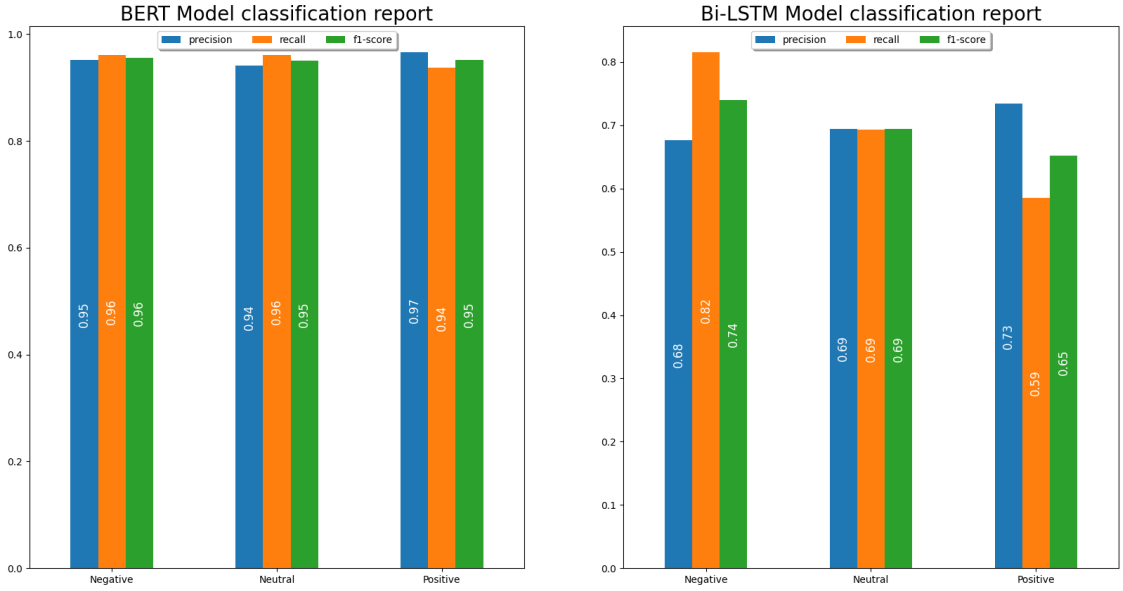

Figure 5.12. Classification report for second iteration

As shown in *Figure 5.12*, the BERT model performed well for the second run that the precision, recall and f1-score across the three sentiments are above 94%. Bi-LSTM with BERT model had high recall (82%) for negative sentiment indicating more correct predictions. High precision for positive (73%) sentiment shows that the model was accurate for it. Neutral sentiment had an equal metric score of 69%. The precision, recall and F1-score metrics for all three sentiments of the BERT model applied to the third set of samples are almost same as shown in *Figure 5.13*. On the other side, the Bi-LSTM with BERT model

shown high recall (85%) and low precision (66%) for negative sentiment. For the neutral and positive sentiments, the model achieved precision (74% and 75%) but recorded low recall (65% and 62%).

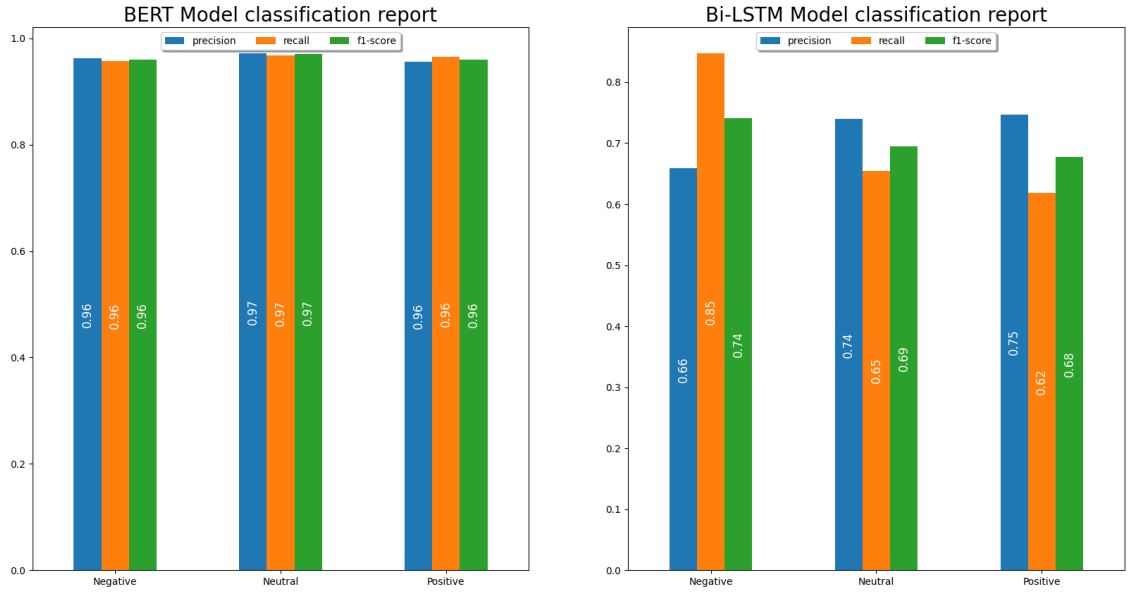

Figure 5.13. Classification report for third iteration

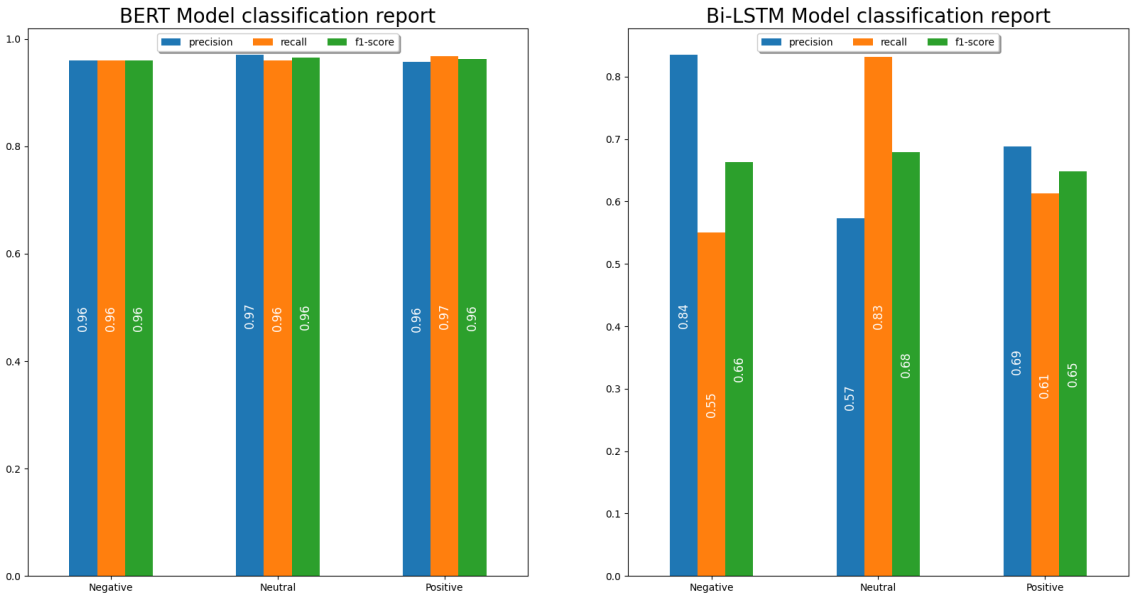

Figure 5.14. Classification report for fourth iteration

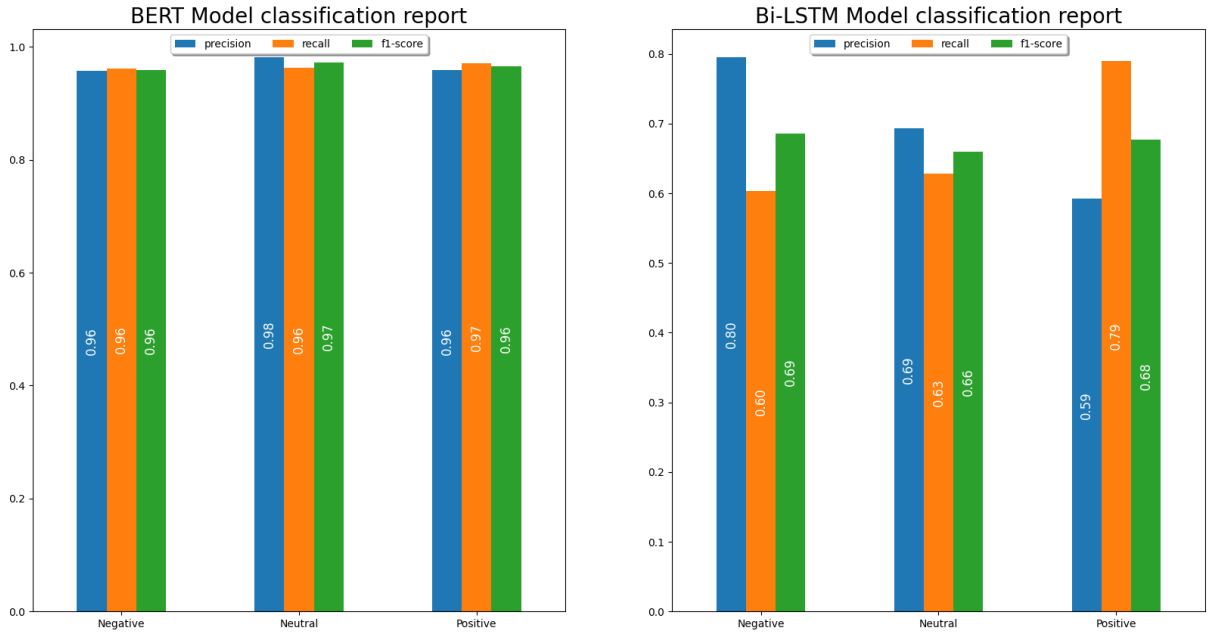

Figure 5.15. Classification report for fifth iteration

The BERT model was performed well in both the fourth and fifth runs as the metrics precision, recall and F1-score are 96% and above across all three sentiments. The report for the fourth and fifth iterations is shown in *Figure 5.14* and *Figure 5.15*. In the fourth iteration, the Bi-LSTM with BERT model showed higher recall (83%) but lower precision (57%) for neutral sentiment. Negative and positive sentiment proved high precision (84% and 69%) but their recall was lower (55% and 61%). But in the fifth iteration, positive sentiment had high recall of 79% and low precision of 59%. The other two sentiments negative and neutral have higher precision (80% and 69%) but lower recall (60% and 63%).

The test accuracy and loss for all five iterations were plotted and *Figure 5.16* represents the same. The BERT model displays a gradual increase in test accuracies from 90% to 96% while the losses decreased from 0.4 to 0.2. Unlike that, the Bi-LSTM with BERT model had lower test accuracies ranging from 67% to 70% and test loss across the runs were around 0.9. The test accuracies of BERT and Bi-LSTM with BERT for each sample sets are shown in Table *5.2*. So, the BERT model outperformed the Bi-LSTM with BERT model. The overall test accuracies for the models shown in *Table 5.3*.

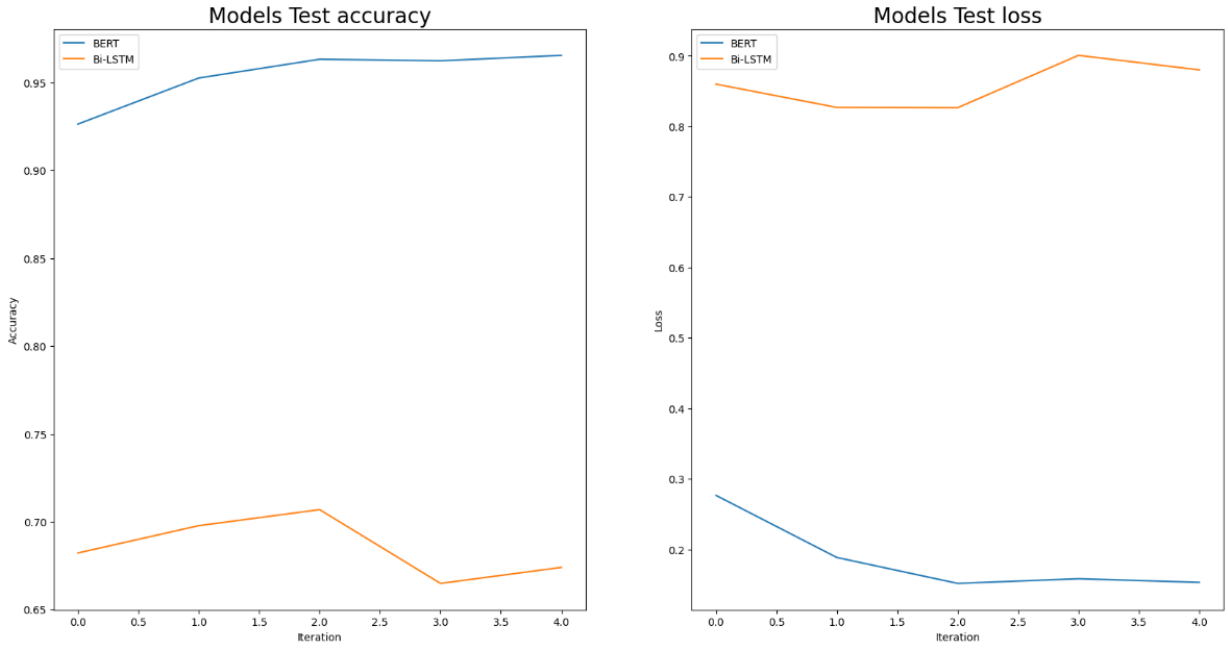

Figure 5.16. Test accuracy of the iterations

Table 5.2. Models test accuracy for each sample sets

Sample set	**BERT test accuracy**	**B-LSTM with BERT test accuracy**
Sample 1	90.22%	66.78%
Sample 2	93.89%	70%
Sample 3	94.33%	67.81%
Sample 4	95.29%	69.22%
Sample 5	95.74%	68.78%

Table 5.3. Models mean test accuracy

BERT – Mean test accuracy	**Bi-LSTM with BERT – Mean test accuracy**
93.9%	68.52%

5.5 Discussion

The majority of the sentiments identified by the VADER tool fall into the neutral category. People had neither positive or negative thoughts initially as the pandemic, lockdown and

vaccination were occurred in the modern era. The health authorities can make use of this trend by increasing vaccine campaigns to convert all these neutral sentiment into positive. The words “emergency”, “case”, “died”, “death”, “embarrassing”, etc. are more prevalent in the negative sentiment so the authorities can find the truth in these tweets and take necessitate actions to mitigate them. Positive sentiments are higher than the negative sentiment that indicates many people recognize that the vaccine could fight against the virus. VADER might have misclassified few tweets because of missing preprocessing steps such as lemmatization without considering POS tagging or tweets written in languages other than English but using English characters. Some of the discrepancies between the context of tweet and its sentiment class as explained in Table *5.4*.

Table 5.4. Misclassified tweets by VADER

Tweets	**Cleaned tweets**	**Sentiment**	**Comments**
Does anyone have any useful advice/guidance for whether the COVID vaccine is safe whilst breastfeeding?… https://t.co/EifsyQoeKN	anyone useful advice guidance whether safe whilst breastfeeding	Positive	Though this seems to be a question, VADER misinterpreted it as positive.
While the world has been on the wrong side of history this year, hopefully, the biggest vaccination effort we've ev… https://t.co/dlCHrZjkhm	world wrong side history year hopefully biggest vaccination effort ev	Negative	Extended version of this tweet could convey a different sentiment, possible positive.
Explain to me again why we need a vaccine @BorisJohnson @MattHancock #whereareallthesickpeople #PfizerBioNTech…	explain need	Neutral	Context of this tweet seems to be negative but it is classified as neutral by VADER.

https://t.co/KxbSRoBEHq			

A study of (Umair and Masciari, 2023) achieved 55%, 69% and 58% of precision, recall and F1-score for positive sentiment and 54%, 85% and 64% for negative sentiment. The BERT model designed in this research outperformed the other study in terms of 93.9% accuracy, more than 96% of precision, recall and F1-score. In another research (Qorib et al., 2023a), BERT achieved 96.71% of accuracy and LSTM model attained 89.93% of accuracy for ten sentiment classes which beats the developed BERT and Bi-LSTM with BERT models with three classes. An ensemble model LSTM-2BiGRU proposed by (Said et al., 2023), had an accuracy of 92.46% which is lower than BERT model accuracy 93.9% but higher than the Bi-LSTM with BERT model accuracy 68.52%. In the research of detecting misinformation in COVID-19 vaccine tweets, (Hayawi et al., 2022) got improved results as they obtained accuracy of 98%, precision 97%, recall 98% and F1-score 98% for the BERT model. The fine-tuned BERT model proved greater performance compared to other ML, hybrid or ensemble models in similar research.

There are few challenges in this study that refined preprocessing steps to achieve better VADER sentiment classification. The deep learning model BERT can be executed with more training data for improved metrics. The hybrid model Bi-LSTM with BERT can be fine-tuned with different hyperparameters to achieve the state-of-the-art results. Also, this approach can be executed on real-time tweets to focus on the latest topic.

5.6 Summary

The performance of the BERT and Bi-LSTM with BERT models were discussed comprehensively. The BERT model demonstrated strong performance across all scenarios and achieved high train and test accuracy scores. Furthermore, its precision, recall and F1-score executed well compared with the Bi-LSTM with BERT model. The validation loss of the Bi-LSTM with BERT model was very close to the training loss across the runs. The model's metrics are compared with few other research metrics. There are few challenges in the designed model and that can be addressed in the future research to get an improved version.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

In this chapter, the detailed explanation of the conducted research has been discussed. The assumptions and limitations of the research was also addressed in the subsequent sections. Furthermore, this research can also provide valuable insights to governing authorities about the common keywords used in the tweets. Finally, the reference for future researchers has been provided to improve the study or use the approach for similar problems.

6.2 Discussion and Conclusion

Since late 2019, the COVID-19 epidemic has been a crucial topic of discussion across various social media platforms. Twitter is one of the popular social media platforms where the people share their thoughts through tweets. After the introduction of COVID-19 vaccines, there has been hesitancy due to concerns about safety and potential side effects. So, understanding the public perception through COVID-19 vaccine related tweets became crucial for addressing similar situations in the future.

This study meant to explore the people's sentiment expressed on Twitter regarding COVID-19 vaccines. There were several previous literatures analyzed the sentiments in the COVID-19 vaccine related tweets but only few studies used lexicon approach, deep learning and hybrid methods. This research initiated from the idea that integrating the approaches could provide more accurate predictions.

The freely available COVID-19 vaccine related tweets that were posted between December 2020 to November 2021 used for this research. The tweets were cleaned and preprocessed for the better model prediction. A lexicon-based model, VADER was used to classify the tweets into three main classes i.e. positive, neutral or negative. The common sentiments among the tweets were neutral and only 15% of tweets projected as Negative. This promises that only small number of people were against COVID-19 vaccine and majority of people have mixed feeling. WordClouds has been plotted based on the sentiments to visualize the common words

use in the tweets. It is observed the words such as “emergency”, “death”, “died” were used in the tweets classified as negative. This shows that people are concerned about the vaccine. Some of the words like “vaccinated”, “shot”, “India” are common across the sentiments.

A deep learning model and a hybrid model considered to predict the sentiments classified by VADER tool. The sentiment classes are imbalanced and training the deep learning model are computationally expensive and time consuming. To address this constraint, bootstrap resampling with sample replacement technique was used to process 5,000 samples from each sentiment. Thus, made a balanced dataset for the deep learning models. The models were executed for five iterations to assess the model performance on different data. The BERT model scored higher training and test accuracy compared to the Bi-LSTM with BERT model across the iterations. Though the Bi-LSTM with BERT model scored lower accuracy, the loss of the validation and test data was always closer to the training data. This specify that the model may perform well if more sample data passed on to the designed Bi-LSTM with BERT model.

The tweets used in the study was of historical rather than recent or real time tweets so the research might not find the latest topics or information. Moreover, the investigation focused only on English language tweets which could bound the applicability of the study’s results to other language. Furthermore, the models were trained and tested only few samples from the original dataset which could have consequences if applied to the complete dataset.

In summary, the study identified the major keywords used in the sentiments that could help the governments, policymakers and other relevant people. Also, it provides an improved method to classify the sentiment expressed in social media posts.

6.3 Contribution to knowledge

The study demonstrates a better approach to categorize the sentiments by using both lexicon and deep learning the models BERT and Bi-LSTM with BERT. Sentiments are classified into three classes i.e. positive, negative and neutral instead of relying on binary class (only positive or negative). Health authorities can monitor the public sentiments to adjust the vaccination drives. Policy makers could make decisions, communicate effectively, address

misinformation and encourage vaccine acceptance based on top keywords found on each sentiment. The same methodology can also be applied for other sentiment analysis on different topics. Also, the researchers can further fine tune the approach to improve the models for better performance.

6.4 Future Recommendations

Sarcasm and misinformation from the tweets can be removed to get a better sentiment classification. The sentiments can be observed for other real time tweets to focus on the current topics. Apart from the tweet texts, the research can be expanded to analysis the sentiments through the images or videos. Other than the sentiment analysis, consider analyzing user engagement metrics such as likes, retweets or comments to measure different types of content on social media platform. The current approach can be extended to multiple language instead relying only on English language. The hyperparameters of the hybrid model Bi-LSTM with BERT can be fined tuned for a better metrics. Additionally, apply these models to a larger dataset for robust results. Importantly, consider ethical implications when analyzing social media data.

REFERENCES

Ainapure, B.S., Pise, R.N., Reddy, P., Appasani, B., Srinivasulu, A., Khan, M.S. and Bizon, N., (2023) Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches. *Sustainability*, [online] 153, p.2573. Available at: https://www.mdpi.com/2071-1050/15/3/2573.

Albahli, S. and Nawaz, M., (2023) TSM-CV: Twitter Sentiment Analysis for COVID-19 Vaccines Using Deep Learning. *Electronics*, [online] 1215, p.3372. Available at: https://www.mdpi.com/2079-9292/12/15/3372.

Amujo, O., Ibeke, E., Fuzi, R., Ogara, U. and Iwendi, C., (2023) Sentiment Computation of UK-Originated COVID-19 Vaccine Tweets: A Chronological Analysis and News Effect. *Sustainability*, [online] 154, p.3212. Available at: https://www.mdpi.com/2071-1050/15/4/3212.

Anon (2024) *Statista - Popular social networks worldwide as of October 2023*. [online] Statista. Available at: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ [Accessed 14 Mar. 2024].

Anon (2024) *WHO - COVID-19 cases reported*. [online] WHO website. Available at: https://data.who.int/dashboards/covid19/cases?n=c [Accessed 14 Mar. 2024].

Anon (2024) *WHO - COVID-19 vaccination, World data*. [online] WHO website. Available at: https://data.who.int/dashboards/covid19/vaccines?n=c [Accessed 14 Mar. 2024].

Anon (2024) *WHO media briefing on COVID-19*. [online] WHO website. Available at: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 [Accessed 14 Mar. 2024].

Aqlan, A.A.Q., Manjula, B. and Lakshman Naik, R., (2019) A Study of Sentiment Analysis: Concepts, Techniques, and Challenges. In: *Lecture Notes on Data Engineering and Communications Technologies*. [online] Springer Science and Business Media Deutschland GmbH, pp.147–162. Available at: http://link.springer.com/10.1007/978-981-13-6459-4_16.

Arora, A., Bansal, S., Kandpal, C., Aswani, R. and Dwivedi, Y., (2019) Measuring social media influencer index- insights from facebook, Twitter and Instagram. *Journal of Retailing and Consumer Services*, [online] 49, pp.86–101. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0969698919300128.

Aslan, S., Kızıloluk, S. and Sert, E., (2023) TSA-CNN-AOA: Twitter sentiment analysis using CNN optimized via arithmetic optimization algorithm. *Neural Computing and Applications*, [online] 3514, pp.10311–10328. Available at: https://link.springer.com/10.1007/s00521-023-08236-2.

Birjali, M., Kasri, M. and Beni-Hssane, A., (2021) A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, [online] 226, p.107134. Available at: https://linkinghub.elsevier.com/retrieve/pii/S095070512100397X.

Cai, R., Qin, B., Chen, Y., Zhang, L., Yang, R., Chen, S. and Wang, W., (2020) Sentiment Analysis About Investors and Consumers in Energy Market Based on BERT-BiLSTM. *IEEE Access*, [online] 8, pp.171408–171415. Available at: https://ieeexplore.ieee.org/document/9200457/.

Chintalapudi, N., Battineni, G. and Amenta, F., (2021) Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models. *Infectious Disease Reports*, [online] 132, pp.329–339. Available at: https://www.mdpi.com/2036-7449/13/2/32.

Chiou, L. and Tucker, C.E., (2018) Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement. *SSRN Electronic Journal*. [online] Available at: https://www.ssrn.com/abstract=3209929.

Chou, W.-Y.S. and Budenz, A., (2020) Considering Emotion in COVID-19 Vaccine Communication: Addressing Vaccine Hesitancy and Fostering Vaccine Confidence. *Health Communication*, [online] 3514, pp.1718–1722. Available at: https://www.tandfonline.com/doi/full/10.1080/10410236.2020.1838096.

Dang, C.N., Moreno-García, M.N. and De la Prieta, F., (2021) Hybrid Deep Learning Models for Sentiment Analysis. *Complexity*, [online] 2021, pp.1–16. Available at: https://www.hindawi.com/journals/complexity/2021/9986920/.

Dhanalakshmi, V., Bino, D. and Saravanan, A.M., (2016) Opinion mining from student feedback data using supervised learning algorithms. In: *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*. [online] IEEE, pp.1–5. Available at: http://ieeexplore.ieee.org/document/7460390/.

Efron, B. and Tibshirani, R.J., (1994) *An Introduction to the Bootstrap*. [online] Chapman and Hall/CRC. Available at: https://www.taylorfrancis.com/books/9781000064988.

Ferdous, Z., Akhter, R., Tahsin, A., Nuha Mustafina, S. and Tabassum, N., (2022) Sentiment Analysis on COVID-19 Vaccine Twitter Data using Neural Network Models. In: *Proceedings of the 2nd International Conference on Computing Advancements*. [online] New York, NY, USA: ACM, pp.435–441. Available at: https://dl.acm.org/doi/10.1145/3542954.3543064.

Funk, C.D., Laferrière, C. and Ardakani, A., (2020) A Snapshot of the Global Race for Vaccines Targeting SARS-CoV-2 and the COVID-19 Pandemic. *Frontiers in Pharmacology*, [online] 11. Available at: https://www.frontiersin.org/article/10.3389/fphar.2020.00937/full.

Gong, W., Parkkila, S., Wu, X. and Aspatwar, A., (2023) SARS-CoV-2 variants and COVID-19 vaccines: Current challenges and future strategies. *International Reviews of Immunology*, [online] 426, pp.393–414. Available at: https://www.tandfonline.com/doi/full/10.1080/08830185.2022.2079642.

Gou, Z. and Li, Y., (2023) Integrating BERT Embeddings and BiLSTM for Emotion Analysis of Dialogue. *Computational Intelligence and Neuroscience*, [online] 2023, pp.1–8. Available at: https://www.hindawi.com/journals/cin/2023/6618452/.

Green, M.S., Abdullah, R., Vered, S. and Nitzan, D., (2021) A study of ethnic, gender and educational differences in attitudes toward COVID-19 vaccines in Israel – implications for

vaccination implementation policies. *Israel Journal of Health Policy Research*, [online] 101, p.26. Available at: https://ijhpr.biomedcentral.com/articles/10.1186/s13584-021-00458-w.

Hayawi, K., Shahriar, S., Serhani, M.A., Taleb, I. and Mathew, S.S., (2022) ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health*, [online] 203, pp.23–30. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0033350621004534.

Huangfu, L., Mo, Y., Zhang, P., Zeng, D.D. and He, S., (2022) COVID-19 Vaccine Tweets After Vaccine Rollout: Sentiment–Based Topic Modeling. *Journal of Medical Internet Research*, [online] 242, p.e31726. Available at: https://www.jmir.org/2022/2/e31726.

Hugging Face, (2024) *BERT*. [online] Hugging Face. Available at: https://huggingface.co/transformers/v2.11.0/model_doc/bert.html# [Accessed 17 Mar. 2024].

Hutto, C. and Gilbert, E., (2014) VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, [online] 81, pp.216–225. Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

Jabalameli, S., Xu, Y. and Shetty, S., (2022) Spatial and sentiment analysis of public opinion toward COVID-19 pandemic using twitter data: At the early stage of vaccination. *International Journal of Disaster Risk Reduction*, [online] 80, p.103204. Available at: https://linkinghub.elsevier.com/retrieve/pii/S221242092200423X.

Joloudari, J.H., Hussain, S., Nematollahi, M.A., Bagheri, R., Fazl, F., Alizadehsani, R., Lashgari, R. and Talukder, A., (2023) BERT-deep CNN: state of the art for sentiment analysis of COVID-19 tweets. *Social Network Analysis and Mining*, [online] 131, p.99. Available at: https://link.springer.com/10.1007/s13278-023-01102-y.

Kathiravan, P., Saranya, R. and Sekar, S., (2023) Sentiment Analysis of COVID-19 Tweets Using TextBlob and Machine Learning Classifiers. [online] pp.89–106. Available at: https://link.springer.com/10.1007/978-981-19-6634-7_8.

Kaur, H., Ahsaan, S.U., Alankar, B. and Chang, V., (2021) A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets. *Information Systems Frontiers*, [online] 236, pp.1417–1429. Available at: https://link.springer.com/10.1007/s10796-021-10135-7.

Kaur, S., Kaul, P. and Zadeh, P.M., (2020) Monitoring the Dynamics of Emotions during COVID-19 Using Twitter Data. *Procedia Computer Science*, [online] 177, pp.423–430. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1877050920323243.

Kumar, A., Singh, R., Kaur, J., Pandey, S., Sharma, V., Thakur, L., Sati, S., Mani, S., Asthana, S., Sharma, T.K., Chaudhuri, S., Bhattacharyya, S. and Kumar, N., (2021) Wuhan to World: The COVID-19 Pandemic. *Frontiers in Cellular and Infection Microbiology*, [online] 11. Available at: https://www.frontiersin.org/articles/10.3389/fcimb.2021.596201/full.

Kwon, S., Joshi, A.D., Lo, C.-H., Drew, D.A., Nguyen, L.H., Guo, C.-G., Ma, W., Mehta, R.S., Shebl, F.M., Warner, E.T., Astley, C.M., Merino, J., Murray, B., Wolf, J., Ourselin, S., Steves, C.J., Spector, T.D., Hart, J.E., Song, M., VoPham, T. and Chan, A.T., (2021) Association of

social distancing and face mask use with risk of COVID-19. *Nature Communications*, [online] 121, p.3737. Available at: https://www.nature.com/articles/s41467-021-24115-7.

Levin, E.G., Lustig, Y., Cohen, C., Fluss, R., Indenbaum, V., Amit, S., Doolman, R., Asraf, K., Mendelson, E., Ziv, A., Rubin, C., Freedman, L., Kreiss, Y. and Regev-Yochay, G., (2021) Waning Immune Humoral Response to BNT162b2 Covid-19 Vaccine over 6 Months. *New England Journal of Medicine*, [online] 38524, p.e84. Available at: http://www.nejm.org/doi/10.1056/NEJMoa2114583.

Ligthart, A., Catal, C. and Tekinerdogan, B., (2021) Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, [online] 547, pp.4997–5053. Available at: https://link.springer.com/10.1007/s10462-021-09973-3.

Luo, Y. and Xu, X., (2021) Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic. *International Journal of Hospitality Management*, [online] 94, p.102849. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0278431920304011.

Ma, Y., Leng, C. and Wang, H., (2024) Optimal Subsampling Bootstrap for Massive Data. *Journal of Business & Economic Statistics*, [online] 421, pp.174–186. Available at: https://www.tandfonline.com/doi/full/10.1080/07350015.2023.2166514.

Malecki, K.M.C., Keating, J.A. and Safdar, N., (2021) Crisis Communication and Public Perception of COVID-19 Risk in the Era of Social Media. *Clinical Infectious Diseases*, [online] 724, pp.697–702. Available at: https://academic.oup.com/cid/article/72/4/697/5858208.

Mose, A., Haile, K. and Timerga, A., (2022) COVID-19 vaccine hesitancy among medical and health science students attending Wolkite University in Ethiopia. *PLOS ONE*, [online] 171, p.e0263081. Available at: https://dx.plos.org/10.1371/journal.pone.0263081.

Nasukawa, T. and Yi, J., (2003) Sentiment analysis. In: *Proceedings of the 2nd international conference on Knowledge capture*. [online] New York, NY, USA: ACM, pp.70–77. Available at: https://dl.acm.org/doi/10.1145/945645.945658.

Paliwal, S., Parveen, S., Afshar Alam, M. and Ahmed, J., (2022) Sentiment Analysis of COVID-19 Vaccine Rollout in India. [online] pp.21–33. Available at: https://link.springer.com/10.1007/978-981-16-5987-4_3.

Pertwee, E., Simas, C. and Larson, H.J., (2022) An epidemic of uncertainty: rumors, conspiracy theories and vaccine hesitancy. *Nature Medicine*, [online] 283, pp.456–459. Available at: https://www.nature.com/articles/s41591-022-01728-z.

Polack, F.P., Thomas, S.J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J.L., Pérez Marc, G., Moreira, E.D., Zerbini, C., Bailey, R., Swanson, K.A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R.W., Hammitt, L.L., Türeci, Ö., Nell, H., Schaefer, A., Ünal, S., Tresnan, D.B., Mather, S., Dormitzer, P.R., Şahin, U., Jansen, K.U. and Gruber, W.C., (2020) Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine*, [online] 38327, pp.2603–2615. Available at: http://www.nejm.org/doi/10.1056/NEJMoa2034577.

Power, R., Robinson, B., Colton, J. and Cameron, M., (2014) Emergency Situation Awareness: Twitter Case Studies. [online] pp.218–231. Available at: http://link.springer.com/10.1007/978-3-319-11818-5_19.

Preda, G., (2021) *COVID-19 All Vaccines Tweets*. [online] Kaggle. Available at: https://www.kaggle.com/dsv/2845240 [Accessed 14 Mar. 2024].

Qorib, M., Oladunni, T., Denis, M., Ososanya, E. and Cotae, P., (2023a) COVID-19 Vaccine Hesitancy: A Global Public Health and Risk Modelling Framework Using an Environmental Deep Neural Network, Sentiment Classification with Text Mining and Emotional Reactions from COVID-19 Vaccination Tweets. *International Journal of Environmental Research and Public Health*, [online] 2010, p.5803. Available at: https://www.mdpi.com/1660-4601/20/10/5803.

Qorib, M., Oladunni, T., Denis, M., Ososanya, E. and Cotae, P., (2023b) Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications*, [online] 212, p.118715. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0957417422017407.

Raheja, S. and Asthana, A., (2021) Sentimental Analysis of Twitter Comments on Covid-19. In: *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. [online] IEEE, pp.704–708. Available at: https://ieeexplore.ieee.org/document/9377048/.

Rahmanti, A.R., Ningrum, D.N.A., Lazuardi, L., Yang, H.-C. and Li, Y.-C., (2021) Social Media Data Analytics for Outbreak Risk Communication: Public Attention on the "New Normal" During the COVID-19 Pandemic in Indonesia. *Computer Methods and Programs in Biomedicine*, [online] 205, p.106083. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0169260721001589.

Rani, S. and Jain, A., (2023) DFM: Deep Fusion Model for COVID-19 Vaccine Sentiment Analysis. [online] pp.227–235. Available at: https://link.springer.com/10.1007/978-981-19-9228-5_20.

De Rosis, S., Lopreite, M., Puliga, M. and Vainieri, M., (2021) The early weeks of the Italian Covid-19 outbreak: sentiment insights from a Twitter analysis. *Health Policy*, [online] 1258, pp.987–994. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0168851021001627.

Sahoo, C., Wankhade, M. and Singh, B.K., (2023) Sentiment analysis using deep learning techniques: a comprehensive review. *International Journal of Multimedia Information Retrieval*, [online] 122, p.41. Available at: https://link.springer.com/10.1007/s13735-023-00308-2.

Said, H., Tawfik, B.S. and Makhlouf, M.A., (2023) A Deep Learning Approach for Sentiment Classification of COVID-19 Vaccination Tweets. *International Journal of Advanced Computer Science and Applications*, [online] 144. Available at: http://thesai.org/Publications/ViewPaper?Volume=14&Issue=4&Code=IJACSA&SerialNo=58.

Saleh, S.N., McDonald, S.A., Basit, M.A., Kumar, S., Arasaratnam, R.J., Perl, T.M., Lehmann, C.U. and Medford, R.J., (2023) Public perception of COVID-19 vaccines through analysis of

Twitter content and users. *Vaccine*, [online] 4133, pp.4844–4853. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0264410X23007430.

Sallam, M., (2021) COVID-19 Vaccine Hesitancy Worldwide: A Concise Systematic Review of Vaccine Acceptance Rates. *Vaccines*, [online] 92, p.160. Available at: https://www.mdpi.com/2076-393X/9/2/160.

Samuel, J., Rahman, Md.M., Ali, G.G.Md.N., Samuel, Y., Pelaez, A., Chong, P.H.J. and Yakubov, M., (2020) Feeling Positive About Reopening? New Normal Scenarios From COVID-19 US Reopen Sentiment Analytics. *IEEE Access*, [online] 8, pp.142173–142190. Available at: https://ieeexplore.ieee.org/document/9154672/.

Singh, M., Jakhar, A.K. and Pandey, S., (2021) Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Social Network Analysis and Mining*, [online] 111, p.33. Available at: https://link.springer.com/10.1007/s13278-021-00737-z.

Thangavel, P. and Lourdusamy, R., (2023) A lexicon-based approach for sentiment analysis of multimodal content in tweets. *Multimedia Tools and Applications*, [online] 8216, pp.24203–24226. Available at: https://link.springer.com/10.1007/s11042-023-14411-3.

Umair, A. and Masciari, E., (2023) Sentimental and spatial analysis of COVID-19 vaccines tweets. *Journal of Intelligent Information Systems*, [online] 601, pp.1–21. Available at: https://link.springer.com/10.1007/s10844-022-00699-4.

Umair, A., Masciari, E. and Habib Ullah, M.H., (2021) Sentimental Analysis Applications and Approaches during COVID-19: A Survey. In: *25th International Database Engineering & Applications Symposium*. [online] New York, NY, USA: ACM, pp.304–308. Available at: https://dl.acm.org/doi/10.1145/3472163.3472274.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., (2017) Attention Is All You Need. [online] Available at: http://arxiv.org/abs/1706.03762.

Voidarou, C., Rozos, G., Stavropoulou, E., Giorgi, E., Stefanis, C., Vakadaris, G., Vaou, N., Tsigalou, C., Kourkoutas, Y. and Bezirtzoglou, E., (2023) COVID-19 on the spectrum: a scoping review of hygienic standards. *Frontiers in Public Health*, [online] 11. Available at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1202216/full.

Wang, Y. and Chen, Y., (2023) Characterizing discourses about COVID-19 vaccines on Twitter: a topic modeling and sentiment analysis approach. *Journal of Communication in Healthcare*, [online] 161, pp.103–112. Available at: https://www.tandfonline.com/doi/full/10.1080/17538068.2022.2054196.

Yadav, A. and Vishwakarma, D.K., (2020) Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, [online] 536, pp.4335–4385. Available at: http://link.springer.com/10.1007/s10462-019-09794-5.

Zhou, X., (2023) Sentiment Analysis of the Consumer Review Text Based on BERT-BiLSTM in a Social Media Environment. *International Journal of Information Technologies and Systems Approach*, [online] 162, pp.1–16. Available at: https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJITSA.325618.

Zulfiker, Md.S., Kabir, N., Biswas, A.A., Zulfiker, S. and Uddin, M.S., (2022) Analyzing the public sentiment on COVID-19 vaccination in social media: Bangladesh context. *Array*, [online] 15, p.100204. Available at: https://linkinghub.elsevier.com/retrieve/pii/S2590005622000534.

APPENDIX A: RESEARCH PROPOSAL

ANALYSING COVID-19 VACCINE-RELATED TWEETS: INSIGHTS INTO PUBLIC SENTIMENTS ON SOCIAL MEDIA PLATFORMS

SARAVANAN KANMANI

Research Proposal

NOVEMBER 2023

Abstract

The COVID-19 epidemic has been a worldwide health disaster that has affected millions of individuals in the global, leading to many deaths. While vaccines have been developed to prevent the spread of the virus, there has been some hesitancy among people to accept them. In this research, the aim is to analyze COVID-19 vaccine-related tweets to gain insights into the common sentiments shared by people on social media platforms like X (formerly Twitter). Openly available dataset that contains tweets related to COVID-19 disease will be used and implement two methods, Lexicon-based and BERT with BiLSTM, to analyze the tweets. In the lexicon method, FastText will be used to compute features and VADER to assign sentiment polarity (i.e., positive, negative, or neutral) of the tweets. Then the classification of sentiments will be performed using a deep learning BERT with BiLSTM model and evaluate the performance using accuracy, recall, precision, and F1-score. The findings of this research could be helpful for public organizations and health authorities to be better prepared for future outbreaks similar to COVID-19.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI……..... Artificial Intelligence
API…..... Application Programming Interface
BBC……..... British Broadcasting Corporation
BERT……..... Bidirectional Encoder Representations from Transformers
BiLSTM……. Bidirectional Long Short-Term Memory
BiGRU……. Bidirectional Gated Recurrent Unit
CNN……..... Convolutional Neural Network
COVID-19…… Corona Virus Disease of 2019
DL……..... Deep Learning
DT……..... Decision Tree
RF……..... Random Forest
KNN……..... K-Nearest Neighbor
GPU……. Graphics Processing Unit
LR……. Logistic Regression
LSTM……... Long Short-Term Memory
ML…… Machine Learning
NB…… Naïve Bayes
NLP……. Natural Language Processing
NLTK……... Natural Language Toolkit
RAM……. Random Access Memory
RMDL……... Random Multimodal Deep Learning
SARS-CoV2… Severe Acute Respiratory Syndrome–related Coronavirus
SVM……. Support Vector Machine
TF-IDF……. Term Frequency–Inverse Document Frequency
TSA……. Twitter Sentiment Analysis
VADER…… Valence Aware Dictionary and sEntiment Reasoner
WHO…… World Health Organization

1. Background

An infectious disease caused by the SARS-CoV2 virus is COVID-19 (Coronavirus 2019). The exact origin of the virus is still unknown but it was first identified in Wuhan, China, in December 2019 and spread to the entire globe (Kumar et al., 2021). The virus can mainly spread through an infected persons when they cough, sneeze, talk, sing or breathe. At the time of writing this proposal, over 700 million of people affected by COVID-19 and over 6 million deaths caused as mentioned in the WHO Dashboard (WHO Coronavirus (COVID-19) Dashboard, 2023). Though different countries followed various measures, the best way to get rid of COVID-19 disease is vaccination (Umair and Masciari, 2023), (Umair et al., 2021). Many pharmaceutical, research institutes, and government put their efforts to find COVID-19 vaccines. After the introduction of vaccine, people hesitated to accept the vaccine because of the fear of its rumor and side-effects (Umair and Masciari, 2023), (Green et al., 2021). Therefore, government and public health authorities in need of understanding the people thoughts about COVID-19 vaccines to plan it better now and in similar pandemic situation in future.

People express their thoughts in social media platform like X (formerly Twitter) as it is publicly accessible by everyone. The tweets in social media can be analyzed to understand the people's reaction on any recent topic (Luo and Xu, 2021). X (formerly Twitter) is a social networking platform that allows users to post their opinions or comments as tweets. Sentiment analysis is a popular method to analyze the tweets and to find the polarization (i.e., positive, negative, neutral) of the tweets (Raheja and Asthana, 2021).

This study is to analyze public feelings and sentiment about COVID-19 vaccines which they have expressed in social media platform X (formerly Twitter). The purpose is to help the government organization and health authorities to consider the insights from the public opinion and design their vaccine policy in such pandemic situation. A dataset from Kaggle website will be used for this research and text will be pre-processed for a proper analysis. Two sentiment analysis approaches Lexicon-based and deep learning are proposed in this research. FastText will be used for feature computation and VADER will be used to get sentiment polarity. Finally, sentiment classification will be performed using a deep learning technique BERT with BiLSTM and the performance of the model will be evaluated. A visualization to get the important topics discussed in COVID-19 will be plotted using WordCloud.

2. Related Work

COVID-19 has been severely affected the life of people around the world. After the roll out of vaccines for the virus COVID-19, public shared their opinions regarding the vaccines in social media like X (formerly Twitter) platform. Many researchers analyzed the sentiments of tweets regarding COVID-19 vaccines to understand the public thoughts. (Albahli and Nawaz, 2023) presented a novel DL approach "TSM-CV" for sentiment analysis of tweets regarding coronavirus vaccines to know the human behavior. The authors used both past and real-time

data from X (formerly Twitter) platform for their analysis. They proposed a RMDL classifier for their sentiment prediction. Based on this research, the authors concluded that analyzing COVID-19 vaccine sentiments can support the health authorities to take actions to get rid of leagues that are against vaccination. People's opinion about COVID-19 vaccination process also important research (Said et al., 2023) that indicates the people have neutral sentiments about vaccines. In this study, the authors collected the tweets from X (formerly Twitter) platform for a month and applied an ensemble deep learning model LSTM-2BiGRU that outperformed other classical models like LR, NB, DT, RF and KNN. Another study from (Umair and Masciari, 2023) to understand the public feelings and sentiments showed that sentiment and spatial analysis helps in identifying the people's attitude towards vaccines. The authors also visualized COVID-19 vaccines tweet data geo-graphically and implemented several geo-spatial methods like hotspot analysis, kernel density estimation. The authors could have compared the BERT model used in this study with other sentiment analysis techniques.

Public feelings and sentiments of COVID-19 vaccination changed over time in response to different BBC news reports was researched in the study of (Amujo et al., 2023). This research is based on the three UK based vaccines that are Pfizer-BioNTech, Moderna, and Oxford AstraZeneca. The authors used six period of data based on the BBC news reports of development of COVID-19 vaccines and implemented BERT model to analyze the public views. In this research the authors found that Moderna vaccine had more positive sentiments compared to other two vaccines and AstraZeneca had more negative sentiment. The paper also found that BBC news report had a significant impact on the public sentiment. The study does not seem to address the impact of misinformation on vaccine sentiment. Vaccine hesitancy is the key concern and various researchers used ML techniques to examine the vaccine uncertainty based on tweets from X (formerly Twitter) platform. To understand the people views before and after vaccination, the authors (Qorib et al., 2023a) used NRCLexicon technique to label tweets into 10 different classes and the significance of the associations among the basic emotions were checked using t-test. This study showed that public feelings turned gradually to positive about COVID-19 vaccination. Same set of authors (Qorib et al., 2023b) conducted another study to investigate COVID-19 vaccine uncertainty through examining three sentiment calculation methods (Azure Machine Learning, VADER, and TextBlob). They found TextBlob with TF-IDF and LinearSVC classification model performed well and the combination of CountVectorizer and TF-IDF decreases the model accuracy. This study as well concluded that hesitancy in Covid-19 vaccine progressively decreased over a period. Despite the vaccine drives, there was an increasing in vaccine hesitancy due to misinformation regarding the vaccines were spread on social media either via humans or bots. Therefore, the authors (Hayawi et al., 2022) researched about detecting misinformation in tweets related to vaccination. The authors collected the tweets from X (formerly Twitter) and misinformation was manually read and labelled with the help of consistent sources and health specialists. The proposed BERT model performed well with on the test set and the model is effective in finding misrepresentation regarding COVID -19 vaccines on public network.

During COVID-19 period, people shared lot of posts and reviews in X (formerly Twitter) about the disease and its vaccines. These posts and tweets were examined in the study of (Kathiravan et al., 2023) to understand the psychological and emotional impacts of people. In this paper, the authors researched the state-of-art automatic extraction of emotions from public tweets related to coronavirus and it provides public mental health insights about COVID-19 for the health authorities. Deep Learning techniques that competently work on unstructured data was not considered in this study. TSA being a popular topic, the authors (Aslan et al., 2023a) proposed a novel approach using CNN optimized via arithmetic optimization algorithm (TSA-CNN-AOA) to understand people's views on COVID-19 epidemic. This study concluded that the approach they proposed is successful for TSA that can help to minimize and eliminate the impact of disease. Tweets by Indian citizens related to COVID-19 pandemic and vaccine drive was analyzed to help the policy makers and health workers to obtain the insights and make right decisions for similar pandemic outbreaks (Ainapure et al., 2023a).

Based on the research analysis conducted, most of the researchers analyzed the sentiment of COVID-19 tweets using either lexicon or deep learning and with limited number of tweets. In this proposed research, the sentiment analysis will be achieved using both lexicon and deep learning methods that helps to classify the COVID-19 vaccine sentiments effectively.

3. Aim and Objectives

The primary aim of this research is to study public sentiments related to COVID-19 vaccines using tweets from X (formerly Twitter) platform. The purpose of this research is to benefit the public organization, health authorities and policymakers to get insights and to consider the public opinion while introducing their vaccination policy for similar upcoming pandemic outbreak.

The research objectives are framed based on the aim of this study which are as follows:

- To analyze the COVID-19 vaccine related tweets posted by public and find the most common sentiments expressed.
- To find the important topics discussed in COVID-19 vaccine related tweets.
- To classify the sentiments using suitable lexicon and deep learning model.
- To evaluate the performance of the classification model developed.

4. Significance of the Study

Sentiment analysis of the tweets about coronavirus vaccine can provide the public thoughts about COVID-19 disease and its vaccines. By understanding the public opinions and attitudes, the government organization and health authorities can be prepared for similar pandemic situation on how to approach the people for vaccination.

5. Scope of the Study

This research explores public sentiments towards COVID-19 vaccine related tweets using lexicon based and deep learning techniques. The study can provide insights to policymakers about COVID-19 vaccines. The study will be conducted in open sources software and models.

Sentiment analysis will be performed only on historical COVID-19 vaccine associated tweets data which is publicly available and no real time data will be used in this study. Also, this study will be conducted only on tweets text data that were posted in English language.

To make this research achievable within given timeframe the scopes are defined. It also helps to identify the targeted audience and methods used for sentiment analysis.

6. Research Methodology

This study proposes a lexicon and deep learning-based approach to analyze the COVID-19 vaccine associated tweets data. The dataset is publicly available in Kaggle website. A sequence of data preprocessing steps will be executed to improve and clean the data. FastText Word Embeddings will be used for feature computation. A rule and lexicon-based sentiment technique called Valence Aware Dictionary and Sentiment Reasoner (VADER) will be used for sentiment analysis. Finally, classification of sentiments will be performed using a transformed based ML-model Bidirectional Encoder Representation from Transformers (BERT) with Bidirectional Long Short-Term Memory (BiLSTM). The performance of the model will be assessed with Accuracy, Precision, Recall and F1-score. Overview of the planned research methodology is shown in Figure 1.

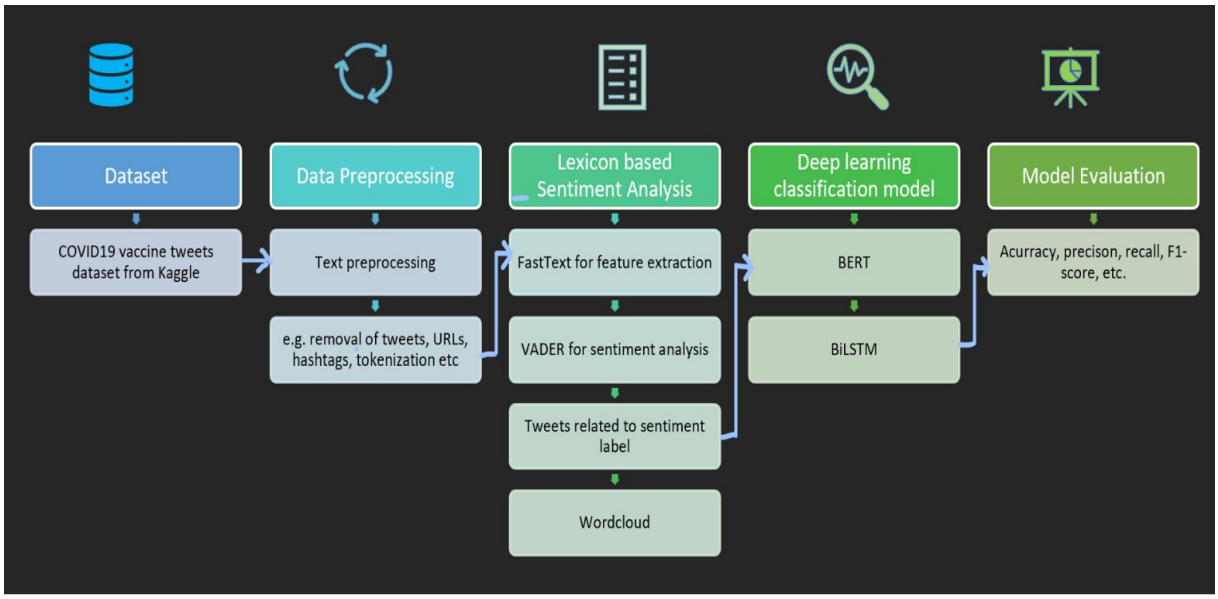

Figure 1. Proposed research methodology overview

6.1. Dataset Description

Freely accessible dataset called "COVID-19 All Vaccines Tweets" from Kaggle (COVID-19 All Vaccines Tweets, 2023) will be used in this research. The dataset contains the tweets associated to COVID-19 vaccines used around the globe like Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin and Sputnik V. There are 228,207 tweets data along with other information of tweeter like name, location, created date, total followers, total friends, number of retweets etc.

6.2. Data Preprocessing

Preprocessing of a text is an important step to train an efficient model. To improve the performance of the sentiment analysis the text preprocessing like removing URLs, punctuations, retweets, hashtags and usernames, lowercase conversion, emojis conversion as words, tokenization, stemming and lemmatization will be performed as followed in (Albahli and Nawaz, 2023), (Qorib et al., 2023b), (Amujo et al., 2023).

6.3. FastText embedding for feature extraction

FastText was developed by Facebook's AI Research team to efficiently handle large amount of text data. It can transform words or text into continuous vectors that can be used in any language spoken in a given task (Aslan et al., 2023b). FastText can outperform the Word2Vec and GloVe word embedding methods because it creates better word embeddings for occasional words or even words not present during training, as the n-gram character vectors are shared with other words (Albahli and Nawaz, 2023). In this research FastText approach will be used to capture both syntactic and semantic information of the text for a better sentiment analysis.

6.4. Sentiment Analysis using VADER

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a tool that analyzes the sentiment of a piece of text by using lexicon of words and their corresponding sentiment scores. It is rule-based tool that determines the sentiment of a piece of text by using a dictionary of words and rules. VADER sentiment analysis relies on the lexicon method to plot vocabulary features to sentiment scores (Ainapure et al., 2023b). Features extracted as vectors from FastText will be passed to VADER so the sentiment analysis will be performed and three labels positive (1), neutral (0) or negative (-1) will be assigned for each sample (Albahli and Nawaz, 2023).

6.5. Sentiment Classification using BERT with BiLSTM

To classify the sentiments, BERT with BiLSTM model will be used in this study. BERT is a neural network-based method which is a pre-trained model and can be fine-tuned to perform various NLP tasks including sentiment analysis. The pre-trained BERT model can be finetuned with a single extra output layer to generate high performing models for various NLP process tasks (Joloudari et al., 2023). It is sequence-to-sequence model based on attention mechanisms for encoding and decoding word-based information (Umair et al., 2023). LSTM is an enhanced model of RNN and BiLSTM composed from Bidirectional LSTM. Combining the models BERT and BiLSTM can improve the model performance as BERT has the capability to learn statistics features of the neighboring words and BiLSTM is capable in learning the context info (Cai et al., 2020). The preprocessed text data is fed into the fine-tuned model, that outputs a possibility distribution over the probable sentiment labels. The sentiment with the maximum probability is then given to the input text. BERT-BiLSTM has become a popular tool for sentiment analysis and has accomplished state-of-the-art results on many sentiments analysis.

6.6. Model Evaluation

The developed model performance will be assessed using several metrics like accuracy, precision, recall, F1-score. The metrics are considered as typical performance analysis methods and most of the researchers use these measures to analyze the public sentiments.

6.7. Data Visualization

A Word cloud is a graphic visualization that used to summarize the textual data by showing frequent keywords occurring in the text in different size and color. Size of individual word signifies its occurrence or importance. A word cloud will be used to visualize the important topics discussed in the COVID-19 vaccine tweets.

7. Required Resources

This section shows the required hardware and software for this research.

7.1. Hardware requirements

A Desktop/laptop computer with GPU with at least 12GB of RAM that can run the deep learning models. The computer should also connect to internet and able to compile and execute the codes.

7.2. Software requirements

The below are the required software for this study.

- Python (3.6 or higher)
- Jupyter notebook (6.1 or higher) for python coding
- Pandas (1.4.2 or higher) for data processing
- Numpy (1.21.5 or higher) for array computation, mathematical functions, etc.
- Matplotlib (3.5.1 or higher), Seaborn (0.11.2 or higher) for visualization
- FastText (0.9.2 or higher). SciPy, NumPy and pybind11 required for FastText.
- VADER (3.3.2 or higher). vaderSentiment package, NLTK, NumPy and SciPy required for VADER.
- BERT (1 or higher). It requires transformers package, PyTorch and TensorFlow.
- WordCloud (1.9.2 or higher). wordcloud package, NumPy, Pillow, matplotlib required to use WordCloud.

8. Research Plan

This section shows the activity and the plan for the research proposed. Also, it explains the contingency plan in case of a risk.

8.1. Gantt Chart

Proposed plan for the research is shown in Figure 2.

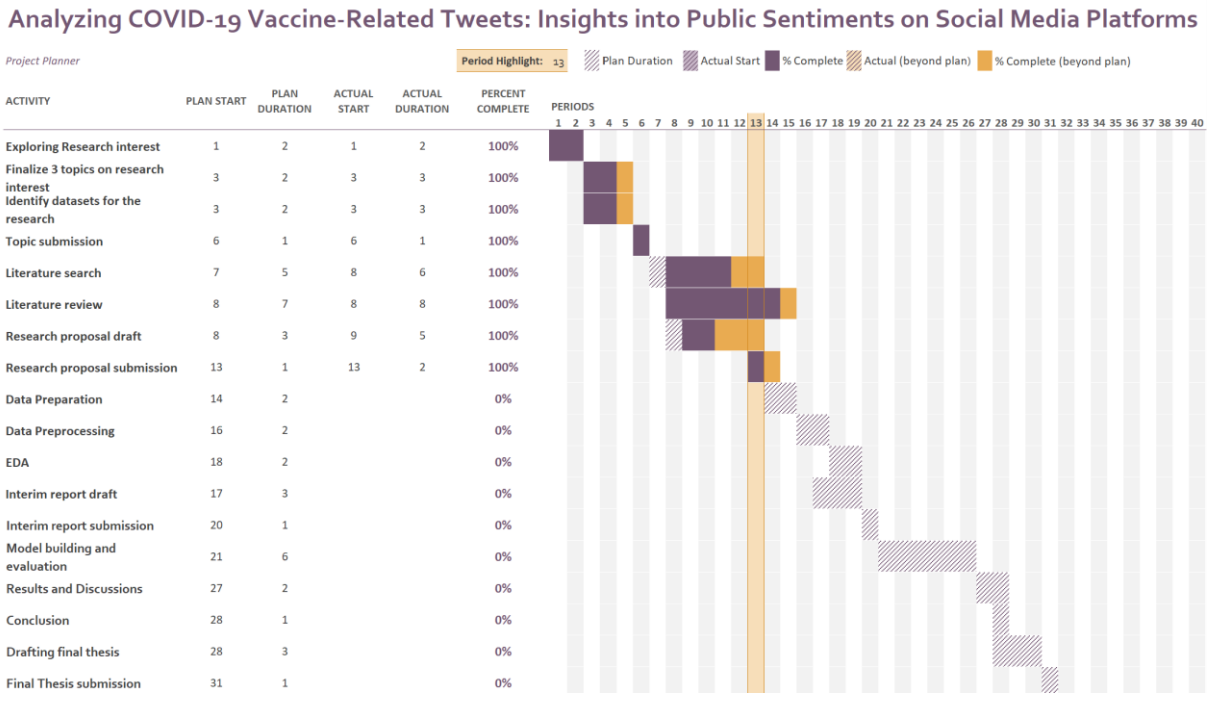

Figure 2. Gantt Chart

8.2. Risk Mitigation and Contingency plan

The contingency plan shown in Table 1 in case of any risk during the period of this research.

Table 1. Risk and Contingency plan

Risk	**Contingency**
Could not perform the sentiment analysis with the proposed computation, lexicon-based or deep learning.	An appropriate suitable sentiment analysis method will be used.
Research cannot be continued or a delay due to health or personal problems.	An extension or an approval for delay from university/upGrad will be requested.

References

Ainapure, B.S., Pise, R.N., Reddy, P., Appasani, B., Srinivasulu, A., Khan, M.S. and Bizon, N., (2023a) Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches. Sustainability (Switzerland), 153.

Ainapure, B.S., Pise, R.N., Reddy, P., Appasani, B., Srinivasulu, A., Khan, M.S. and Bizon, N., (2023b) Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches. Sustainability (Switzerland), 153.

Albahli, S. and Nawaz, M., (2023) TSM-CV: Twitter Sentiment Analysis for COVID-19 Vaccines Using Deep Learning. Electronics (Switzerland), 1215.

Amujo, O., Ibeke, E., Fuzi, R., Ogara, U. and Iwendi, C., (2023) Sentiment Computation of UK-Originated COVID-19 Vaccine Tweets: A Chronological Analysis and News Effect. Sustainability (Switzerland), 154.

Anon (2023) COVID-19 All Vaccines Tweets. [online] Available at: https://www.kaggle.com/datasets/gpreda/all-covid19-vaccines-tweets [Accessed 20 Nov. 2023].

Anon (2023) WHO Coronavirus (COVID-19) Dashboard. [online] Available at: https://covid19.who.int/ [Accessed 20 Nov. 2023].

Aslan, S., Kızıloluk, S. and Sert, E., (2023a) TSA-CNN-AOA: Twitter sentiment analysis using CNN optimized via arithmetic optimization algorithm. Neural Computing and Applications, 3514, pp.10311–10328.

Aslan, S., Kızıloluk, S. and Sert, E., (2023b) TSA-CNN-AOA: Twitter sentiment analysis using CNN optimized via arithmetic optimization algorithm. Neural Computing and Applications, 3514, pp.10311–10328.

Cai, R., Qin, B., Chen, Y., Zhang, L., Yang, R., Chen, S. and Wang, W., (2020) Sentiment analysis about investors and consumers in energy market based on BERT-BILSTM. IEEE Access, 8, pp.171408–171415.

Green, M.S., Abdullah, R., Vered, S. and Nitzan, D., (2021) A study of ethnic, gender and educational differences in attitudes toward COVID-19 vaccines in Israel – implications for vaccination implementation policies. Israel Journal of Health Policy Research, 101.

Hayawi, K., Shahriar, S., Serhani, M.A., Taleb, I. and Mathew, S.S., (2022) ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. Public Health, 203, pp.23–30.

Joloudari, J.H., Hussain, S., Nematollahi, M.A., Bagheri, R., Fazl, F., Alizadehsani, R., Lashgari, R. and Talukder, A., (2023) BERT-deep CNN: state of the art for sentiment analysis of COVID-19 tweets. Social Network Analysis and Mining, 131.

Kathiravan, P., Saranya, R. and Sekar, S., (2023) Sentiment Analysis of COVID-19 Tweets Using TextBlob and Machine Learning Classifiers: An Evaluation to Show How COVID-19 Opinions Is Influencing Psychological Reactions of People's Behavior in Social Media. Lecture Notes in Networks and Systems, [online] 552, pp.89–106. Available at: https://link.springer.com/chapter/10.1007/978-981-19-6634-7_8 [Accessed 7 Nov. 2023].

Kumar, A., Singh, R., Kaur, J., Pandey, S., Sharma, V., Thakur, L., Sati, S., Mani, S., Asthana, S., Sharma, T.K., Chaudhuri, S., Bhattacharyya, S. and Kumar, N., (2021) Wuhan to World: The COVID-19 Pandemic. Frontiers in Cellular and Infection Microbiology, .

Luo, Y. and Xu, X., (2021) Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic. International Journal of Hospitality Management, 94.

Qorib, M., Oladunni, T., Denis, M., Ososanya, E. and Cotae, P., (2023a) COVID-19 Vaccine Hesitancy: A Global Public Health and Risk Modelling Framework Using an Environmental Deep Neural Network, Sentiment Classification with Text Mining and Emotional Reactions from COVID-19 Vaccination Tweets. International Journal of Environmental Research and Public Health, 2010.

Qorib, M., Oladunni, T., Denis, M., Ososanya, E. and Cotae, P., (2023b) Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Systems with Applications, 212.

Raheja, S. and Asthana, A., (2021) Sentimental Analysis of Twitter Comments on Covid-19. In: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). pp.704–708.

Said, H., Tawfik, B.S. and Makhlouf, M.A., (2023) A Deep Learning Approach for Sentiment Classification of COVID-19 Vaccination Tweets. [online] IJACSA) International Journal of Advanced Computer Science and Applications, Available at: www.ijacsa.thesai.org.

Umair, A. and Masciari, E., (2023) Sentimental and spatial analysis of COVID-19 vaccines tweets. Journal of Intelligent Information Systems, 601, pp.1–21.

Umair, A., Masciari, E. and Ullah, M.H., (2023) Vaccine sentiment analysis using BERT + NBSVM and geo-spatial approaches. Journal of Supercomputing, 7915, pp.17355–17385.
Umair, A., Masciari, E. and Ullah, M.H.H., (2021) Sentimental Analysis Applications and Approaches during COVID-19: A Survey. In: ACM International Conference Proceeding Series. Association for Computing Machinery, pp.304–308.