Sentiment Analysis on Covid-19 Booster shots Vaccinations

MSc Research Project
Data Analytics

Jacob Mamman
Student ID: X21103330

School of Computing
National College of Ireland

Supervisor: Taimur Hafeez

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Jacob Mamman
Student ID:	X21103330
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Taimur Hafeez
Submission Due Date:	14/12/2022
Project Title:	Sentiment Analysis on Covid-19 Booster shots Vaccinations
Word Count:	5560
Page Count:	20

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

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

Signature:	
Date:	31st January 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

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

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

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

Sentiment Analysis on Covid-19 Booster shots Vaccinations

Jacob Mamman
X21103330

Abstract

The rapid growth of social networks and the ease of internet access have accelerated the spread of incorrect information and rumours on social media platforms. By putting people's mental and physical well-being in jeopardy, this misinformation has made the COVID-19 epidemic much more severe. Whether or whether booster shots of the COVID-19 vaccine will be required is also uncertain at this time. Consequently, in order to clear up this misunderstanding and minimize the widespread anxiety, the sentiment analysis of the sentiment analysis on the covid-19 booster shots vaccination was conducted using a dataset from the Twitter online platform. The study analyzes the sentiment of online discussions around the covid-19 booster shot vaccines using TEXTBLOB as a baseline model and the BERT and RoBERTa algorithms to predict those sentiments and evaluate their performance. The 46,398-item data set utilized was obtained through the Twitter API. The BERT algorithm had better results than the ROBERTA with 0.1 margins, with an F1 accuracy score of 100%.

1 Introduction

1.1 Background of Study

Finding fake or misleading news is a difficult topic in the realm of natural language processing because of the prevalence of online communities and the ease with which information may be spread. A large number of online sites undoubtedly include a wealth of useful information, but there is also a lot of misinformation and myths out there as well. The spread of misleading information must therefore be stopped now more than ever, especially in light of the ongoing COVID-19 pandemic. COVID-19 is without a doubt one of the deadliest tragedies in recorded human history. A global epidemic has been declared by the World Health Organization (WHO). The "infodemic" is being fought while the world is in chaos. With the development of modern communication methods, misleading information and rumours have been circulated widely, which has resulted in worry and anxiety. Because COVID-19 has a significant impact on so many people's lives, the general public has paid great attention to material relating to it. Many adversaries had seized upon this opportunity to disseminate false information for evil motives, whether they be financial, political, or otherwise. A tsunami of false information about the condition's causes, effects, prevention measures, and treatment has engulfed the internet Tasnim et al. (2020) . Contrary to popular belief, drinking alcohol does not destroy the coronavirus, nor can you stop the virus from spreading by dousing yourself in chlorine Islam et al.

(2020). (both of which are urban legends). Those who had depended on this incorrect information had suffered, and in some cases died. According to a study Islam et al. (2020), after accepting the false information circulated on social media, over 5800 people were hospitalized and 800 of them passed away Coleman (2020). People's emotional and physical health have both been impacted by the worry and anxiety this misleading information causes Xiong et al. (2020). During the Covid-19 period, a substantial corpus of research on collective emotions was developed, and Twitter has since become a popular technique for gathering data for many of these studies.Saha et al. (2020) research team mined Twitter for data using a variety of hashtags. On this date, sentiment analysis is performed using the LDA machine learning method. The main incentive in conversations about COVID-19 is determined to be fear.

One way to deal with this is to learn how to spot and avoid websites that spread false information. Because gathering proof can take so much time and effort, it can be difficult for people to tell the difference between true and fake news. In light of these considerations, developing models that can recognize all the misleading information is a viable alternative. The authors of this study set up and compare five transformer-based models for COVID-19 fake news detection: BERT, BERT without LSTM, ALBERT, RoBERTa, and a Hybrid of BERT & ALBERT. To find COVID-19 false report cases, numerous studies have been done. The use of a Decision Tree, Logistic Regression, Gradient Boost, and Support Vector Machine (SVM) to identify COVID-19-related hoaxes was proven by Patwa et al. (2021). in 2021. To do this, they gathered and checked the bias and validity of more than 10,000 news articles and social media posts about COVID-19. What was true news and what was fake news was made abundantly clear. The Support Vector Machine obtained the greatest F1 score (93.32%) when evaluated against the validation data out of all the models they looked at. In an example of automatic COVID-related tweet, authenticity recognition Das et al. (2021). To achieve this, the researchers used previously trained models in an ensemble technique. They also employed an advanced heuristic technique based on username and link domains. The collection includes both "real" news items and "fake" news tales. These last ones were chosen from tweets, posts, and stories that exaggerated the effects of COVID-19.

1.2 Research Question

To what extent are individuals likely to take COVID-19 more booster shots of the vaccine?

1.3 Aim and Objectives

The proposed study's objective is to conduct a sentiment analysis of COVID-19 booster shots. The following are the primary aims toward that end:

- To conduct a critical evaluation of the literature on sentiment analysis related to Covid-19 booster shots vaccinations.
- To evaluate the performance of BERT and ROBERTA algorithms in predicting sentiments
- To determine the positive, neutral and negative sentiments

2 Related Work

2.1 Overview

The Covid-19 dilemma has had widespread repercussions on businesses and the lives of people around the world. Social media serves as a crucial information and communication route for people who are socially isolated in these extraordinary situations. Section 2.1: General Sentiment Analysis Overview of the Covid-19 booster injection vaccinations are the subject of Section 2.2. Section 2.3 focuses on a comparison of the RoBERTa model and other sentiment analysis models, while Section 2.4 focuses on an overview of the research findings.

2.2 General Overview of Sentiment Analysis

Computationally analyzing people's feelings and thoughts about a subject is also known as "sentiment analysis" (SA) or "opinion mining." Some scholars argue that OM and SA take slightly different vantage points. Algorithmic sentiment analysis is a branch of NLP that examines how people feel, think, and express themselves through written text. The fields of construction, emotion identification, and transfer learning all benefit from sentiment analysis. When recognizing and analyzing the sentiment expressed in a text, opinions about a subject are identified, looked into, and retrieved as part of sentiment analysis. Effective techniques for identifying feelings and classifying the attitudes they signify include feature selection, product reviews, sentiment categorization, sentiment polarity labelling, sentiment identification, and sentiment categorization Zhang et al. (2018). Although, one difficulty with sentiment analysis is defining the investigation's objects, which are judgments and subjectivity. These linguists coined the word "subjectivity" for the first time Quirk et al. (1985). It is impossible to see or objectively verify "private states," which include thoughts, feelings, and intuitions. Even the idea of a private state raises concerns about sentiment analysis. Depending on the setting, talks' subjective connotations can vary widely and frequently take on odd shapes. However, keep in mind that just because something is perceived subjectively doesn't mean it's necessarily incorrect Wiebe et al. (2004). Using text analysis and natural language processing techniques, sentiment analysis identifies and extracts evaluation information from the textual content. In this essay, Hussein (2018) , investigated the issues that have been connected to various sentiment analysis methodologies. Although Mary has strong opinions on chocolate, it's not always accurate to say that she loves it. The same is true for claims that seem to be unbiased. Successful targeted COVID-19 immunization efforts require an understanding of the factors that contribute to vaccine scepticism in low-coverage areas. Lanyi et al. (2021) investigated the use of AI and soft intelligence together in the field of public health. As a proof of concept, we used a set of geo-located tweets from London, UK to analyze using natural language processing (NLP) technology, and we found and evaluated key hurdles to vaccine uptake. Based on the algorithm's selection of topics and sentiment, the platform then selected 913 tweets with negative sentiment from the top 12 subject clusters. So that more qualitative research could be conducted, it was decided to parse apart these tweets. Much research has revealed that a large number of Twitter users disseminate erroneous information concerning vaccines. Despite how crucial this is, authors still offer reviews, ratings, comments, and recommendations as the internet fast expand to include websites, social networks, blogs, and online portals. This author's work is heartfelt and covers a wide range of topics, including but not limited to people, places, and things. For businesses, governments, and society as a whole, these mindsets are extremely helpful. The majority of this content was created by writers, thus any value must be derived through text mining and sentiment analysis. However, rating and sentiment analysis are not without their challenges. These challenges provide barriers to precise polarity assessment and thorough emotion analysis.

2.3 Covid-19 booster shots vaccines

Since many people lacked education and have COVID anxiety, they did not respond well to vaccines. This reluctance hinders global attempts to stop COVID-19. Social media can help distribute information and affect public opinion, according to studies.Cascini et al. (2022) developed sentiment analysis software to evaluate how social media platforms may affect COVID-19 vaccination views and public health strategies to overcome vaccine reluctance. Sarlan et al. (2014) described a sentiment analysis that could collect data from several tweets. After separating positive and negative tweets, opinions were shown in a pie chart and HTML webpage.Ansari and Khan (2021) used theme sentiment analysis, emotional analysis, and demographic interpretation to investigate how the general community thinks about the COVID-19 vaccination. We also ran sentiment categorization experiments to better understand gender and geography. Although most tweets were unfavourable, scientific research shows a consistent worldwide health trend toward vaccination.Sediqin (2021) scraped 340 million COVID-19 tweets between December 2020 and March 2021. They studied how people react to tweets regarding masks and immunizations. This study sought reliable information about COVID-19 vaccinations. People's reactions to government orders to get vaccinations and wear masks have been studied extensively. The author was interested in tweets with high retweets and reply rates. Starting with retweets, the dataset was separated into three categories. Otherwise, a tweet is inactive. Topic modelling determined the most popular tweet topics across all categories. VADER analyzed tweets without the hashtags "vaccine" and "mask." The fraction of favourable to negative tweets fluctuates moderately but substantially with the hashtags "vaccine" or "mask." Some people oppose the COVID-19 vaccine and face masks and refuse to get the shot or wear one. Twitter users' sentiments varied from cheerful to negative over time. Due to credible organizations mentioning COVID hazards and the popularity of "vaccine" and "mask" tweets, people are less likely to get a third booster shot. Using a convolutional neural network (CNN) model using respiratory sound parameters, we categorized COVID-19 and immunization sounds for recognizing COVID-19-positive symptoms, enhancing diagnosis accuracy. The proposed solution outperforms the state-of-the-art on IEMOCAP, RAVDESS, and EMODB.Kwon et al. (2021) used MLP and CNN to better recognize emotions in spoken English. This MLT-SER system uses a 1D CNN to learn emotional functions from speech inputs. This model requires time to study and test, but it's trustworthy and effective, making it ideal for real-time speech processing. Table 1 summarizes the main advantages and disadvantages of machine learning and deep learning methods for categorizing text/sentiment. This table summarizes the model's merits, weaknesses, sources, and publication year. The literature describes Machine Learning models using low-level lexical characteristics, high-frequency word features, syntactic features, support vector machines, naive Bayes, boosted trees, and random forests. ANN, CNN, Capsule networks, DenseNet, VGG-16, and BERT perform better than ordinary ML models (Table 1). Catal and Nangir (2017) employed a voting mechanism among many classifiers to increase accuracy. But Deep Learning strategies have improved. Our research uses hybrid methodologies, but to improve accuracy, we used multiple deep-learning models.

2.4 Comparative Analysis between RoBERTa Model and other Model in sentiment analysis

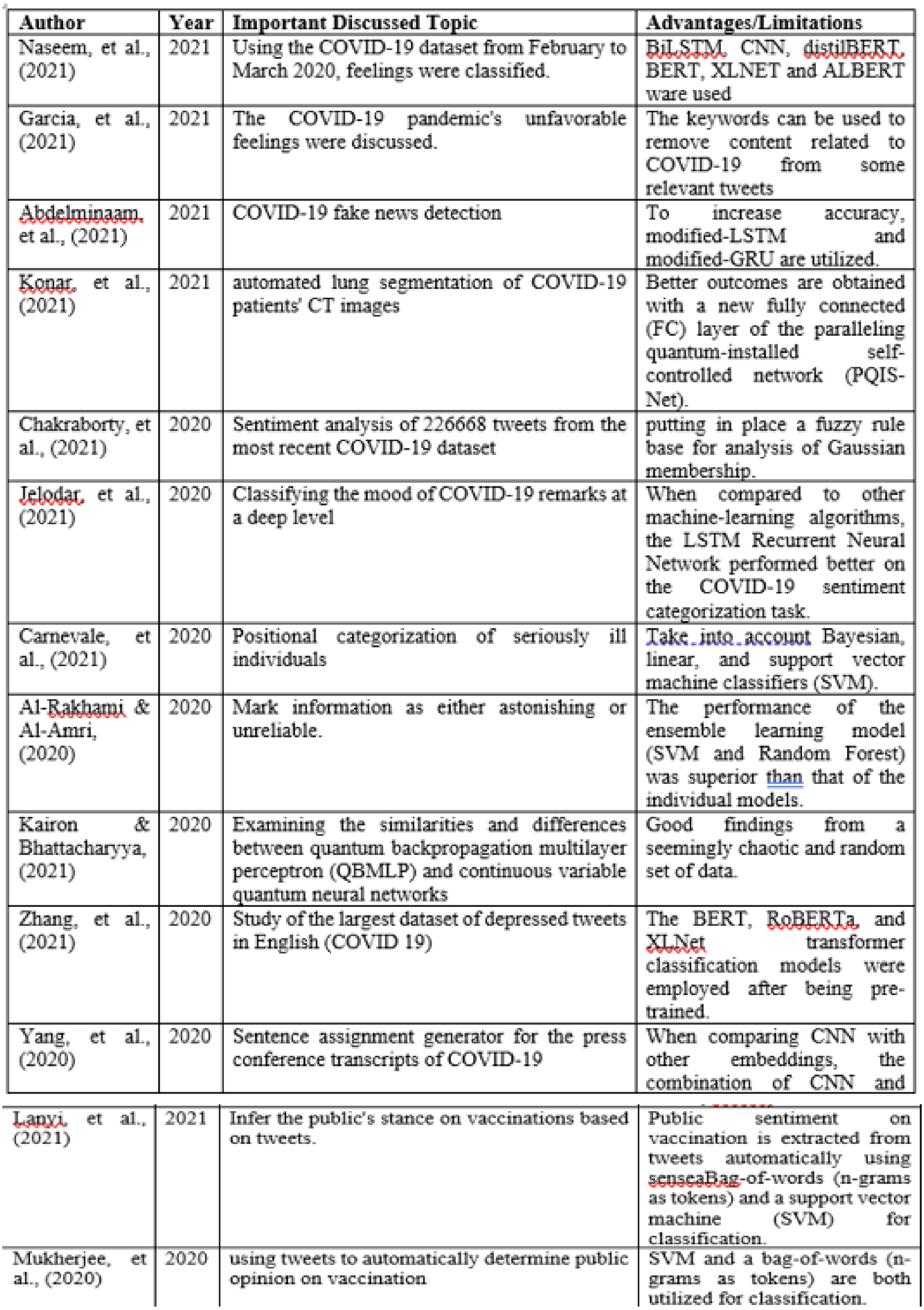
The sentiment analysis systems that were surveyed are summarized in Table 1 in relation to COVID-19 booster doses and immunizations. The development of Arabic-language sentiment analysis systems is depicted in the table below as progressing slowly. This issue is exacerbated by a number of challenges, including the diversity of Arabic dialects and the paucity of pertinent data. These barriers make deep learning techniques for Arabic sentiment analysis uncommon, particularly at the micro level.

2.4.1 BERT

The BERT pre-trained model and bagged-SVM classifier were used by Kazameini et al. (2020) to extract contextualized word embedding from text data. It was fuelled with essays. Divided preprocessed essays were fed into a BERT base model. Ten SVM classifiers made predictions based on the feature vectors of the document. A majority decided. The group's output went up by 1.04 per cent. In 2020, Yang et al. (2020). suggested using the AraBERT model to categorize Arabic COVID-19 tweets. A BERT-based transformer that was trained using Arabic corpus makes up the suggested model. An analysis of the SenWave dataset produced an F1-Macro score of 0.52. The 11 emotion classes in 10,000 Arabic tweets about COVID-19 were annotated. Another multi-label emotion detection technique is suggested by COVID-19. Using the suggested methodology, tweets about India during the COVID-19 pandemic were examined. BERT and RoBERTa have received training using English SenWave. Evaluation techniques for Arabic COVID-19 tweets are scarce. Machine learning algorithms were used in two methods to categorize tweets as good, bad, or neutral. SVMs were trained on 10,623 tweets with positive, negative, or neutral labels by Aljameel et al. (2021) 0.84 Fi-score (TF-IDF).

2.4.2 RoBERTa

In recent work, two deep learning methods for perfectly alright sentiment classification were created Kabir and Madria (2021). 10 different emotional labels were utilized to categorize COVID-19 tweets using these models. In the first model, the key term responsible for a tweet's emotional impact was identified using a customized Q&A RoBERTa head. Since the second model makes use of BiLSTM and incorporates information from the attention layer and additional features, it is advised for the classification of emotions. Their data show a transition during the epidemic from sadness to a more upbeat view. Some have even claimed that the improved level of language comprehension offered by the Deep Bidirectional Representations from Transformers (BERT) pre-trained model led to the rediscovery of the Natural Language Processing (NLP) pipeline Devlin et al. (2018). This is because it uses the vanilla Transformer word embeddings from Google, which was introduced in 2018.Al-Rfou et al. (2019) Tenney et al. (2019) . The BERT model has issues with word-piece embedding, computational complexity, and fixed input-length size constraints. To address the numerous underlying issues with BERT, the Extended Auto Regression Pre-training for Language Understanding (XLNet), the Robustly Improved BERT Pre-Training Approach (RoBERTa), and the DistilBERT pre-trained models were



Author	Year	Important Discussed Topic	Advantages/Limitations
Naseem, et al., (2021)	2021	Using the COVID-19 dataset from February to March 2020, feelings were classified.	BiLSTM, CNN, distilBERT, BERT, XLNET and ALBERT ware used
Garcia, et al., (2021)	2021	The COVID-19 pandemic's unfavorable feelings were discussed.	The keywords can be used to remove content related to COVID-19 from some relevant tweets
Abdelminaam, et al., (2021)	2021	COVID-19 fake news detection	To increase accuracy, modified-LSTM and modified-GRU are utilized.
Konar, et al., (2021)	2021	automated lung segmentation of COVID-19 patients' CT images	Better outcomes are obtained with a new fully connected (FC) layer of the paralleling quantum-installed self-controlled network (PQIS-Net).
Chakraborty, et al., (2021)	2020	Sentiment analysis of 226668 tweets from the most recent COVID-19 dataset	putting in place a fuzzy rule base for analysis of Gaussian membership.
Jelodar, et al., (2021)	2020	Classifying the mood of COVID-19 remarks at a deep level	When compared to other machine-learning algorithms, the LSTM Recurrent Neural Network performed better on the COVID-19 sentiment categorization task.
Carnevale, et al., (2021)	2020	Positional categorization of seriously ill individuals	Take into account Bayesian, linear, and support vector machine classifiers (SVM).
Al-Rakhami & Al-Amri, (2020)	2020	Mark information as either astonishing or unreliable.	The performance of the ensemble learning model (SVM and Random Forest) was superior than that of the individual models.
Kairon & Bhattacharyya, (2021)	2020	Examining the similarities and differences between quantum backpropagation multilayer perceptron (QBMLP) and continuous variable quantum neural networks	Good findings from a seemingly chaotic and random set of data.
Zhang, et al., (2021)	2020	Study of the largest dataset of depressed tweets in English (COVID 19)	The BERT, RoBERTa, and XLNet transformer classification models were employed after being pre-trained.
Yang, et al., (2020)	2020	Sentence assignment generator for the press conference transcripts of COVID-19	When comparing CNN with other embeddings, the combination of CNN and
Lanyi, et al., (2021)	2021	Infer the public's stance on vaccinations based on tweets.	Public sentiment on vaccination is extracted from tweets automatically using senseaBag-of-words (n-grams as tokens) and a support vector machine (SVM) for classification.
Mukherjee, et al., (2020)	2020	using tweets to automatically determine public opinion on vaccination	SVM and a bag-of-words (n-grams as tokens) are both utilized for classification.

Figure 1: Summary of Covid 19 related Text Sentiment classification based papers

all put forth. No progress has been made in identifying emotions from texts, despite the fact that BERT and its offshoots have been used thoroughly for information retrieval (QA), natural language inference (NLI), text summarization (TS), and other natural language processing (NLP) tasks dealing with human-related and environmental problems. Using RoBERTa models, sentiment analysis of tweets about covid-19 booster injections is examined here.

2.4.3 ISEAR

Alotaibi suggested using supervised logistic regression to identify written sentiment. Both practice and evaluation used ISEAR's data. Training a logistic regression model with emotional utterances and labels. The trained classifier only saw unseen emotion-labelled texts during testing. The Researchers used precision, recall, and F1-score to assess their model. Happiness, fear, sadness, humiliation, and guilt all had F1 values of 0.76, 0.64, 0.73, 0.62, and 0.57, respectively. They indicated that designing categorization characteristics would be easier with a deep learning model. Bi-LSTM, Self-attention, and Convolutional Neural Networks were merged by Polignano et al. (2019) in 2019. To recognize text emotions more accurately, word embedding extraction was essential. A Bi-LSTM, a CNN ensemble, and a Self-Attention model were used to test Google, GloVe, and Fast-Text embeddings. Data from ISEAR, SemEval-2018 Task 1, and SemEval-2019 Task 3 were used to test their model. Performance for all datasets was enhanced using Fast Text embedding. To enhance model performance, they suggested using a powerful pre-trained embedding.

2.4.4 Data Processing

According to Scherer and Wallbott (2017), the ISEAR dataset is a freely accessible dataset built from 37 country-specific questionnaire studies aimed at bridging cultural gaps. There are a total of 7666 sentences categorized according to seven feelings: happiness, rage, sadness, humiliation, guilt, surprise, and fear. It was chosen for this analysis because its balanced class feature is so useful for establishing broad, generalized predictions. The ISEAR dataset's data distribution with its quantity is anger at 1096; Disgust at 1096; sadness at 1096; shame at 1096; fear at 1095; joy at 1094; and guilt at 1093, with a total of 7666.

2.5 Summary of the Literature Reviewed

The objective of this study is to analyze public sentiment toward the third COVID-19 vaccine dose. The RoBERTa model was used to evaluate the majority of the data, which was collected via Twitter. However, raw data for the process will also come from tweets on Covid-19 vaccinations on Twitter. To begin with, these tweets will be cleaned up using natural language processing. The text data will be examined before being converted into polarity and subjectivity for categorization using the RoBERTa model. Data that is neutral or positive will be separated by the program. The tone of tweets addressing the use of additional doses of the COVID-19 vaccination, if necessary, will be influenced by these classifications. It would be crucial to educate people about the benefits of having a third dose of the COVID-19 vaccine as a booster shot. This study will be the starting point for an explanation of our ideas.

3 Methodology

3.1 System Overview

The methodologies used to reach the study's primary conclusions, as well as the infrastructure's design requirements and the results of data analysis, are presented here. Additional information on these models can be found in the subsequent paragraphs. Tweet Collecting Data & Building a Model Sections 3.2, 3.3, 3.4, and 3.5 cover the pre-processing, exploratory analysis, sentiment analysis of Covid-19, and design definition, respectively. Section 3.6 expands on the RoBERTa Model's explanation, whereas Section 3.7 analyzes the RoBERTa Model. Figure 2 shows the architectural diagram of RoBERTa and BERT model.

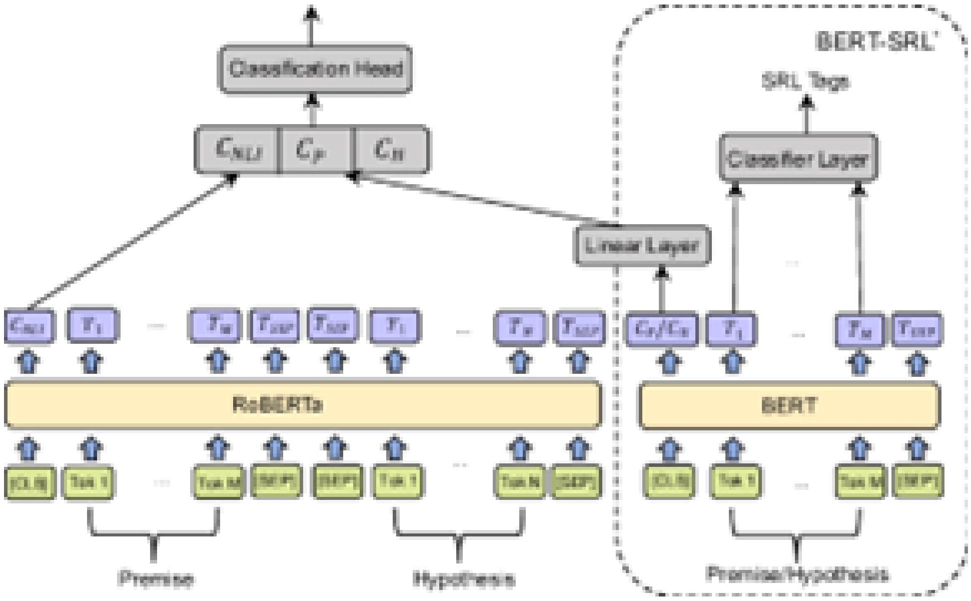


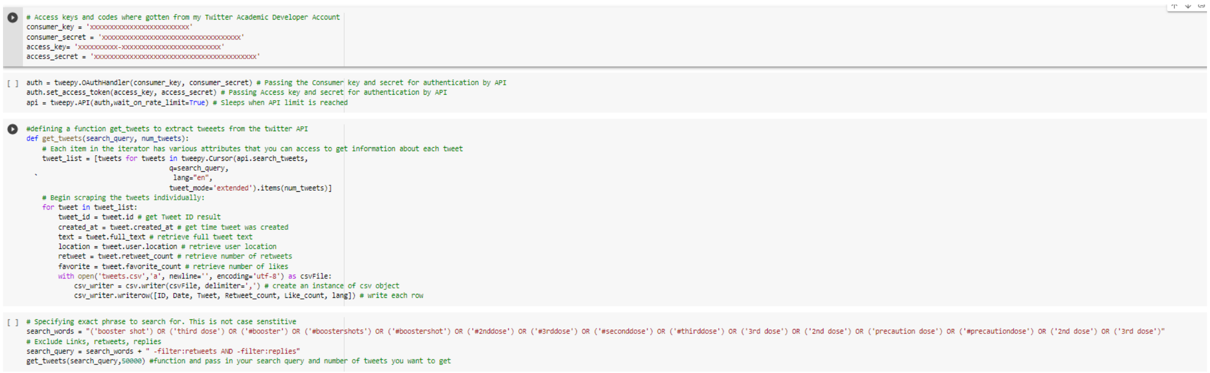
Figure 2: Architecture diagram of RoBERTa model and BERT Model

3.2 Tweet Data

3.2.1 Data Collection

Twitter's Academic API was used to retrieve tweet media data. Tweet data was extracted from 25 October 2022 to 1st December using hashtags like boosters, covid19, covid, vaccine, flu, vaccines etc. A Python script was written to retrieve tweets from the Twitter Academic API by authenticating the access codes and keys, and then a function get tweets were used to extract the tweets needed based on the search words passed. The results of the script written were saved as a CSV file. Although the virus was found in China in November 2019, the outbreak didn't start until January 2020. The recent timing was planned to coincide with the aftermath of the second and third-dose vaccine rollout. This investigation required the collection of 46,398 data points. The acquired tweets are then prepared for further analysis by following the steps below. Figure 1 depicts the sequential processes required to perform sentiment analysis on booster doses of the COVID-19 vaccine. TextBlob was utilized as a starting point for processing textual data.

It provides an easy-to-use API for developing simple sentiment analysis and classification natural language processing (NLP) applications. In addition, the following two data sources have undergone data transformation: A big data frame with the relevant columns for analysis was created by concatenating CSV files with the gathered and saved tweet data. The finding in this study debuted the feature decision.



```
# Access keys and codes where gotten from my Twitter Academic Developer Account
consumer_key = 'xxxxxxxxxxxxxxxxxxxxxxxxx'
consumer_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_key= 'xxxxxxxxxxx-xxxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
```

```
auth = tweepy.OAuthHandler(consumer_key, consumer_secret) # Passing the Consumer key and secret for authentication by API
auth.set_access_token(access_key, access_secret) # Passing Access key and secret for authentication by API
api = tweepy.API(auth,wait_on_rate_limit=True) # Sleeps when API limit is reached
```

```
#defining a function get_tweets to extract tweeets from the twitter API
def get_tweets(search_query, num_tweets):
    # Each item in the iterator has various attributes that you can access to get information about each tweet
    tweet_list = [tweets for tweets in tweepy.Cursor(api.search_tweets,
                                    q=search_query,
  `                                  lang="en",
                                    tweet_mode='extended').items(num_tweets)]
    # Begin scraping the tweets individually:
    for tweet in tweet_list:
        tweet_id = tweet.id # get Tweet ID result
        created_at = tweet.created_at # get time tweet was created
        text = tweet.full_text # retrieve full tweet text
        location = tweet.user.location # retrieve user location
        retweet = tweet.retweet_count # retrieve number of retweets
        favorite = tweet.favorite_count # retrieve number of likes
        with open('tweets.csv','a', newline='', encoding='utf-8') as csvFile:
            csv_writer = csv.writer(csvFile, delimiter=',') # create an instance of csv object
            csv_writer.writerow([ID, Date, Tweet, Retweet_count, Like_count, lang]) # write each row
```

```
# Specifying exact phrase to search for. This is not case senstitive
search_words = "('booster shot') OR ('third dose') OR ('#booster') OR ('#boostershots') OR ('#boostershot') OR ('#2nddose') OR ('#3rddose') OR ('#seconddose') OR ('#thirddose') OR ('3rd dose') OR ('2nd dose') OR ('precaution dose') OR ('#precautiondose') OR ('2nd dose') OR ('3rd dose')"
# Exclude Links, retweets, replies
search_query = search_words + " -filter:retweets AND -filter:replies"
get_tweets(search_query,50000) #function and pass in your search query and number of tweets you want to get
```

Figure 3: Authenticating Twitter's Academic API

3.2.2 Data Preprocessing

Covid-19 subject pre-processing is a crucial first step. In this process, stop words and other meaningless words in the text are parsed out. Dimensional savings in feature storage were achieved by omitting the superfluous term. As a starting point for preparing the text, TextBlob was used. Twitter's pre-processing procedure is very similar. Before tokens can be lemmatized, they must be cleaned of stop words, this means taking the text from the English language. It assumes that people on social media utilize a more relaxed, emotion-laden variety of English. Achieving clean data and discovering actionable insights depends on the quality of test pre-processing. Python library created to pre-process tweets by removing hashtags and URL. The following further data cleaning operations were executed after the initial clean by the python library: Getting rid of URLs and punctuation, changing the text to lowercase, getting rid of numbers and words with numbers, getting rid of extra spaces and single-character words, and getting rid of English "halt words" Separate a sentence into its constituent words (tokenize); and Figure 3 depicts the lemmatization process, which groups many inflected forms of words into the root form.

3.3 Exploratory Analysis

To learn more about the gathered text, an exploratory analysis in Python was performed. This comes before later processes like topic extraction and sentiment analysis. Tweets length lesser than four words were dropped to ensure the accuracy of the sentiment analysis for positive, neutral, and negative

3.3.1 Word Frequencies

Using a word cloud, how often certain words appeared in the Twitter data was analyzed. Word Cloud is a visual representation of text data, where the relative size of individual

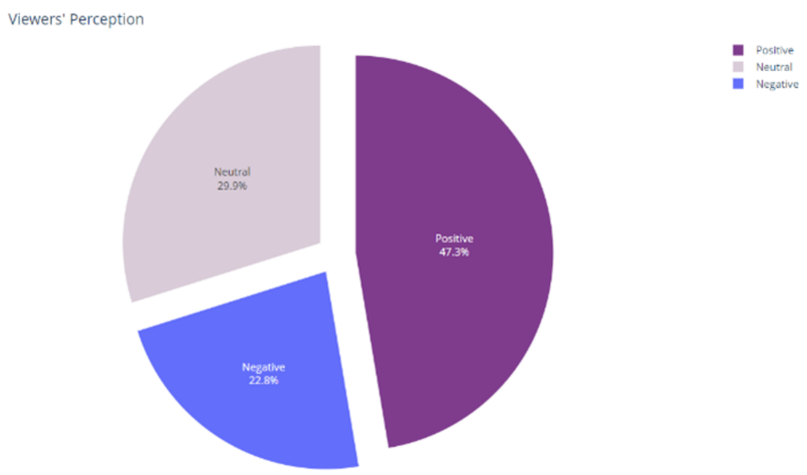


Figure 4: Pie chart of the classification of sentiments using TEXTBLOB

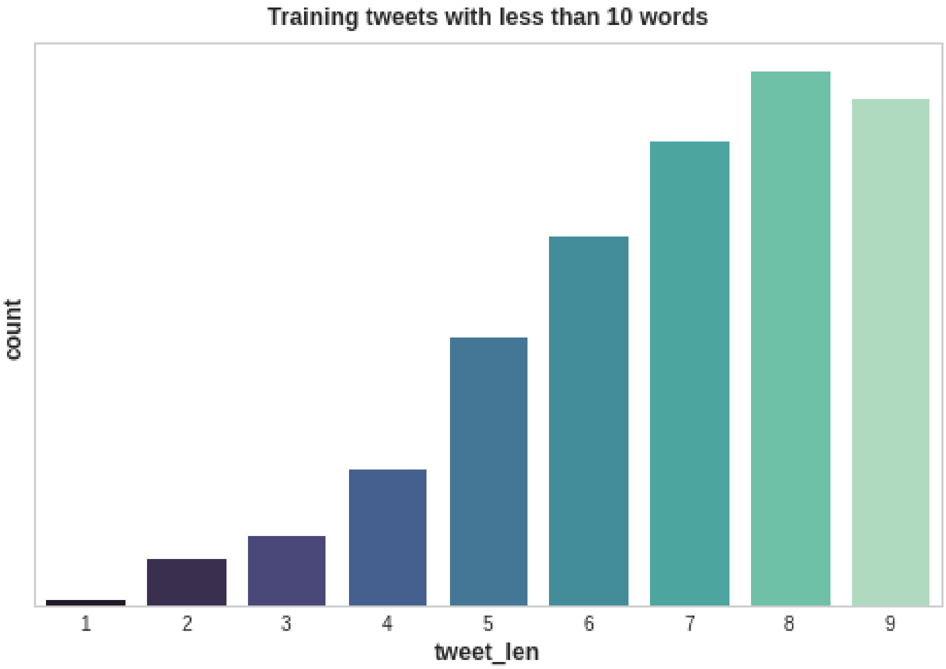


Figure 5: Tweets Length

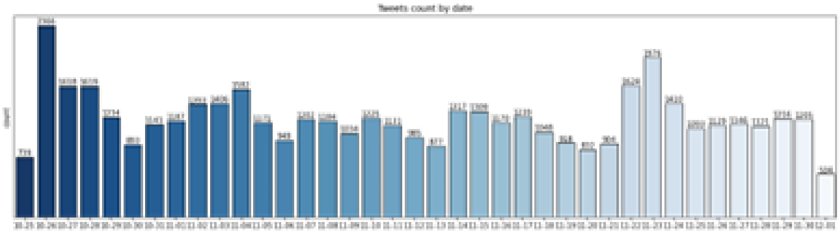


Figure 6: Diagram of a Twitter Tweets by date

words or phrases indicates their frequency of occurrence in the text. It's a straightforward way to see which words appear most often in the text. The analysis of social media text using the Word Cloud paradigm is quite common. The word "the" has the highest word count of 47142, while the word "covid" has the lowest word count of 15218, both numbers amongst others were found by using the Built-in Python collections module, this is shown in figure 5.

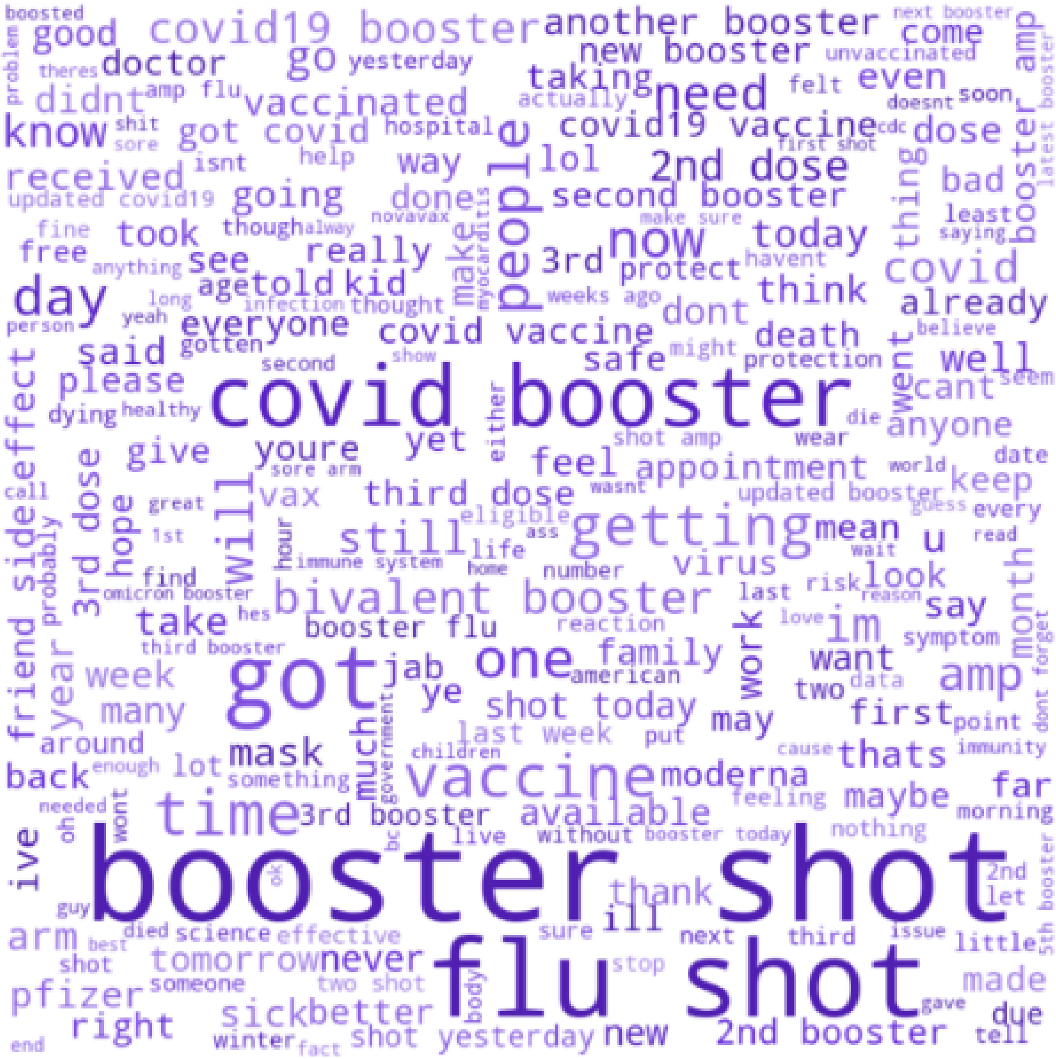


Figure 7: Word Cloud

3.3.2 Emotion Detection

To process textual information, you can use TextBlob, a Python (2 and 3) package. It offers a straightforward programming interface for applications for standard natural language processing tasks including part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, and others. The data on emotions was annotated with positive, neutral, and negative labels using TextBlob. Textblob is able to categorize and maintain a tally of the various feelings conveyed by the words extracted from processed text. To accomplish this, it follows these procedures: Get the right words out to describe how you feel; Analyze the feeling associated with each word, and keep a running tally of the words and the feelings associated with them.

3.4 Sentiment Analysis of Covid-19

Extraction of sentiment from the Tweets data was followed by a comparison of pre-and post-preprocessed sentiment results (polarity scores) to determine the efficacy of text pre-processing. Each Tweet's polarity score was determined with Python TextBlob's sentiment analysis capabilities. TextBlob was developed on top of NLTK to facilitate rapid experimentation. In this study, the unsupervised method was utilized to extract the mood from the text and to probe the relationship between the terms. The opinion is formed by linking the text's keywords to pre-existing vocabulary categories.

4 Design Specification

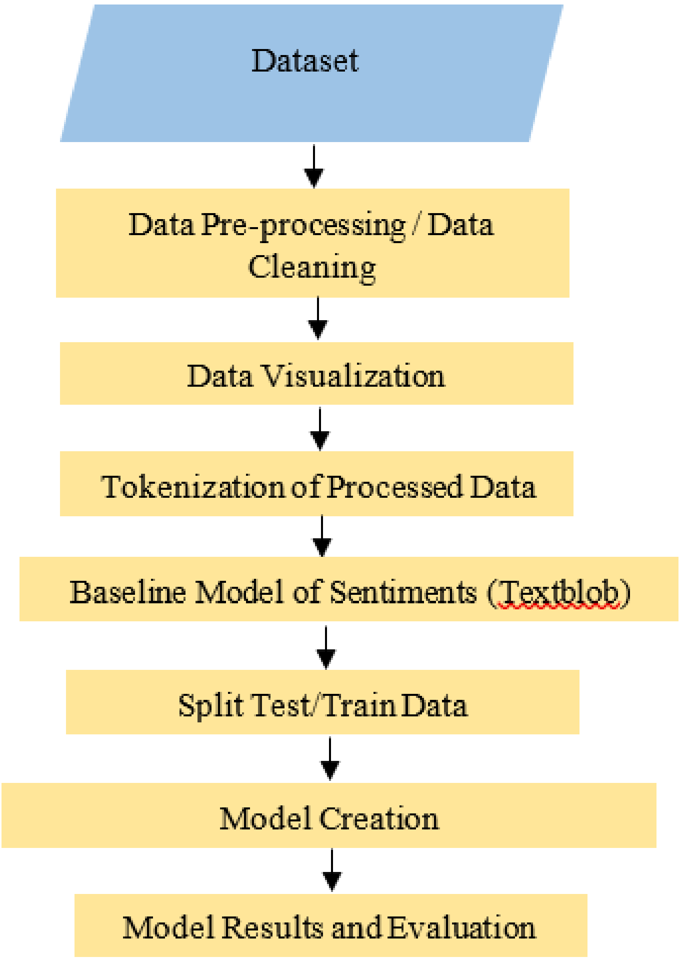

Dataset
Data Pre-processing / Data Cleaning
Data Visualization
Tokenization of Processed Data
Baseline Model of Sentiments (Textblob)
Split Test/Train Data
Model Creation
Model Results and Evaluation

Figure 8: Design Specification Diagram

In order to anticipate the sentiment of a tweet, this project performs a Sentiment Analysis utilizing the BERT and roBERTa algorithms. The tweets in question pertain to the event known as ”covid-19” (Positive, Negative or Neutral). In particular, the provided dataset will be used to fine-tune BERT in addition to ROBERTA, hence enhancing the model's overall performance. In particular, the submitted dataset will be used to fine-tune BERT and ROBERTA, hence improving the model's overall performance. Before the data is fed to the algorithms, the tweets will be thoroughly cleaned by removing links, hashtags at the end of the sentences, and punctuation. This will help the operating systems better grasp the content and enhance their ability to make accurate predictions. After the Textblob model was used to extract sentiments from tweets about covid-19 booster shots, the RoBERTa and BERT machine learning algorithm was used to forecast the sentiments of the basic model. Overall, there were 21730 tweets reflecting a happy mood, 1333 tweets reflecting a neutral disposition, and 10534 tweets reflecting an upbeat outlook. Figure 8 shows the design specification.

4.1 RoBERTa Model

Liu et al., (2020) say Roberta utilized the masking technique to predict omitted material. RoBERTa does this by rapidly increasing training mini-batches and learning speed. RoBERTa boosted BERT's output to expand hidden language modelling. The RoBERTa study used more data than the initial BERT study. During model training, batch size and learning rate were applied to the COVID-19 dataset (Figure 7). Precision, Recall, F1-Score, and Support analyze each combination's results. This RoBERTa model used 46,397 tweets.

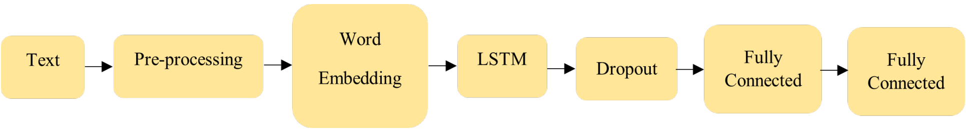
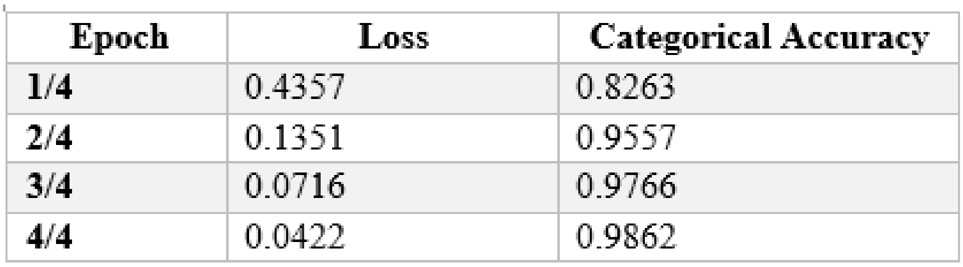


Figure 9: Block Diagram of the RoBERTa-based model

Figure 8 demonstrates the data validation, that is the data correctness and loss of the sentiment on covid-19 booster doses vaccinations validation rules are repeating programming sequences that verify the accuracy, relevance, and security of data.



Epoch	Loss	Categorical Accuracy
1/4	0.4357	0.8263
2/4	0.1351	0.9557
3/4	0.0716	0.9766
4/4	0.0422	0.9862

Figure 10: Demonstrating the data validation

4.2 BERT Model

A popular attention-based language interpretation model is BERT. By selecting left and right parameters in each layer, BERT may train bidirectional representations from unlabeled input. With one additional output layer, the pre-trained BERT model can be used to construct cutting-edge models for a variety of applications without substantial task-specific engineering changes. Before creating the model, tokenization and padding were applied. Together, token and mask inputs were considered model inputs. The text feature vectors were produced with BERT and sent to a 128-unit LSTM layer followed by a 20% dropout layer. After a fully linked layer with sigmoid activation, the model formed the output using characteristics.

`figures/figure 8 Block Diagram of Bert Model.png`

Figure 11: Block Schematic of the model's architecture.

5 Evaluation

Chrome browser was used for all of the experiments in this study through the Google CoLab interface. Graphics processing unit (GPU) type GP100-893-A1; 84GB RAM and

220GB ROM; 100% performance; NVIDIA Tesla P100-PCIE-16GB graphics card. In this section, we discuss datasets, parameter definitions, and model assessments. The proposed solution is also compared to industry norms. Practical implementation was carried out utilizing Python and its Huggingface module (Binder, et al., 2021). In order to fine-tune the initial models, the "ktrain" package (Maiya, 2020) was employed. We use the "AdamW" optimizer with varying learning rates (between 1e5 and 5e5) and list batch sizes (8, 16, 32). The models were tuned for 25 rounds before selecting the optimal checkpoint based on its ability to predict validation set data.

5.1 Dataset

The total data set used is 46,397 which was collected from Twitter API. This dataset has been used to determine the sentiment analysis on covid-19 booster shots vaccines. The complete details for test results for the BERT classifier and RoBERTa Classifier are shown in Tables 5 & 6. The support for the BERT and RoBERTa classification with respect to positive, neutral, and negative is 13337, 21730, and 10534 respectively.

5.2 Experiment Setup

For the model to be useful, a classifier must be used. As a result, we run a number of experiments with the deep learning classifiers RoBERTa and BERT to see how they perform. The performance of a model can be evaluated with the use of precision, F1-score, accuracy, and recall. When seen through the maze of confusion, their meaning becomes clear. 4.3 Sentiment Analysis Comparison Confusion Matrix RoBERTa's classification outcomes demonstrate a 99% accuracy. The accuracy of negative tweets is one hundred per cent, that of neutral tweets is one hundred per cent, and that of positive tweets is ninety-eight per cent. The closer the outcome is before and after preprocessing, the higher the percentage. Nonetheless, BERT's categorization findings demonstrate a 99% accuracy. If we further dissect this, we find that negative tweets have a 99% accuracy rate, neutral tweets have a 99% accuracy rate, and positive tweets have a 97% accuracy rate. The closer the outcome is before and after preprocessing, the greater the percentage.

5.2.1 Confusion matrix

There is a wide variety of names for what is commonly known as the confusion matrix or mistake matrix. The outcomes of the test algorithm can be summed up. This confusion matrix's rows stand for the tweets' predicted value, while the matrix's columns show what those tweets actually amounted to. The positive result is shown in the first line of the first column of Figure 10. (TP). For any given topic, it is possible to estimate the total number of tweets with high precision. An FP appears in the matrices in the two columns, first row. The number of tweets concerning a certain piece of data is often estimated in a way that is not entirely accurate. In the confusion matrix, false negatives are located in the second row of the first column (FN). It can't predict helpful tweets with any degree of accuracy. In the confusion matrix, the genuine negative appears in the second row of the second column (TN). It reliably estimates the number of meaningless tweets that will be sent out.

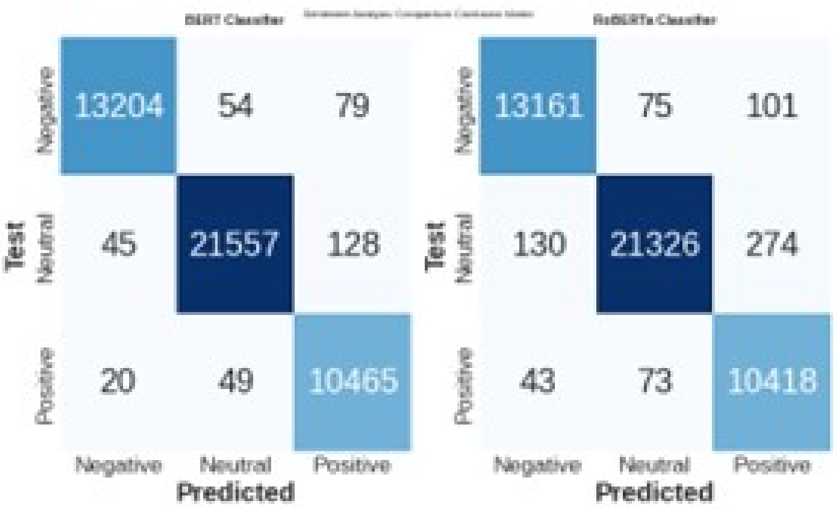


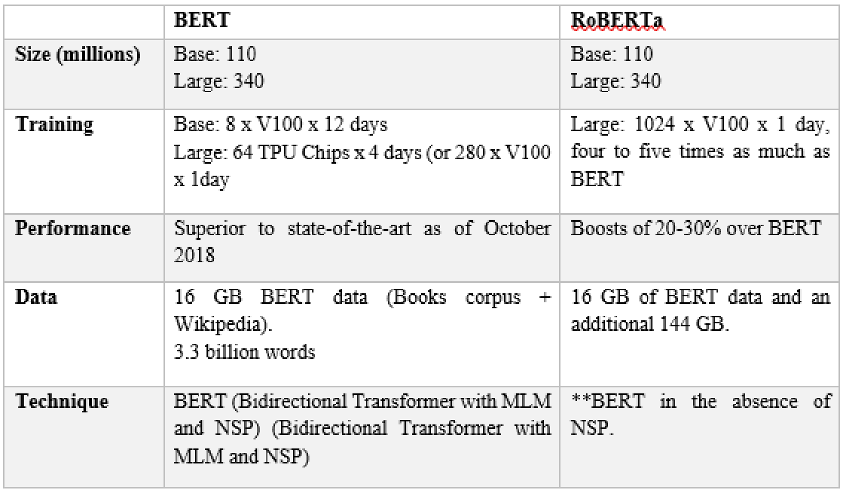
Figure 12: Sentiment Analysis Comparison Confusion Matrix with respect to positive, neutral and negative

5.2.2 Performance analysis

This section contains four subsections. In the first subsection, we describe how the machine learning model's performance stacks up against other similar solutions. In the following part, we'll look at how several deep learning models stack up against one another in terms of performance. In the third chapter, different ensemble deep learning models were examined, and in the last section, the proposed model was contrasted, majority voting-based ensemble deep learning, with the current approaches.

5.2.3 Comparison of RoBERTa and BERT Model

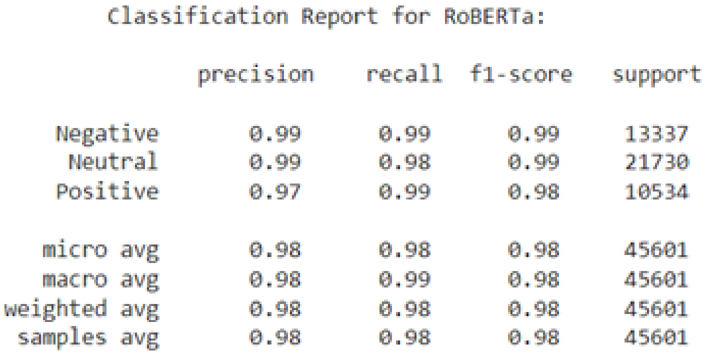
Here, Figure 13 compares the proposed model RoBERTa to the baseline model BERT in terms of accuracy, precision, F1-score, and recall.



	BERT	**RoBERTa**
Size (millions)	Base: 110 Large: 340	Base: 110 Large: 340
Training	Base: 8 x V100 x 12 days Large: 64 TPU Chips x 4 days (or 280 x V100 x 1day	Large: 1024 x V100 x 1 day, four to five times as much as BERT
Performance	Superior to state-of-the-art as of October 2018	Boosts of 20-30% over BERT
Data	16 GB BERT data (Books corpus + Wikipedia). 3.3 billion words	16 GB of BERT data and an additional 144 GB.
Technique	BERT (Bidirectional Transformer with MLM and NSP) (Bidirectional Transformer with MLM and NSP)	**BERT in the absence of NSP.

Figure 13: Comparison of RoBERTa and BERT Model

BERT is a powerful tool for fine-tuning machine-learning applications. It is a bidirectional transformer for pre-training over large amounts of unlabeled text input. BERT did better than the NLP state-of-the-art on a number of difficult tasks by using a bidirectional transformer, the new pre-training tasks of Masked Language Model and Next Structure Prediction, a lot of data, and Google's computing power. On the other hand, Facebook's RoBERTa (Robustly Optimized BERT Technique) retrains BERT with a better training method and a thousand times more data and computing power. RoBERTa improves BERT's training method by adding dynamic masking. This makes the masked token change over the course of the training epochs, which gets rid of the need for a Next Sentence Prediction (NSP) assignment before training. It was also found that training with bigger groups worked better. In particular, RoBERTa uses the same 16GB Books Corpus and English Wikipedia as BERT does in its pre-training phase. The 76 GB CommonCrawl News dataset and the 38 GB Stories from Common Crawl dataset were also used (31 GB). This combination and an unbelievable 1024 V100 Tesla GPUs allowed us to complete the pre-training for RoBERTa in a single day. This means that when it comes to analyzing sentiment, RoBERTa is more effective than BERT. Figure 12 from the RoBERTa Classification Report. Obtaining 99% F1-Score, 99% Precision, and 99% Recall for the positive measures, 99% F1-Score, 99% Precision, and 98% Recall for the neutral metrics, and 98% F1-Score, 97% Precision, and 99% Recall for the negative measures are all part of the report. Training with larger batch sizes was found to be more effective.



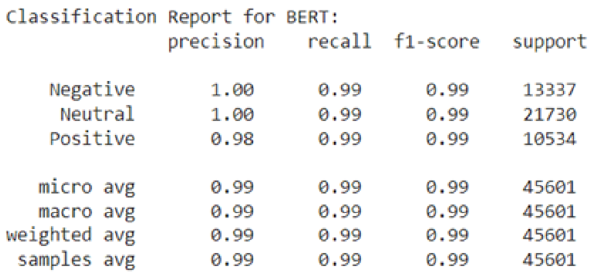
```
Classification Report for RoBERTa:

              precision    recall  f1-score   support

    Negative       0.99      0.99      0.99     13337
     Neutral       0.99      0.98      0.99     21730
    Positive       0.97      0.99      0.98     10534

   micro avg       0.98      0.98      0.98     45601
   macro avg       0.98      0.99      0.98     45601
weighted avg       0.98      0.98      0.98     45601
 samples avg       0.98      0.98      0.98     45601
```

Figure 14: Classification Report for RoBERTa

The following metrics are included in the report: sample average, weighted average, micro average, macro average, positive, neutral, and negative. Obtaining 99% F1-Score, 100% Precision, and 99% Recall for Positive Metrics, 99% F1-Score, 100% Precision, 99% Recall for Neutral Metrics, and 99% F1-Score, 98% Precision, and 99% Recall for Negative Metrics. To fine-tune for specific machine learning applications, As a bidirectional transformer, BERT is put to use in the pre-training phase of the learning process on large amounts of unlabeled textual input.

5.3 Real-time Application

Millions of people have died due to this epidemic, which has also caused a severe health crisis and economic downturn worldwide. It would have been simpler at this point to



```
Classification Report for BERT:
              precision    recall  f1-score   support

    Negative       1.00      0.99      0.99     13337
     Neutral       1.00      0.99      0.99     21730
    Positive       0.98      0.99      0.99     10534

   micro avg       0.99      0.99      0.99     45601
   macro avg       0.99      0.99      0.99     45601
weighted avg       0.99      0.99      0.99     45601
 samples avg       0.99      0.99      0.99     45601
```

Figure 15: Classification Report for BERT

acquire organized social media data for the victims, governments, and NGOs. It is imperative that strict quality control be performed on tweets in order to guarantee that only valuable content from the most popular sites gets shared. The study aims to employ sentiment analysis to identify inaccurate and misleading material regarding covid-19 booster injectable vaccines and halt its spread online. RoBERTa and the BERT Model are used to extract informative tweets, and real-time tweet classification using SAP is taken into consideration. One of the finest approaches for expressing desires is sentiment analysis, which converts information from unstructured words into a structured manner.

This SAP seeks to extract the most common words from informative tweets and categorize the moods into good, negative, and neutral tweets. The apt text processor is the Natural Language Toolkit package.

6 Conclusion and Future Work

If false information spreads about a worldwide pandemic like COVID-19, it might have devastating effects on people's health and well-being. Therefore, understanding individual perspectives behind Covid -19 Vaccinations would clarify people's drawbacks to getting vaccinated. Identifying bogus news by hand can be a laborious and challenging process. The study shows how Both BERT and ROBERTA models are excellent models for distinguishing between sentiments behind a text, all with the goal of facilitating the straightforward, efficient, and trustworthy identification of information and reason behind a text. Even though both models performed admirably, this study found that BERT Algorithm had a better score in predicting sentiments behind a tweet. Additionally, BERT and ROBERTA models are incorporated into this study, and results are shown to be comparable to those of state-of-the-art models.

, In essence, based on the research carried out in this project, we can agree that a good number of individuals are on board with taking booster shots if the need arises. A 4-epoch-trained test model was then applied to the Tweet data to perform sentiment analysis on covid-19 booster injections. More tweets can be considered for future studies. The potential impacts of COVID-19 on the economy, the job market, and individuals' daily lives could be analyzed and predicted using machine learning techniques. Furthermore, understanding the sentiments around Covid-19 Booster shot Vaccination could serve as a blueprint for vaccination administration for any other pandemic if the need arises.

References

Al-Rfou, R., Choe, D., Constant, N., Guo, M. and Jones, L. (2019). Character-level language modeling with deeper self-attention, *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, pp. 3159–3166.

Aljameel, S. S., Alabbad, D. A., Alzahrani, N. A., Alqarni, S. M., Alamoudi, F. A., Babili, L. M., Aljaafary, S. K. and Alshamrani, F. M. (2021). A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent covid-19 outbreaks in saudi arabia, *International journal of environmental research and public health* **18**(1): 218.

Ansari, M. T. J. and Khan, N. A. (2021). Worldwide covid-19 vaccines sentiment analysis through twitter content., *Electronic Journal of General Medicine* **18**(6).

Cascini, F., Pantovic, A., Al-Ajlouni, Y. A., Failla, G., Puleo, V., Melnyk, A., Lontano, A. and Ricciardi, W. (2022). Social media and attitudes towards a covid-19 vaccination: A systematic review of the literature, *EClinicalMedicine* p. 101454.

Catal, C. and Nangir, M. (2017). A sentiment classification model based on multiple classifiers, *Applied Soft Computing* **50**: 135–141.

Coleman, A. (2020). Hundreds dead" because of covid-19 misinformation, *BBC News* **12**.

Das, S. D., Basak, A. and Dutta, S. (2021). A heuristic-driven ensemble framework for covid-19 fake news detection, *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, Springer, pp. 164–176.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges, *Journal of King Saud University-Engineering Sciences* **30**(4): 330–338.

Islam, M. S., Sarkar, T., Khan, S. H., Kamal, A.-H. M., Hasan, S. M., Kabir, A., Yeasmin, D., Islam, M. A., Chowdhury, K. I. A., Anwar, K. S. et al. (2020). Covid-19–related infodemic and its impact on public health: A global social media analysis, *The American journal of tropical medicine and hygiene* **103**(4): 1621.

Kabir, M. Y. and Madria, S. (2021). Emocov: Machine learning for emotion detection, analysis and visualization using covid-19 tweets, *Online Social Networks and Media* **23**: 100135.

Kazameini, A., Fatehi, S., Mehta, Y., Eetemadi, S. and Cambria, E. (2020). Personality trait detection using bagged svm over bert word embedding ensembles, *arXiv preprint arXiv:2010.01309* .

Kwon, S. et al. (2021). Att-net: Enhanced emotion recognition system using lightweight self-attention module, *Applied Soft Computing* **102**: 107101.

Lanyi, K., Green, R., Craig, D. and Marshall, C. (2021). Covid-19 vaccine hesitancy: analysing twitter to identify barriers to vaccination in a low uptake region of the uk, *Frontiers in Digital Health* **3**.

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., Ekbal, A., Das, A. and Chakraborty, T. (2021). Fighting an infodemic: Covid-19 fake news dataset, *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, Springer, pp. 21–29.

Polignano, M., Basile, P., de Gemmis, M. and Semeraro, G. (2019). A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention, *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pp. 63–68.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985). A contemporary grammar of the english language.

Saha, S., Nie, Y. and Bansal, M. (2020). Conjnli: Natural language inference over conjunctive sentences, *arXiv preprint arXiv:2010.10418* .

Sarlan, A., Nadam, C. and Basri, S. (2014). Twitter sentiment analysis, *Proceedings of the 6th International conference on Information Technology and Multimedia*, IEEE, pp. 212–216.

Scherer, K. R. and Wallbott, H. (2017). International survey on emotion antecedents and reactions (isear)(1990).

Sediqin, M. (2021). *Semantic Analysis of Vaccine and Mask Sentiments in COVID-19 Twitter Data*, PhD thesis, The University of Wisconsin-Milwaukee.

Tasnim, S., Hossain, M. M. and Mazumder, H. (2020). Impact of rumors and misinformation on covid-19 in social media, *Journal of preventive medicine and public health* **53**(3): 171–174.

Tenney, I., Das, D. and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline, *arXiv preprint arXiv:1905.05950* .

Wiebe, J., Wilson, T., Bruce, R., Bell, M. and Martin, M. (2004). Learning subjective language, *Computational linguistics* **30**(3): 277–308.

Xiong, J., Lipsitz, O., Nasri, F., Lui, L. M., Gill, H., Phan, L., Chen-Li, D., Iacobucci, M., Ho, R., Majeed, A. et al. (2020). Impact of covid-19 pandemic on mental health in the general population: A systematic review, *Journal of affective disorders* **277**: 55–64.

Yang, Q., Alamro, H., Albaradei, S., Salhi, A., Lv, X., Ma, C., Alshehri, M., Jaber, I., Tifratene, F., Wang, W. et al. (2020). Senwave: Monitoring the global sentiments under the covid-19 pandemic, *arXiv preprint arXiv:2006.10842* .

Zhang, L., Wang, S. and Liu, B. (2018). Deep learning for sentiment analysis: A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(4): e1253.