ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 02, February 2025

# Sentiment Analysis of COVID-19 Variant Omicron from Twitter Tweets

[1] Hrithick Kanagaraj, [2] K Pratheksha, [3] N Kanagaraj, [4] OB Niviya

[1] Department of Metallurgical and Material Science Engineering National Institute of Technology Surathkal, Karantaka, India
[2] Software Engineer Maersk Global Service Centre Bangalore, Karnataka, India
[3] Senior Software Architect Cleveland Cliffs West Chester, Ohio, USA
[4] Department of Electronics and Communication Engineering, Kongu Vellalar Institute of Technology Erode, Tamil Nadu, India

Corresponding Author Email: [1] hrithick.kanagaraj@gmail.com, [2] prathe.sk@gmail.com, [3] n.kanagaraaj@gmail.com, [4] niviya.ob@gmail.com

***Abstract*— COVID-19 is a life changing incident for the whole world. Its variant Omicron has caused vivid disturbances and health issues for millions of people around the world. To get more idea about the people's thoughts, there is a need for sentiment analysis module to help the government and health sectors for improvising to stop the spread of Omicron. Twitter is a high level micro blogging site to gather public's views about Omicron. The tweets are gathered from twitter for the keyword Omicron, these text data are labeled according to its sentiment using TextBlob algorithm. This labeled data undergoes various trainings using a few classifier models. The deep learning models like LSTM, BiLSTM, BERT and Roberta are trained for tweets dataset. Roberta outperforms due to its unique feature of dynamic masking and the usage of transformer models in it with an accuracy of 91%.**

***Index Terms*— Omicron; sentiment analysis; classification; accuracy; sentiment prediction.**

## I. INTRODUCTION

Covid-19 is an illness in humans caused by corona virus that causes respiratory illness and deaths in large amounts. It was identified in China originally in 2019 and caused a pandemic across the world in 2020. It has led to a dramatic loss of human life worldwide and unprecedented challenge to public health. This pandemic has affected millions of people in risking the loss in their livelihoods. Informal economy workers are particularly vulnerable because the majority lacks social protection and access to quality health care and have lost access to productive assets.

Tracking the pandemic, advising on critical interventions, shipping diagnostics and personal protective equipment, and disseminating information to the public on how to protect themselves and others are totally necessary. WHO is also working in collaboration with scientists, manufacturers, and global health organizations through the ACT-Accelerator on development, production, and equitable access to COVID-19 tests, treatments, and vaccines.

The mutation of the virus that is the variant had started to emerge and cause more problems. For around more than two years, many new variants of corona viruses have been detected in risking the lives of people. A variant is where the virus contains at least one new change to the original virus. The variants that had caused severe damage to the public are Delta and Omicron. These variants are spreading across the people one after the other in large amounts causing more damage. Omicron was first spotted in South Africa in the early 2022. The Omicron variant of the corona virus SARS-CoV-2 has spread around the world faster than any other previous versions.

Twitter sentiment analysis is an area that has recently captured researchers' interest. Twitter is a famous micro blogging website where users may share their thoughts and opinions. Sentiment analysis in Twitter addresses the issue of analyzing tweets in terms of the statements made in them. These corporations frequently monitor user reactions and respond to them on micro blogs. Building technology to detect and summarize overall sentiment is one challenge. But the views of a large pandemic in this technological world are necessary to understand and judge the mental health of people about the new variants.

The tweets are considered as the emotions and views of people that is been put out publicly. These tweets are analyzed for understanding the sentiments of people about the pandemic. Models are trained for labeling "tweets" as positive, negative, or neutral and classification models are applied. This type of a system can help researchers analyze the mentality of people about the pandemic and health departments incur about the needs of the people. The journalists can provide quality helpful information about the disease by verifying it with the reality of the public's views.

Recent Artificial Intelligence induced health industry needs preventive measures for their advancement in the technology of their interest. Twitter is a platform that allows the usage of tweets for analyzing the sentiments of people about any disease or trauma caused worldwide using deep learning approaches.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 02, February 2025

## II. RELATED WORK

Sentiment analysis is an essential part for the medical industry and the health sectors for a pandemic from the most popular micro blogging platform. Python libraries like tweepy allows collecting tweets from twitter for our exploratory data and unique feature of analyzing the sentiments and the need for prediction of emotion of people based on their views published. Twitter data is the sole data to analyze people's sentiments about a pandemic and its functioning will upgrade the performance of the health sectors. Hence predictive sentiment analysis becomes the routine workflow in the pandemic control maintenance. The data collected from twitter about a variant is examined for sentiment analysis using deep learning and artificial intelligence techniques. The main issue in application of these techniques is to categorize them based on sentiment and need pre-processing before the application of the classification algorithms.

Employing natural language processing, clustering, and sentiment analysis techniques (Abraham C. Sanders et al. 2021) organized tweets relating to mask-wearing into high-level themes, then relay narratives for each theme using automatic text summarization. Topic clustering based on mask-related Twitter data offers revealing insights into societal perceptions of COVID- 19 and techniques for its prevention. It is a valuable tool that can aid health providers and policy makers in understanding public response to health interventions in the ongoing global health crisis. Thematic clustering and visualization based on mask-related Twitter data can offer distinct insights into societal perceptions of COVID-19, complementary to findings from more traditional epidemiological data sources.

Analysing the sentiments of different people's opinion (Chinder Kaur et al. 2020) fetched the twitter streaming tweets related to corona virus using twitter API and analyzed these tweets using machine learning techniques and tools as positive, negative, and neutral. Experiments through Python programming on different tweets using twitter API and NLTK library is used for pre-processing of tweets and then analyze the tweets dataset by using Textblob and after that show the interesting results in positive, negative, neutral sentiments through different visualizations.

Sentiments of the people from the USA and India had been analyzed by (Swati Sharma et al. 2020) using R Studio by text mining. The pre-processed tweets were scored and classified by polarity (positive or negative) and categorized into 10 different types of emotions using the R package titled "Gutenberger" and NRC emotion-based dictionary. Sentiment analysis was done by identifying the polarity. Emotions of words in tweets were analyzed as expressed in overall corpus review data by using NRCWord-Emotion Association Lexicon. The findings provide guidance to the policy makers to tailor their support policies in response to the emotional state of their people and also assist the marketers to tailor the communication strategies in the light of the emotional state of the target market.

During the global pandemic, many individuals as well as organizations and government agencies are posting their viewpoints regarding the corona virus. (Bishwo Prakash Pokharel et al. 2020) performed sentimental analysis of tweets of the twitter social media using python programming language with tweepy and textblob library. The tweets have been collected, pre-processed and then used for text mining and sentiment analysis using google colab. A graphical representation on the provided data has also been done after sentimental analysis.

Studying the eagerness and opinions of people to understand their mental state, (Mrityunjay Singh et al. 2021) performed sentiment analysis using the BERT model on tweets. A sentiment analysis on two data sets; one data set is collected by tweets made by people from all over the world, and the other data set contains the tweets made by people of India. The MCC validation accuracy is based on the Matthews correlation coefficient (MCC) that is a widely used statistical rate that generate a high score prediction results. Five metrics for performance analysis that are Average Likes over the period, Average Re-tweets over the period, Intensity Analysis, Polarity & Subjectivity, and Wordcloud were used.

Sentimental Analysis mines the opinion of human being for a viewpoint. In this Pandemic condition of Corona, whole world is sharing their opinions on the social media. (Supriya Raheja et al. 2021) showed an analysis of sentiments to get the opinion of people if they are positivity during this situation or not. The paper is using the technique of polarity to know the opinion is positive, negative, or nonpartisan. Three main keywords "COVID", "Corona virus" and "COVID-19" are used to check the polarity. The key point while performing sentimental analysis at as a rule is to recognize either sentence is subjective or objective. If the given line is delegated as objective, no other major undertakings are necessary, whereas if the line is delegated as subjective, its polarity (positive, negative, nonpartisan) must be known.

With the rapid increase in the use of the Internet, sentiment analysis has become one of the most popular fields of natural language processing (NLP). A study conducted by (Prasoon Gupta et al. 2021) to mine the sentiments of Indian citizens regarding the nationwide lockdown enforced by the Indian government to reduce the rate of spreading of Corona virus. The sentiment analysis of tweets posted by Indian citizens has been performed using NLP and machine learning classifiers. Data have been extracted from Twitter using Tweepy API, annotated using TextBlob and VADER lexicons, and preprocessed using the natural language tool kit provided by the Python. Eight different classifiers have been used to classify the data.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 02, February 2025

## III. EXISTING SYSTEM

Predictive analysis of pandemic using tweets in a large commercial way involves earlier detection of mentality of the public. Existing literature uses machine learning algorithms for the classification of sentiment analysis detection. The datasets used in the existing works does not suffer from sentiment classification problem. But in the proposed work, the real-world dataset of tweets suffer from classification problem. The positive, negative and neutral classes are not identified and categorized uniformly. The trained models are not compared sequentially to obtain the accurate results. This may leads to misclassification and analysis of people's sentiments detection about pandemic. So sentiment analysis techniques are applied to avoid data incorrect prediction and then apply the classification algorithms. The objectives of the proposed work are summarized as follow:

- To analyze the sentiments of people about Omicron.
- Analyze the sentiments and manifestations of a greater pandemic's varieties to keep an account of the fears among people.
- Sentiment analysis algorithm TextBlob is explored to address the polarity of the data and data labeling problem.
- Various deep learning classification models like LSTM, bi-directional LSTM, GRU, Bert and Roberta are trained for analyzing higher classification accuracy.

## IV. SYSTEM ARCHITECTURE AND PROPOSED WORK

### A. Sentiment Analysis Model

Users create "tweets" that are status messages on Twitter, a popular micro blogging site. These tweets occasionally convey opinions on various topics. The data about Omicron comprises of information gathered from twitter tweets in a regular basis. These are disappointment instances of the mentality during a negative cause worldwide, and the proposed work is to anticipate whether the received tweets are disappointment that is caused because of a particular variant of corona virus or a better improvement from the sentiment of the public. This might help in staying away from disappointment during pandemic and along these lines diminishing fear among public.

The obtained information about the Omicron tweets provides an unlabeled data distribution. So, there is a need to categorize the data among the given labels. The proposed work concentrates on the sentiment analysis by labeling into positive, negative and neutral categories according to its polarity and followed by the classification of the obtained labeled data about analyzing the sentiment on COVID-19 variant of Omicron in the given tweets. The overall flow of the proposed framework is shown in Figure1. To have the data labeled into three different classes like positive, negative and neutral, dataset containing tweets obtained from twitter has to undergo polarity analysis technique like TextBlob. After labeling the data, the data is trained for classification of prediction. Deep learning models like LSTM, Bi-directional LSTM, GRU, Bert and Roberta are tested for the prediction of the sentiments in the data sequence. The multi classification is employed for the prediction of sentiments in the tweets for further predictive classification and reduced disappointments and mishaps among people. The classes of data obtained from the tweets are positive, negative and neutral.

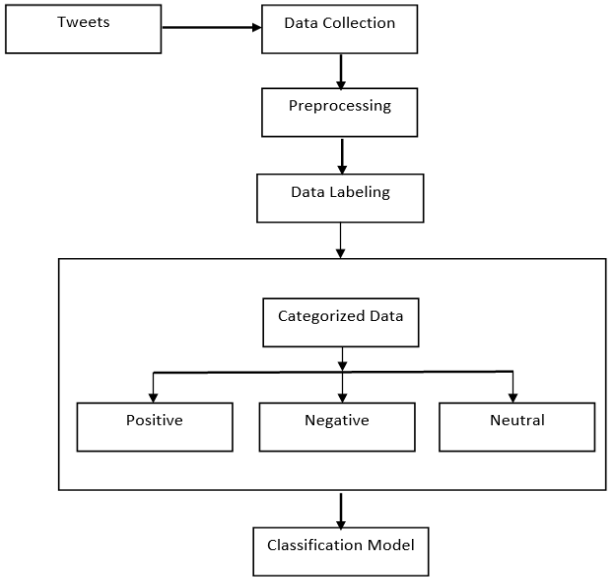


**Figure 1.** General process flow diagram

### B. Sentiment Analysis Labeling Techniques

Sentiment analysis typically refers to classifying tasks with respect to polarity where the data is split into classes are not represented before in the dataset. As the tweets are real time, categorizing data sets are a special case for sentiment analysis classification problem where the categorized class definition is not uniform among the dataset. Typically, they are composed of three classes for Omicron based analysis: positive class, negative class and neutral class. When observation is done among the tweets, it is categorized based the polarity of it. The following techniques can be employed to categorize the data based on sentiment. The data is categorized as shown in Figure 2.


**Figure 2.** Categorizing data of Omicron tweets

#### 1. TextBlob

TextBlob is a simple library which supports complex analysis and operations on textual data. For lexicon-based approaches, a sentiment is defined by its semantic orientation

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 02, February 2025

and the intensity of each word in the sentence. This requires a pre-defined dictionary classifying negative and positive words. It has a basic API for standard NLP tasks like part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation. TextBlob uses the averaging technique to determine a polarity score for a single word, and so a similar procedure is applied to every single word, resulting in a combined polarity for bigger texts as shown in Figure 3.


**Figure 3.** TextBlob labeling

Features of TextBlob are noun phrase extraction, part-of-speech tagging, sentiment analysis, classification (Naive Bayes, Decision Tree), tokenization (splitting text into words and sentences), word and phrase frequencies, parsing, n-grams, word inflection (pluralization and singularization) and lemmatization, spelling correction, add new models or languages through extensions and WordNet integration. Textblob can assist in getting started with NLP operations. The intensity is another parameter used by TextBlob to calculate subjectivity. The TextBlob library detects a text's language and utilises Google Translate to translate TextBlobs, sentences, and words into other languages.

### C. Sentiment Analysis Classification Model

The data is labeled by TextBlob based on its polarity. Once the data is labeled as positive, negative and neutral based on its sentiment analysis by polarity, the classification models are trained to check for predictions. The classification models are used to predict the analysis done by TextBlob is accurate and preferable. The deep learning classification models are LSTM, Bi-directional LSTM, GRU, Bert and Roberta.

### 1. Long Short-Term Memory

RNNs are a form of RNN that uses long short-term memory networks, or LSTMs for short. They were made to deal with the problem of long-term reliance. In a conventional RNN, the problem frequently emerges when connecting old knowledge to new information. If they could achieve this, RNN would be really useful. This problem is referred to as long-term dependency. In the repeating module of a traditional RNN, there is only one layer. The default behavior of the LSTM is to remember information for long periods of time. LSTM networks' memory module or repeating module differs from RNNs' memory module or repeating module.

A cell, an input gate, an output gate, and a forget gate make up a typical LSTM unit. The three gates control the flow of information into and out of the cell, and the cell remembers values across arbitrary time intervals. As there may be lags of undetermined duration between critical occurrences in a time series, LSTM networks are well-suited to categorizing, processing, and making predictions based on time series data. LSTMs were created to solve the problem of vanishing gradients that might occur when training traditional RNNs.

In order to change each weight of the LSTM network in proportion to the derivative of the error (at the output layer of the LSTM network) with respect to corresponding weight, an RNN using LSTM units can be trained in a supervised fashion on a set of training sequences, using an optimization algorithm like gradient descent combined with back propagation through time to compute the gradients needed during the optimization process.

LSTM networks are a sort of recurrent neural network that can learn order dependence in sequence prediction issues. This is a necessary characteristic in complicated problem fields such as machine translation, speech recognition, and others. LSTM networks are a sort of recurrent neural network that can learn order dependence in sequence prediction issues. This is a necessary characteristic in complicated problem fields such as machine translation, speech recognition, and others. Deep learning's LSTMs are a complicated topic.

LSTM functions similarly to an RNN cell at a high level. The LSTM network's internal workings are shown here. As indicated in the Figure 4 below, the LSTM is made up of three sections, each of which serves a distinct purpose.

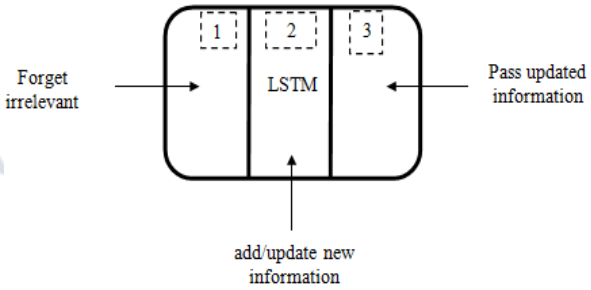


**Figure 4.** LSTM architecture

The first portion determines whether the information from the preceding timestamp is important to remember or can be ignored. The cell attempts to learn new information from the input in the second phase. Finally, the cell passes updated information from the current timestamp to the next timestamp in the third component.

The gates are the three elements of an LSTM cell. The Forget gate is the first section, the Input gate is the second, and the Output gate is the third as shown in Figure 5. The initial step in an LSTM network cell is to select whether to maintain or discard information from the preceding timestamp.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 02, February 2025

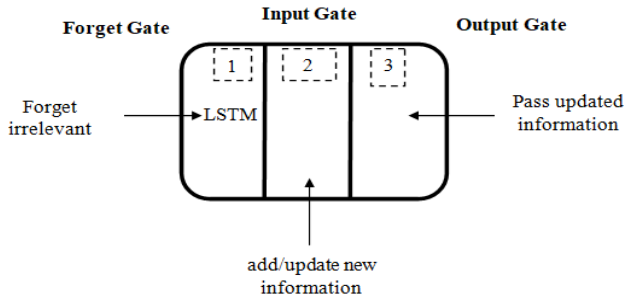


**Figure 5.** LSTM gates structure

Learning rates input and output biases and other parameters are all available with LSTMs. As a result, no fine modifications are required. With LSTMs, the complexity of updating each weight is decreased to O (1), similar to Back Propagation Through Time (BPTT), which is a benefit. To run at and for each sequence time-step, LSTM requires four linear layers (MLP layers) per cell. To be computed, linear layers demand a lot of memory bandwidth; in reality, they can't use a lot of compute units since the system doesn't have enough memory bandwidth to feed them.

### 2. Bidirectional Long Short-Term Memory

Bidirectional long-short term memory is the process of constructing a neural network that can store sequence information in both directions (future to past) and forward (ahead to future) (past to future). Our input runs in two directions in a bidirectional LSTM, which distinguishes it from a conventional LSTM. The input flow in one way, either backwards or forwards, with a normal LSTM. However, with bi-directional input, there information low in both directions, preserving both the future and the past.

The first model learns the input sequence, whereas the second model learns the opposite of that sequence. Given two models trained, must provide a mechanism to integrate them. It's commonly known as the Merge stage. One of the following functions can be merged sum, multiplication, concatenation and average. This short presentation emphasizes the fact that bidirectional LSTMs are effective time series forecasting models – in this case, utilizing the Bitstamp dataset for Bitcoin as input data for the network.

Bidirectional LSTMs are a type of LSTM that can be used to increase model performance on sequence classification issues. Bidirectional LSTMs train two instead of one LSTM on the input sequence in problems where all time steps of the input sequence are known. The input travels in both directions, and it can use information from both sides, unlike ordinary LSTM. It's also a useful tool for simulating the sequential relationships between words and phrases in both directions.

Bidirectional LSTMs vary from conventional LSTMs in that their input flows in both directions. The input can flow in one way, either backwards or forwards, with a normal LSTM. However, with bi-directional input, the information flows in both directions, preserving both the future and the past.

A bidirectional LSTM, often known as a biLSTM, is a sequence processing model that consists of two LSTMs, one of which takes input in one way and the other in the other. BiLSTMs effectively increase the quantity of data available to the network, giving the algorithm better context (e.g. knowing what words immediately follow and precede a word in a sentence) as shown in Figure 6.

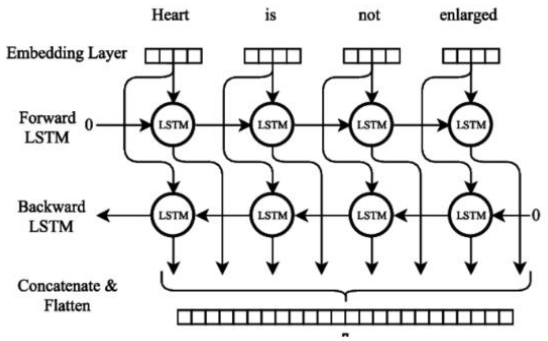


**Figure 6.** Bi-directional LSTM working

In contrast to forward and backward hidden sequences, a BiLSTM calculates the input sequence from the opposite direction. The encoded vector is created by concatenating the final forward and backward outputs, where the first hidden layer is's output sequence. Bidirectional LSTMs are a type of LSTM that can be used to increase model performance on sequence classification issues. Bidirectional LSTMs train two instead of one LSTM on the input sequence in problems where all time steps of the input sequence are known.

### 3. BERT

Bidirectional Encoder Representations from Transformers is a transformer-based machine learning technique for natural language processing pre-training developed by Google. BERT is an open source machine learning framework for natural language processing (NLP). BERT is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context. The application of Transformer's bidirectional training to language modeling is an important technical advance for BERT. Previously, researchers looked at a text sequence from left to right or a combination of left-to-right and right-to-left training.

Transformer, an attention mechanism that learns contextual relationships between words (or sub-words) in a text, is used by BERT. Transformer incorporates two different mechanisms in its basic form: an encoder that reads the text input and a decoder that generates a job prediction. Only the encoder technique is required because BERT's purpose is to construct a language model. The Transformer encoder reads the complete sequence of words at once, unlike directional models that read the text input sequentially

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 02, February 2025

(left-to-right or right-to-left). As a result, it is classified as bidirectional, however it is more correct to describe it as non-directional. This property enables the model to deduce the context of a word from its surrounds (left and right of the word).

BERT is unquestionably a watershed moment in the application of machine learning to natural language processing. Since, it is user-friendly and enables for quick fine-tuning, it'll probably find a wide range of practical uses in the future.

Language models could previously only interpret text input sequentially, from left to right or right to left, but not both at the same time. BERT is unique in that it can read in both directions at the same time. Bidirectionality is a capacity made possible with the development of Transformers. BERT is pre-trained on two different but related NLP tasks using this bidirectional capability: Next Sentence Prediction and Masked Language Modeling. The goal of Masked Language Model (MLM) training is to hide a word in a sentence and then the program guess which word was hidden (masked) based on the context of the hidden word. The goal of Next Sentence Prediction training is to have the software predict whether two provided sentences have a logical, sequential relationship or are just random.

The purpose of any NLP technique is to comprehend human language in its natural form. In BERT's example, this usually entails guessing a word from a list. Models are often trained utilizing a vast reservoir of specialized, labeled training data to do this. This demands linguists working in teams to manually label data. BERT, on the other hand, was trained with simply an unlabeled plain text corpus (namely the entirety of the English Wikipedia, and the Brown Corpus). Even when it is utilized in actual applications, it continues to learn unsupervised from unlabeled text and improve. Its pre-training acts as a foundation of "knowledge" on which to build. BERT can then be fine-tuned to a user's specifications and adapt to the ever-growing corpus of searchable content and queries. Transfer learning is the term for this procedure.

BERT, as previously stated, is made possible by Google's Transformers research. BERT's greater capacity for grasping context and ambiguity in language is due to the transformer, which is part of the model. Instead of processing each word individually, the transformer processes each word in connection to all other words in the phrase. The Transformer helps the BERT model to understand the whole context of a term by looking at all surrounding words, allowing it to better understand searcher intent.

BERT is also the first NLP techniques that rely purely on the self-attention mechanism, which is enabled by the bidirectional Transformers at its core. This is essential because a word's meaning can frequently alter as a phrase progresses. Each new word adds to the overall meaning of the word that the NLP algorithm is focusing on. The more words in a statement or phrase, the more unclear the emphasis word becomes. By reading bidirectionally, BERT provides for the augmented meaning by accounting for the effect of all other words in a phrase on the focus word and reducing the left-to-right momentum that biases words towards a given meaning as a sentence proceeds.

BERT receives a series of words as input, which continue to flow up the stack. Each layer performs self-attention, transfers the results through a feed-forward network, and then passes the information on to the next encoder. Until now, this has been similar to the Transformer in terms of architecture. The divergence of the output is clearly seen in the architecture as shown in Figure 7.

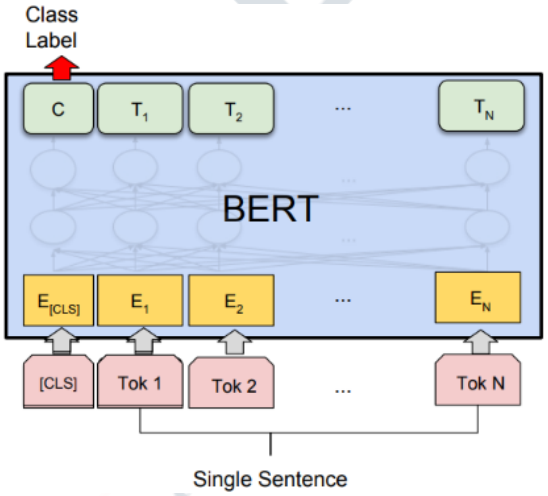


**Figure 7.** BERT architecture

Bert's model can merge mental traits with short text, according to model comparison. Bert can increase the accuracy of categorization results by better capturing mental aspects. This will contribute to the growth of brief text classification. BERT provides AI with a number of significant advantages, including: Model performance is significantly improved over traditional methods. It has the ability to deal with huge amounts of text and language. Using pre-trained models in a simple way (transfer learning). Because of the training framework and corpus, the model is quite huge. It takes a long time to train because it is so large and there are so many weights to update and it is not cheap. The nicest part is that pre-trained BERT models are free and open source. This means that anyone can use BERT to solve NLP tasks and develop models.

**4. Roberta**

Roberta stands for Robustly Optimized BERT Pre-training Approach. It was presented by researchers at Facebook and Washington University. The goal of this paper was to optimize the training of BERT architecture in order to take lesser time during pre-training. A robustly optimized method for pre training natural language processing (NLP) systems that improves on Bidirectional Encoder Representations from

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 02, February 2025

Transformers, or BERT, the self-supervised method released by Google in 2018.

Roberta employs dynamic masking, in which various parts of the sentences are veiled for different Epochs. This improves the model's stability. BERT is a ground-breaking technique that obtained cutting-edge results on a variety of NLP tasks by using a non annotated text from the web rather than a language corpus tagged particularly for a given task. Since then, the method has gained popularity as both an NLP research baseline and final task architecture. BERT also exemplifies the collaborative aspect of AI research, to undertake a replication study of BERT thanks to Google's open release, exposing chances to increase its performance. On the widely used NLP benchmark, General Language Understanding Evaluation, our optimized technique, Roberta, achieves state-of-the-art results (GLUE). The working architecture is shown in Figure 8.

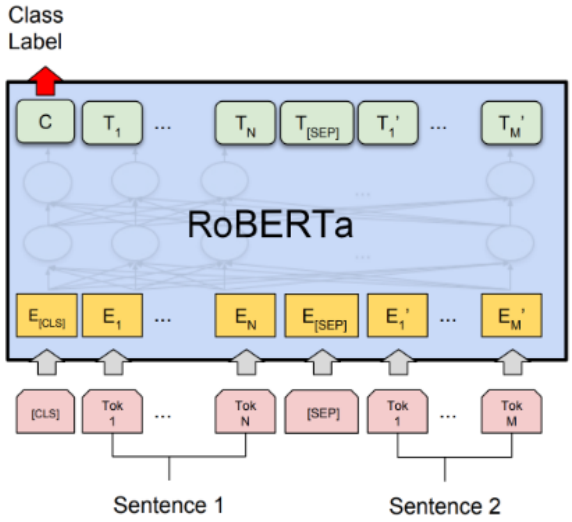


**Figure 8.** Roberta working principle

Roberta builds on BERT's language masking method, which teaches the system to predict purposely hidden content within otherwise the non annotated language instances. Roberta modifies critical hyper parameters in BERT, such as deleting BERT's next-sentence pre training target and training with considerably bigger mini-batches and learning rates, which was implemented in PyTorch. As a result, Roberta outperforms BERT on the masked language modeling objective, resulting in improved downstream task performance. Both existing non annotated NLP datasets, a new set of public news stories are employed.

The model achieved state-of-the-art performance on the tasks after adopting these design improvements, as well as a significant performance improvement on the GLUE benchmark. Roberta tied for first place in the GLUE leader board with an 88.5, matching the performance of the previous leader, XLNet-Large. These findings emphasize the importance of hitherto unstudied design choices in BERT training and help to separate the relative contributions of data size, training time, and pre training objectives.

### 5. Gated Recurrent Unit

Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) have been proposed (RNNs) to address the issue of vanishing / exploding gradients in typical Recurrent Neural Networks. LSTM and GRU are comparable, however GRU has fewer gates. There is no separate cell state because it relies only on a hidden state for memory transfer across recurrent units. A gated recurrent unit (GRU) is a type of recurrent neural network that uses connections between nodes to accomplish machine learning tasks such as memory and clustering in speech recognition.

Certain NLP tasks employing classic neural networks, such as text categorization and sentiment analysis, and we've completed them successfully. However, it ran into some issues with standard neural networks. In recurrent neural networks, gated recurrent units (GRUs) are a gating mechanism. The GRU is similar to an LSTM with a forget gate, but it has fewer parameters because it does not have an output gate. GRU's performance on polyphonic music modeling, speech signal modeling, and natural language processing tasks was found to be comparable to LSTM's. On some smaller and less frequent datasets, GRU has been proven to perform better. There are multiple variations on the full gated unit, including gating that uses the prior hidden state and bias in various combinations, as well as a simpler version known as the minimal gated unit. The working is shown in Figure 9.

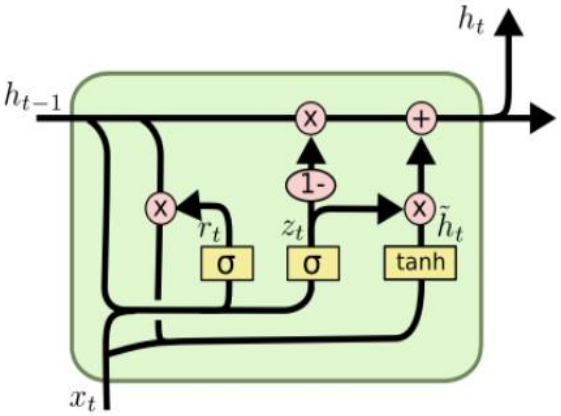


**Figure 9.** GRU working

Aside from its internal gating mechanisms, the GRU acts similarly to an RNN, with sequential incoming data and memory being consumed by the GRU cell at each time step, also known as the hidden state. The concealed state is subsequently transmitted back into the RNN cell along with the next set of input data. GRU is faster and consumes less memory than LSTM, although LSTM is more accurate when working with datasets with longer sequences.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 02, February 2025

## V. SYSTEM REQUIREMENTS

### A. Software Specifications

- Google Drive for dataset storage.
- Google Colaboratory.

### B. Software Description

#### 1. Scikit-learn

Scikit-analyze (Sklearn) is the most beneficial and strong library for machine learning in Python. It functions various type, regression and clustering algorithms which includes aid vector machines, random forests, gradient boosting, k-manner and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS fiscally backed challenge.

#### 2. Tensorflow

TensorFlow is an open-source library for numerical computation and large-scale machine learning. It is created by the Google brain team. TensorFlow bundles together a slew of machine learning and deep learning (aka neural networking) models and algorithms and makes them useful by way of a common metaphor. TensorFlow uses Python to provide a convenient front-end API for building applications with the framework, while executing those applications in high-performance C++.

### C. Functional Requirements

The consumer can be able to provide the dataset to the machine. The loaded dataset has no class category with sentiment bifurcation. The dataset is preprocessed and split as per its meaning with sentiments with the sentiment analysis algorithm techniques. The dataset is categorized and evaluation achieved using more than one classifiers.

### D. Non - Functional Requirements

1. Availability - The system can be used each time
2. Effectiveness – The balancing is finished the usage of sampling techniques through sklearn.
3. Throughput - Throughput of the system is notably adequate to offer a service.

## VI. EXPERIMENTAL ANALYSIS

### A. Environmental Setup

Google Collaboratory workspace is used to implement the Sentiment Analysis of COVID-19 variant Omicron from Twitter tweets.

### B. Dataset Description

*Dataset Collection* - The first and the foremost step is the dataset collection. The dataset has been downloaded from twitter using tweepy library from a particular date and with the number of tweets to be retrieved. Tweepy is a Python module for accessing the Twitter API that is open-sourced and simple to use. It provides a way for your Python application to interact with the API. The training set contains in 60,000 tweets. The test set includes 16000 tweets. The total number of classes that are distributed is 3.

*Data Preprocessing* - The dataset consists of sequential textual tweets. The tweets are preprocessed before training the model for sentiment analysis. The tweets downloaded contains junk characters, to extract the text without junk characters, data preprocessing is needed. Data preprocessing for this type of data required is cleaning the text. Text cleaning is done to remove emojis, punctuations, hashtags, special characters and multiple spaces using built-in libraries such as strip_emoji, strip_all_entities, clean_hashtags, filter_chars and remove_mult_spaces.

### C. Implementation Module

*Sentiment Analysis Labeling Module* - The sentiment analysis module is the module for splitting the retrieved dataset that is the tweets into three categories such as positive, negative and neutral. These classes are the sentiment analyzed from a particular tweet about the Omicron. TextBlob is used as an algorithm to categorize the tweets into the three different sentiments. The number of positive tweets is 42%, the number of negative tweets is 15.1% and the number of neutral tweets is 42.3%.

*Classification Module* - For training, the collected dataset along with the labeled sentiment is given as input and the system is trained from the input, the final values of the parameters used in this training is obtained after empirical attempts of combinations. The classifier models used for this module are LSTM, BiLSTM, BERT, Roberta and GRU. Once the classifier models are trained, the results are analyzed for predictions.

## VII. RESULT ANALYSIS

### A. Evaluation Metrics

#### 1. Accuracy

Accuracy is one metric which gives the bit of prognostications our model got right. Formally, accuracy has the following description,

$$Accuracy = \frac{\text{Number of correct prediction}}{\text{Total number of predictions}}$$

When performing classification prognosticate, there are four types of outcomes that might occur,

- True Positive (TP): When prognosticate an observationbelong to a class and it actually does belong to that class.
- True Negative (TN): When prognosticate an observation does not belong to a class and it actually does not belong to that class.
- False Positive (FP): When prognosticate an observation belong to a class and it actually does not

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 02, February 2025

belong to that class.

- False Negative(FN): When prognosticate an observation does not belong to a class and it actually does belong to that class.

**2. Precision**

As shown by the red rounded-rectangle in the confusionmatrix above, Precision is the ratio of the number of true positives to the total number of predicted positives. It is the fraction of predicted positives that were correctly classified. It measures how precise a model is whenit classifies an observation as being positive. It also tell us how well our model reduces type I error (False positive).

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

**3. Recall**

Recall is the ratio of the number of true positives to the total number of actual positives as shown by the blue rounded-rectangle in the confusion matrix above. It is the fraction of actual positives that were correctly classified. It measures how well a model recalls the actual positive classes. It also tells us how well our model reduces type II error (False negatives). It is also called sensitivity because it measures how sensitive a model is to the positive class.

$$Recall = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

**4. F1-score**

It is defined as the harmonic mean of the model's precision and recall. Harmonic mean is used because it is not sensitive to extremely large values, unlike simple averages. Consider a model with a precision of 1, and recall of 0 gives a simple average as 0.5 and an F1 score of 0. If one of the parameters is low, the second one no longer matters in the F1 score. The F1 score favors classifiers that have similar precision and recall. Thus, the F1 score is a better measure to use if you are seeking a balance between Precision and Recall.

$$\text{F1 Score} = \frac{\text{TP}}{\text{TP} + 1/2(\text{FN} + \text{FP})}$$

**B. Comparitive Analysis of Classification Models**

TextBlob algorithm is implemented on the given dataset for labeling before classification and prediction. The classification metrics reports like precision, F1-score, recall, support and accuracy. The classifier models such as LSTM, bi-directional LSTM, GRU, BERT and Roberta as discussed are implemented and their results are tabulated in the Table1.

**Table I:** Classifier performance

| Classifier | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| LSTM | 73.37 | 71.39 | 70.26 | 66.49 |
| BiLSTM | 80.24 | 76.11 | 76.77 | 72.86 |
| BERT | 85.65 | 83.16 | 84.95 | 84.07 |
| Roberta | 93.13 | 92.41 | 92.42 | 92.20 |

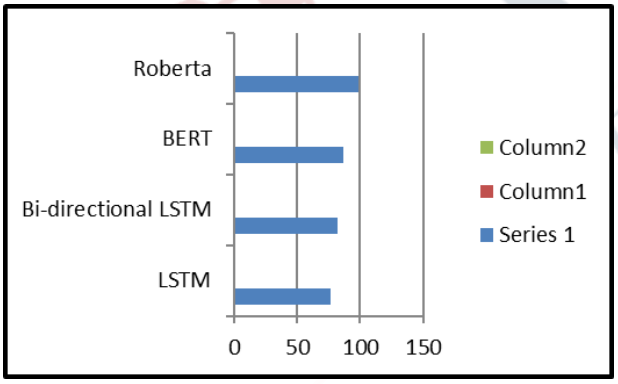


**Figure 10.** Comparative Study Accuracy Analysis

## VIII. CONCLUSION

Sentiment Analysis of COVID-19 variant Omicron from twitter using deep learning techniques provides prediction of sentiment and mentality of the public and its allied with the health units to provide more support. The predictive analysis of sentiment about a pandemic is the maintenance of secured society and increase the needs required and improves the quality of medical facilities needed as per the real data. The real world data received from twitter are not labeled based on the emotion and polarity and there is a need for categorizing using sentiment analysis techniques to provide the labeled sentiment for each tweet. TextBlob outperforms other sentiment analysis techniques because change in sentiment analysis methodology instances of each tweet. Roberta provides higher accuracy of 98% with a categorized sentiment analyzed data. The next nearer accuracy classifier is BERT.

The proposed work can be further enhanced for predictive maintenance of each tweet by analyzing the sentiment in each of it by choosing the contributing features that is the type of variant using supervised algorithms. The top picked tweets that contribute to the prediction of sentiment analysis about a disease explore the real emotion of the public. This provides scope for inducement of advanced technologies in medical field to provide more helps in terms of health and for the government to take effective actions as per the people's real views in pandemic situations.

## IX. ACKNOWLEDGMENT

The authors would wish to acknowledge the supports provided by Department of CSE of Kongu Engineering College for the completion of the work.

## REFERENCES

[1] Abraham and Sanders (2021). " Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse." AMIA Joint Summits on Translational Science: 555-564.

[2] Chinder Kaur (2020). " Twitter Sentiment Analysis on Coronavirus using Textblob. " Research Gate Publications.

[3] Imamah (2020). " Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion." Information Technology International Seminar.

[4] Swati Sharma and Aryaman (2020). " Twitter Sentiment Analysis During Unlock Period of COVID-19." Sixth International Conference on Parallel, Distributed and Grid Computing.

[5] Bishwo Prakash Pokharel (2020). "Twitter Sentiment Analysis during COVID-19 outbreak in Nepal." Social Science Research Network.

[6] Mrityunjay Singh (2021). " Sentiment analysis on the impact of coronavirus in social life using the BERT model." Socail Netwrok Analysis and Mining 11: 33.

[7] Supriya Raheja (2021). " Sentimental Analysis of Twitter Comments on Covid-19." 11th International Conference on Cloud Computing, Data Science and Engineering, IEEE.

[8] Prasoon Gupta (2021). " Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter." IEEE Transactions on Computational Social Systems, Vol:8, No:4.

[9] Mohammed Ehsan Basiri (2021). " A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets" Science Direct Knowledge Bases Systems.

[10] Shivam Behl (2021). "Twitter for Disaster Relief Through Sentiment Analysis for COVID-19 and Natural Hazard Crisis." Internaltional Journal of Disaster Risk Prediction 55: 02.