An Ensemble Deep Learning Model for Mental Depressive Disorder Classification and Suicidal Ideation through Tweets

[1*]Vani Rajasekar, [2]S. Kanimozhi, [3]S. Sankarananth, [4]R. Sharan, [5]R. Nitiish
[1*,4,5] Department of CSE, Kongu Engineering College, Erode, Tamil Nadu, India
[1] Email: vanikecit@gmail.com
[4] Email: sharanr.21cse@kongu.edu
[5] Email: Nitiish.21cse@kongu.edu
[2]Assistant Professor, Department of Biomedical Engineering,
Sri Shanmugha College of Engineering and Technology, Sankari, Tamil Nadu, India
Email: dayakanimozhi@gmail.com
[3]Assistant professor, Department of Electronics And Communication Engineering,
Bannari Amman Institute of Technology, Erode, Tamil Nadu, India,
Email: sankarananth@bitsathy.ac.in

Abstract:

Early detection of mental depression prevents the severe repercussions of long-term depressive symptoms such as suicidal thoughts and ideation. With the widespread use of social media and the internet these days, prompt identification of emotional reactions is essential. Therefore, monitoring social media texts like Facebook comments or tweets could be highly helpful in detecting the mental depression. With the advent of Artificial Intelligence (AI) techniques it is possible for the early detection and classification of mental depressive detection and suicidal ideation. The proposed approach uses the labeled twitter tweets to classify the depression intensity. The performance evaluation is done based on the four ensemble models known as CNN, LSTM, LSTM+ RNN and BERT for the classification of tweets based on depressive and non-depressive classes. The parameters used for the evaluation are accuracy, precision, recall and F1-Score. From the result analysis it is inferred that average high accuracy and precision is obtained as 97% for LSTM. Similarly average high recall and F1-Score is obtained as 96% and 97% respectively. Furthermore, the optimization helps to enhance the proposed classification as well as it makes suitable for identification of suicidal ideation. The suggested method thus achieves the better performance for the earlier identification mental depression based on emotions in many social media users, demonstrating the viability of CNN, BERT, RNN and LSTM.

Key words: Mental depression, Suicidal ideation, Deep learning, CNN, LSTM, BERT.

1) Introduction

The startling global prevalence of depressive disorders is highlighted by recent statistics. According to the World Health Organization (WHO), depression is one of the main factors contributing to disability, affecting over 264 million people worldwide based on the most recent data. Global concerns about mental health are particularly acute in the majority of developed nations and many emerging economies. WHO report on mental health states that 1 in 4 persons globally experience some form of mental illness. Based on data statistics from the Global Health Metrics, depression is one of the psychiatric conditions that contribute significantly to the global disability cause. This might influence a person's performance at job, in school, or even in their relationships with family and friends. Approximately 4.4% of the global population, or more than 350 million people, suffer from depression. Furthermore, two thirds of patients among them don't ask for assistance. The main issue is that depression can have an unintentional negative impact on a person's social and personal life. If the treatment for severe mental illnesses is not provided at the earlier stage, the issue will get worse. Existing research indicates that the mental problems are typically associated with an increased risk of suicide. Individuals with depressive illnesses are primarily characterized by anhedonia, or trouble finding pleasure, and a state that leans towards depression. The following requirements are looked at if these traits manifest [1]. There may be some physical complaints, like difficulties sleeping, sharp drops in appetite and weight, blurred vision, mental exhaustion, and a poorer reaction to emotions such as fatigue, worry, or irritation. Non-somatic disorders include feelings of helplessness or sadness, pleasure loss, worthlessness, guilt, or suicidal thoughts, which can also affect some people. Furthermore, the COVID-19 pandemic has exacerbated the situation by creating a spike in the number of documented cases of anxiety and depression as a result of the difficulties it has caused with regard to social, economic, and health issues. however, many victims were unable to receive appropriate treatment, despite the fact that some of them were suffering from serious mental illnesses, as a result of the lack and inequality of public resources in health services [2,3]. People began posting emotional posts in forums and looking for online assistance as social network services developed. An estimated 900, Suicide is the second leading cause of death for them worldwide, accounting for one in every ten deaths of young people. A portion of these deaths are caused by serious mental illnesses. A study using linguistic and interactional measures examined the transition from mental health to thoughts of suicide in individuals who have attempted suicide.

---

[1]Tankut, Ü., et al., Analysis of tweets regarding psychological disorders before and during the COVID-19 pandemic.

[2] Pavlova, A., & Barkers, P. (2022). “Mental health” as defined by Twitter: Frames, emotions, stigma. Health communication, 37(5), 637-647.

[3] Cedeno-Moreno, D., et al., Automatic Classification of Tweets Identifying Mental Health Conditions in Central American Population in a Pandemic.

There are reports of mental health issues among those who attempt suicide [4]. The transition from mental diseases to suicidal thoughts and actions is a gradual process. Suicide risks were categorized into four categories non-suicidal, suicidal ideas or desires, suicidal acts or plans, and suicidal intentions. Prior to experiencing suicidal thoughts, individuals may have several other mental health issues. Meta-analyses show that underlying mental problems had a 90% chance of contributing to suicide, mostly in high-income nations. One of the most useful resources for giving people with mental health concerns assistance and feedback is the social networking site. With the limited resources for help, effective early prevention of suicide requires automatically assessing various levels of risk and providing ethical support to alleviate victims' concerns. The goal of the proposed approach is to apply deep learning approaches for earlier identification and determine individuals' risk levels. By doing so, specialists or social workers will be better equipped to grasp the circumstances of the people they are assisting with their mental health problems [5]. Online support can be facilitated and mental health monitoring implemented with the help of the automatic detection technology. Social workers may find it easier to prioritize and distribute resources to individuals with varying needs and conditions based on the severity of those situations with the use of suicide risk category and mental health rating. Therefore, it is possible to block the transition from mental health discourse to suicidal ideas by effective preventive efforts. When considering a multi-class classification problem, mental health disorders and suicide could be classified at multiple levels.

Research on affective computing is used to attribute feelings to suicide notes, the majority of which have a lot of depressing language. Those blogs frequently touch on topics including personal crises, family problems, and work stress. Suicidal thoughts and other mental health disorders must therefore be classified with consideration for the subtle variances between each of those criteria. Contextualized pre-trained language models or transformers, are a recent trend in NLP that have drawn a lot of interest for a range of text processing applications [6]. Transformer provides a range of modelling languages to enable various word representations and has demonstrated its capacity to comprehend global contexts. While previous investigations have demonstrated that LSTM models can classify mental health conditions Text classification has undergone a revolution to recent efforts that use deep neural networks. Categorizing mental health and suicidal ideation, on the other hand, is a more specialized process that necessitates paying attention to the language used by potential victims. In online social material, there are similarities between suicidal thoughts and mental diseases such as depression, anxiety, and bipolar disorder in terms of language usage, topic distribution, and emotion polarity.

---

[4] Ji, S., Li, X., et al., Suicidal ideation and mental disorder detection with attentive relation networks. Neural Computing and Applications.

[5] Rajasekar, V., Efficient multimodal biometric recognition for secure authentication based on deep learning approach.

fairly well, domain-specific pre-trained language models, which incorporate new advances in contextual word representations such as Mental BERT, may increase efficiency and accuracy. The LSTM, BERT classification model, and models that combine the LSTM and Transformer word representation models are some of the different deep learning classification models that will be looked at in this research. For the classification of mental depressive disorders, combining Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) can be a useful strategy. For every text token, create contextualized word embeddings using BERT. Word semantic meaning is captured by BERT embeddings according to the sentence's context.

2) Literature Review

In the subject of health informatics, social media, and particularly social networks, have grown to be extremely significant information sources. A component of health informatics is the design and use of information technology-based innovations to address issues with health care and public health. The various approaches of machine learning and deep learning to promote advancements linked to psychiatric issues has been the subject of extensive investigation. [7] tested a number of models, including LSTM and Bi-LSTM models that combine LSTM and Bi-LSTM with an attention layer. The model was able to select educational terms since attention mechanisms were added in the research that was done. Applying the Bi-LSTM model carefully yielded the better accuracy rate of about 74.1% and F1-score of about 74.3%.

Deep Learning is considered as important topic to analyze EEG information, because it can be taught over the massive quantity of data created by EEG. It also incorporates automated FE by utilizing raw EEG signals, increasing results. 46 carefully chosen primary studies having been employed in extensive mapping studies [8]. The author aimed to provide a comprehensive overview of the key areas of research in the diagnosis and treatment of mental illnesses utilizing EEG and DL, along with recommendations for more study. [9] provides a useful model text-based depression prediction that combines Recurrent Neural Network (RNN) with two dense layers and two hidden layers. This model can be used to prevent mental disorders that leads to the suicidal ideations.

---

[6] Subramanian, M., et al., Effectiveness of decentralized federated learning algorithms in healthcare: a case study on cancer classification

[7] Anindyaputri, N. A., et al., A Comparative Study of Deep Learning Models for Detecting Depressive Disorder in Tweets. [8] Rivera, M. J., et al., Diagnosis and prognosis of mental disorders by means of EEG and deep learning: a systematic mapping study.

[9] Amanat, A., et al., Deep learning for depression detection from textual data. Electronics, 11(5), 676.

Using text, semantics, and content from writing, researchers train an RNN to detect depression in textual data. 33 publications on the medical classification of schizophrenia, anxiety,

depression, bipolar disorder, PTSD, an eating disorder nervousness, as well as attention deficits hyperactivity disorder (ADHD) are taken from multiple database searches with the most commonly used managing criteria for meta-analyses and systematic reviews technique in [10].

These papers were selected through an individual assessment of their application of deep learning and machine learning technology, and their suggested approaches were subsequently categorized into the different illnesses covered by this study. Additionally, a list of numerous publicly accessible datasets is provided along with a discussion of the challenges the researchers faced. Supervised learning is the process that involves acquiring the relationship between a set of input factors as well as an outcome factor, using it to forecast the results of never-before-seen data. A useful example of supervised learning for solving regression and classification issues is the support vector machine (SVM). This approach divides n-dimensional space into distinct classes by locating the hyperplane, or ideal decision line or border [11]. It is based on the idea of margin calculation. This entails classifying fresh data items going forward into the appropriate groups. SVM has several benefits, such as its capacity to handle structured and semi-structured data. Ensemble learning is a popular machine learning technique. It entails teaching multiple students how to tackle an issue on their own. With this approach, a single model is created by merging several learners, each of which functions as a separate standard machine learning method. Three classes make up ensemble learning: bagging, boosting, and stacking. By employing random sampling to construct several datasets, bagging simultaneously builds multiple learners and then aggregates all of the students employing a majority vote or average method [12],[13]. The author identified neurophysiological markers of depression in two of the biggest resting-state datasets for MDD and distinguished MDD patients from healthy subjects using advanced deep learning techniques in addition to conventional machine learning. Graph convolutional neural networks (GCN) and support vector machines (SVM) were used to classify functional connectivity matrices, and 5-fold cross-validation was used to assess outcome [14],[15] suggested an end-to-end integrated DL model based on resting-state electroencephalography (EEG) data for the purpose of categorizing MDD patients and healthy controls. A fully connected layer was then used to complete the classification. Initially, the model used a multi-head self-attention algorithm to autonomously acquire these relevant connectivity relationships.

[10] Iyortsuun, N. K., et al., A review of machine learning and deep learning approaches on mental health diagnosis

[11] Squires, M., Tao, X., et al., Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment.

[12] Rajasekar, V., et al., Ensemble machine learning methods to predict the balancing of ayurvedic constituents in the human body.

[13] Bondi, E., Maggioni, E., et al., Neuroscience & Biobehavioral Reviews, 144, 104972.

[14] Gallo, S., et al., Functional connectivity signatures of major depressive disorder: machine learning analysis of two multicenter neuroimaging studies.

[15] Xia, M., Zhang, Y., et al., An end-to-end deep learning model for EEG- based major depressive disorder classification, IEEE Access.

[16] suggested a deep learning methodology that uses EEG data to produce images. From EEGs, two pictures are created using functional and spectral connectivity properties. The images produced are then used to train a two-stream convolutional neural network, and the results are then mixed. After that, the combined result is put into effect to a sequential model of long-short-term memory, completely connected, and softmax layers, which determines which samples belong to the MDD and healthy control (HC) classes. A public EEG dataset comprising EEG information from 34 MDD patients and 30 HC-matched subjects was utilized to validate the suggested methodology. The electroencephalogram (EEG) is a commonly used technique for real-time recording of the electrical brain dynamics brought on by neurological activity.

This allows for the analysis of mental processes, brain health, or dysfunction, and potential signs of mental diseases. Several sensors, or electrodes, are often positioned on various parts of the scalp in an EEG. These types of sensors capture the voltage differential among two separate electrodes, which detect neuronal activity in various brain regions [17],[18]. It has demonstrated efficacy in the diagnosis of illnesses related to the nervous system, cognitive psychology, and psychophysiology. In terms of the systematic literature review (SLR) strategy for depression identification, [19] reviewed a number of cutting-edge machine learning and deep learning algorithms. We also mention a few important issues from the body of research that could be investigated further in the future. Lastly, they hope that this poll will aid readers and ML and DL researchers in understanding important approaches to depression diagnosis. [20] suggested a three-part approach: There are three types of CNN models: one that is trained only on textual features, one that is trained only on audio characteristics, and a hybrid model that employs LSTM techniques to integrate textual and audio features. An enhanced variant of the LSTM model called the Bi-LSTM model is also employed in the suggested methodology. The models that are discussed in the results have their training accuracy, training loss, validation accuracy, and validation loss determined.

[16] Afzali, A., Khaleghi, A., et al., Automated major depressive disorder diagnosis using a dual-input deep learning model and image generation from EEG signal.

[17] Rafiei, Aet al., Automated detection of major depressive disorder with EEG signals: a time series classification using deep learning.

[18] Venu, K., et al., EEG Signal Classification for Motor Imagery Tasks.

[19] Hasib, K. M., et al., Depression detection from social networks data based on machine learning and deep learning techniques.

[20] Marriwala, N., & Chaudhary, D. (2023). A hybrid model for depression detection using deep learning. Measurement: Sensors, *25*, 100587.

3) Proposed Methodology

Finding people who have suicidal thoughts is the goal of the work of categorizing posts about suicide. As a result, it's common to find suicidal thoughts in social media posts. For this role, various machine learning and deep learning approaches are utilized in the earlier researches. Using suicidality keywords taken from earlier research, we first gathered Twitter data in order to detect suicidal attitude on social media. In order to identify suicidal content, we first eliminated duplicate tweets, then annotated and labeled the data, performed feature extraction and preprocessing, then used machine learning and deep learning models trained on various feature representation sets. Lastly, we examined how well the employed strategies performed. The ensemble models used in the proposed approach are LSTM, RNN+LSTM, CNN and BERT model. The proposed methodology contains the following modules a) Data collection b) Data Pre-processing c) Feature extraction d) Ensemble model classification d) Performance evaluation. The flow of the methodology is shown in the Figure 1. Creating a framework and uploading a Twitter database are the first steps toward predicting depression. Experts in mental health text analysis categorize tweets as depressed or not depressed in order to detect sentiments associated with depression or not. The supplied data must first undergo pre-processing steps in order to remove noise. Following the parameters when processing the data has a major positive effect about the standard of extraction of features. Pre-processing approaches including data normalization, encoding, removal of punctuation, stop word removal, and others are processing in the supplied textual information. Clean and noise-free information is produced in this process, which is then utilized for feature extraction. The pre-processed data is subjected to feature extraction processes in order to extract pertinent and significant features. The features that are extracted determines the pertinent data dimensions to let classification algorithms function more effectively.

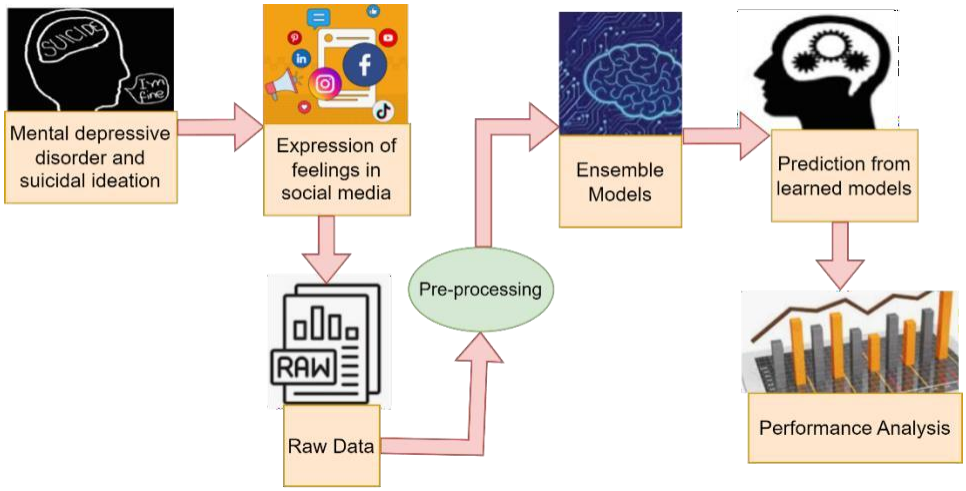


Fig. 1. Architecture of the proposed model

3.1) Data Preprocessing

Preparing the context is an important step in the data mining process. Real-world data is gathered in a variety of ways and isn't domain-specific, which leads to incomplete, unstructured, and untrustworthy data that contains errors. If such data is immediately evaluated, it produces inaccurate and inappropriate forecasts. Several techniques are employed in our framework's pre-processing stage. The user-specified text patterns are removed using the first approach. This function is to be used to eradicate the patterns like "user handles (@username)", "hashtags (#hashtag)", "web addresses", "characters, representations, and numerals", "empty strings", "drop rows with NaN in the column", "duplicate rows", and so on. Using this procedure, all URLs in each tweet are removed and the dataset is cleaned up. URLs are not considered account since removing them will decrease and they are not helpful for forecasting computational intricacy. Time and period are ineffective in predicting depression; hence this data is eliminated from the tweets. Following this, stop words are removed and stemming is carried out. One possible approach to return an information in its base procedure is through stemming. Porter Using a stemmer, one can create a word's root by eliminating any prefix or suffix from the term (−ize, −ed, −s, −de, etc.). Once every tweet has been cleaned, the cleaned versions are sent back and used as input for the tokenizer, which is the next stage. Programs called tokenizers divide a text file into fewer lines or words, then use normal expressions to split the input string into tokens. Tokenization starts with providing the cleaned positive and negative tweet datasets to the tokenizer.

3.2) Feature Extraction

The unique features are extracted from the database using Principal Component Analysis (PCA). Rather of using the standard Term Frequency–Inverse Document Frequency (TF-IDF) approach, PCA was utilized to identify and normalize the frequencies of all significant depressed terms. PCA uses the covariance matrix to perform decomposition and provide eigenvalues that optimize inter-class scattering while minimizing sample inner scattering. The inner class scattering matrices are shown in equation 1 and inter class matrices are represented by equation 2

$$M_i = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x_i})\,(y_i - \bar{y_i})^T \qquad (1)$$

$$M_{it} = \frac{1}{n-1}\sum_{xi}\sum_{i=1}^{n}(x_i - \bar{x_i})\,(y_i - \bar{y_i})^T \qquad (2)$$

Where M_i represents the inner class matrices and M_{it} represents the inter class matrices. x_i represents all the feature vector and $\bar{x_i}$ represents the mean of all the feature vectors. The optimized feature vector is shown in equation 3.

$$S_{opt} = \frac{|S^T M_i S|}{|S^T M_{it} S|} \quad (3)$$

3.3) Ensemble Model for Depression Classification

The convolution is the representation form the sliding W-grams for input sequences consists of 'n' entries such as a_1, a_2, an. The convolution weight and bias are included for generation of w-grams.

The one-dimensional convolution operation of filter size 250, kernel size 3 and filter length is denoted by 'l'. The global max function is denoted by equation 4.

$$w_i = conv1d(a_i) \quad (4)$$

In order to anticipate distressed persons on Twitter with better classification performance, a method that combines CNN with bidirectional-LSTM is proposed which is a kind of RNN. After a lot of testing, we found that CNN performs better in contextual knowledge from the preceding information is not necessary and is efficient at retrieving spatial features. In contrast, RNNs are useful for information extraction when the surrounding elements' context is crucial for categorization. Sentences that follow one another in textual material can be used to convey feelings when speaking with others. It is possible to encode time-sequential data using machine-learning methods. Recurrent links between hidden states and states in the past and present make up an RNN. Gradient vanishing issues can occasionally affect memory, which is a crucial component of neural networks. Problems with memory processing have an answer in LSTM technology. A RNN with 60 LSTM units is depicted in Figure 2 as its fundamental construction.

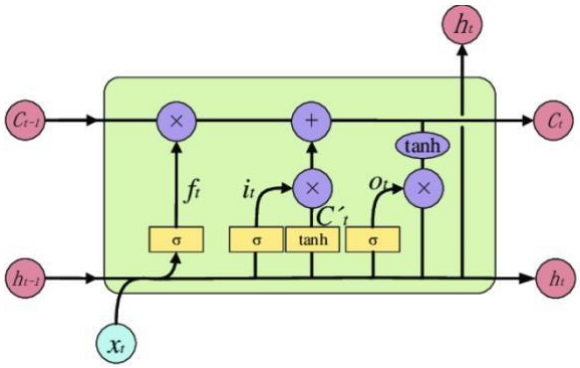


Fig. 2. Basic Architecture of LSTM

Each and every LSTM unit contains three types of gates a) Input gate b) Forget gate and c) Result gate. The attention layer on the LSTM is denoted in the equation 5.

$$A(attention)_t = LSTM(h_1, A(attention)_{t-1}) \quad (5)$$

Specifically, the reasoning behind using the attention-based technique on the layer above of LSTM is employed to increase the accuracy of the model. The attention strategy prioritizes

important information above unimportant information in the current goal. The "SoftMax" feature might be utilized to ascertain the output outcomes.

$$Result = softmax(UA_t + b_t) \quad (6)$$

Where U is the weight function, A_t is the attention mechanism at the timestamp t and b_t is the bias at the timestamp t.

Transformers serve as the foundation for the pre-trained, unsupervised Natural Language Processing model known as BERT (Bidirectional Encoder Representations from Transformers). The process of continuously determining the weightings between each output element and each input component of a transformer, based on their relationship, is referred to as attention in natural language processing. It is bidirectional because it makes use of a transformer design, which enables it to take into account a word's left and right context at the same time. It is possible to record a word's complete context within a sentence or textual sequence thanks to the bidirectional functionality. The subtleties of natural language, such as context, semantics, and sentiment communicated in text, can be understood and captured using BERT. Because of its bidirectional characteristic, it has a conceptual grasp of catching both explicit and implicit signals associated to depression. Additionally, it possesses a semantic understanding that allows it to recognize the commonalities among the many depression manifestations. It operates on the transformer architecture, a deep learning model that Vaswani et al. presented. This model is made up of several layers of encoders, one feed-forward neural network and one self-attention mechanism per layer. In the pre-training stage, BERT refines its capacity to predict absent words in a sentence by examining the surrounding context. This is made possible through the method named as Masked Language Modeling (MLM), in which some input keywords are purposefully hidden. The proposed model is then taught to foresee the concealed tokens by using the context that remains. This model makes use of Next Sentence Prediction (NSP), an extra pre-training assignment. When two sentences are entered into NSP, the model is trained to determine if the first and second sentences in the original text make sense together. This exercise helps BERT understand the relationship between sentences, which improves its capacity to handle problems that need contextual comprehension that goes beyond individual sentences. The flow of the BERT model is shown in the Figure 3.

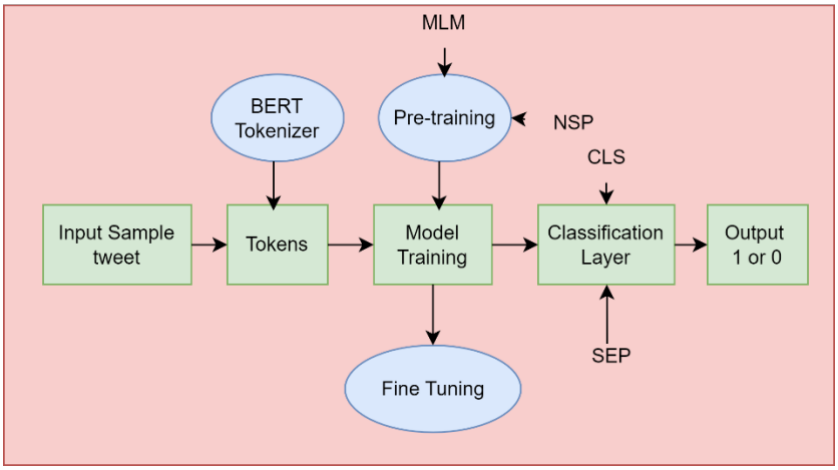


Fig. 3. BERT Architecture

Similar to the transformer paradigm, BERT has a multilayer bidirectional transformer encoder architecture. The transformer architecture is defined by the encoder-decoder system that is based on self-attention on the encoder's part along with focus on the decoder part. In BERT, there are two variants:

LARGE variety has 24 layers in the ENCODER stack, whereas the BASE variant has 12. This is superior than the six encoder layers of the original transformer system. Furthermore, BERTLARGE amplifies this with 1024 hidden units, surpassing the 512 hidden units of the original study, and BERTBASE provides larger feedforward networks with 768 hidden units. Furthermore, the original paper suggested 8 attention heads; both BERT variations had 16 and 12, respectively, more attention heads. With 340 million parameters, BERTLARGE has more parameters than BERTBASE, which has 110 million.

Three crucial keys are included in the map that the BERT models return: pooled output, sequence output, and encoder outputs.

- **pooled output**: Each series of input is represented as a whole by pooled_output. It is [batch size, H] in shape. Consider this to be an embedding of the whole review of the film.
- **sequence output**: Every input token in the context is represented by sequence output. [batch size, seq_length, H] is the shape. Consider this to be a embedded context for every character in the movie review.
- **encoder output**: The L Transformer blocks' intermediate activations are called encoder outputs. "Encoder outputs" in outputs for every 0 <= i < L, [i] is a Tensor of shape [batch size, seq_length, 1024] containing the outputs of the i-th Transformer block. Sequence output is the value of the list at the end.

3.4) Fine tuning using BERT

Bert can be used as a feature extractor by aggregating or combining the output of the training algorithm's last layer, depending on how many layers it has and a meaningful sentence representation can be created. BERT was designed to anticipate words and sentences, so it is essential to fine-tune it for a particular task. Evaluation of sentiment with BERT can be carried out by superimposing a layer of categorization applied to the [CLS] token's Transformer result. The last concealed layer of this token functions as the "sentence vector" for series categorization. In the event that the algorithm is optimized, the [CLS] token representations becomes an appropriate sentence representation.

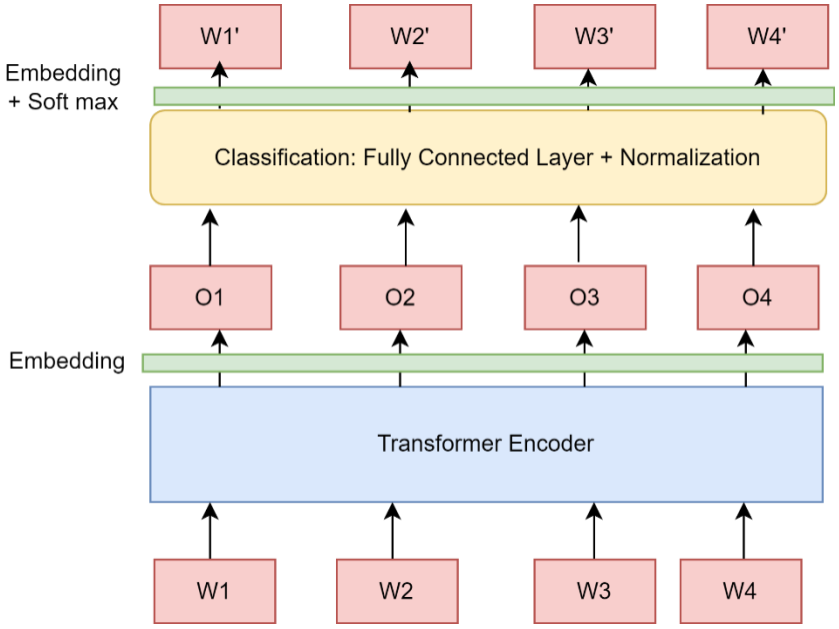


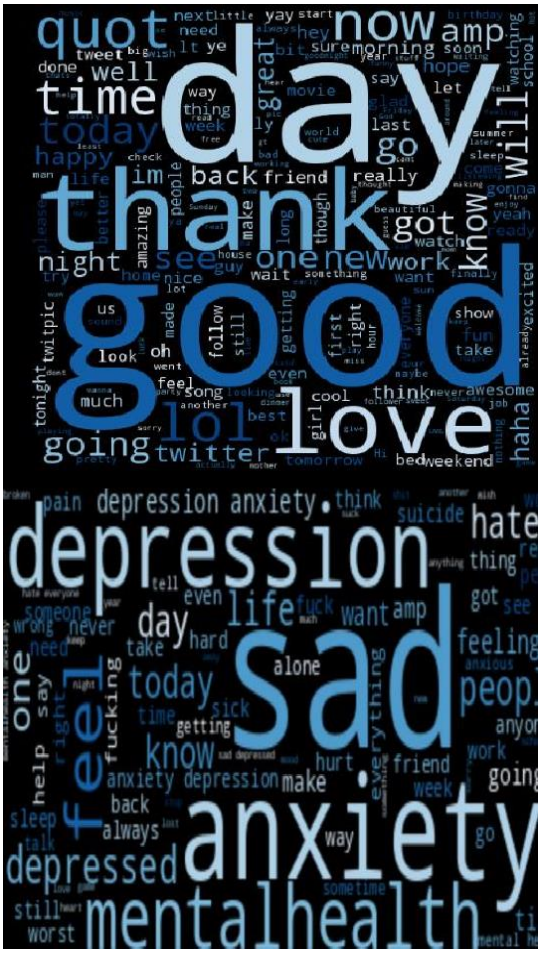
Fig. 4. BERT Fine Tuning Model

- **Pre-trained Model**: bert-base-uncased (12-layer, 768-hidden, 12-heads, 110M parameters)
- **Processor:** Custom or Cola Processor with Label list: [“Neutral”, “Negative”, “Positive”]
- **Fine-tuning parameters:** Maximum Sequence Length = 30, Batch Size = 32 Number of train epochs = 4.0, warmup fraction = 0.1, learning rate = 2e-5, 5e-5, and 8e-5 (During the warm-up stage, the development rate rises linearly). In the first batch repetition, the development rate is 1*p/n if the desired training rate is p and the time spent warming up is n. In the second batch, the training rate is 2*p/n, and so on, until we reach the actual rate at iteration n.

- The steps below can be used to perform a similar analysis on the self-attention layers in BERT Sentiment Analysis:
    1. Assign the attention weights for the [CLS] token to the final multi-head attention layer.
    2. Calculate each token's average over several heads.
    3. Consistency among tokens

3.5) Performance Evaluation

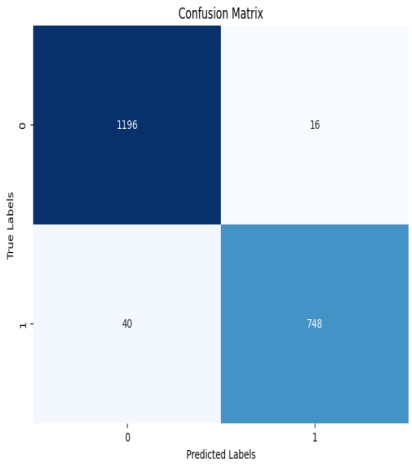
To guarantee a uniform distribution of each class throughout all datasets, in that the tweets that are labeled are divided into training dataset and test dataset in this stage. The Python Sklearn package is utilized for splitting and the stratified division of data is supported by the 'train_test_split' function. The models employed in this study are CNN, LSTM, LSTM+RNN and BERT models for the categorization of tweets based on depressive and non-depressive classes. The sample dataset images and class label encoding are shown in the Figure 5.



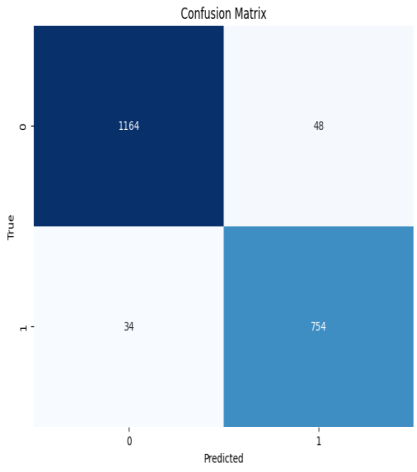
	text	label
8740	rt david cooks idol dies ais season seven winn...	0
54577	happy everything place home long process thoug...	0
1742	mental health taken serious something tragic h...	1
49258	minutes later light crashed floor broke pieces	0
34475	yeah totally would nice see get point	0
...	...	...
39571	wish manlier could grow beard reasonable speed	0
3201	made foreign friend ghosting sad	1
226	dam bored lost followers depressed	1
31505	hurt ankle really bad playing soccer haha laug...	0
20413	oh watch videos people dying right going sleep	0

Fig.5. Tweet Dataset and Label Encoding

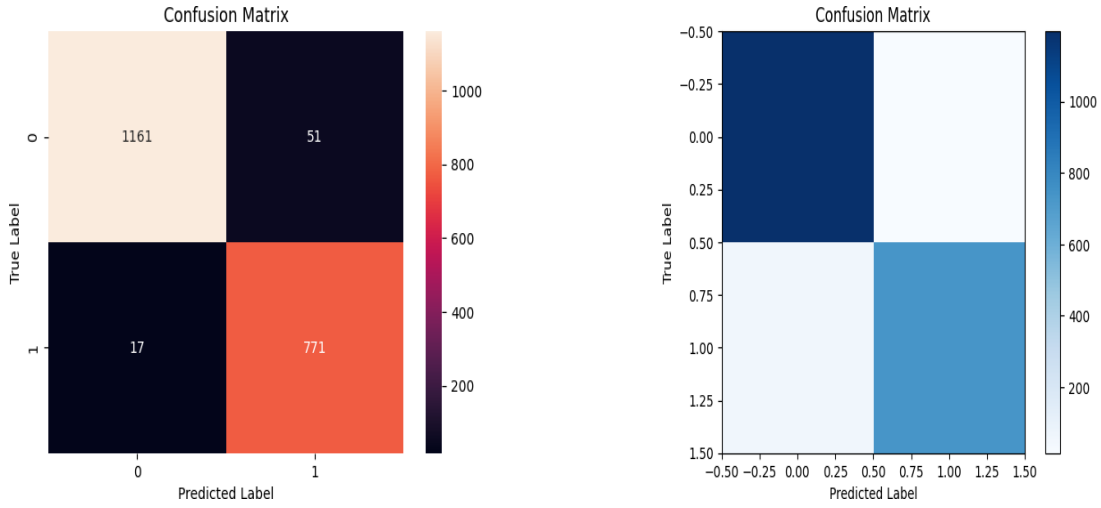
'Hugging Face' offers tokenizers for each model, which translate tokens to corresponding IDs. These tokenizers are used to tokenize tweets. The longest possible token length in the dataset is found to be 62 characters, hence all tweets have been extended to an established maximum of 64 characters which is split into three parts: 15% goes toward the test set, which has 11000 tweets, 70% goes toward the training set, which has 51000 tweets, and 15% goes toward the validation set, which has 11000 tweets. The layer used for classification includes a dropout layer that represents the degree of depressive disorders, with a softmax of size three. To prevent models from being overfitted too soon during the training phase, a dropout layer is added. Electra has received pre-training from Bookcorpus and Wikipedia. The pre-trained model is fed information for training to train the classification layer, which has all configurable parameters, on a particular depression intensity categorization. Here's how to fine-tune using the hyperparameters. Using instructional rates of three different values—2e-5, 5e-5, and 8e-5—the models' ability is evaluated at low, average, and high learning rates. For every analysis, the optimizer named "Adam" is utilized. For each experiment, the batch size is 64. Model training is done using the TensorFlow and Keras deep learning framework. Effectiveness a single cycle of learning policy, which simplifies training to get the optimal learning rate, is employed for effective training. The learning rate progressively rises in the first half of training and progressively falls in the second. All of the trials are conducted on an Ubuntu-based computer with an Nvidia Tesla P100 GPU and 12 GB of RAM.



a) LSTM



b) BERT



c) RNN + LSTM d) CNN

Fig. 6. Confusion Matrix of the Ensemble Models

3.6) Evaluation Metrics

The trained model is applied to the test dataset using the soft max layer to predict class labels. For every experiment, a three-dimensional confusion matrix between the true and anticipated labels is generated. The confusion matrix indicates misclassification in addition to providing accuracy for each class. All models are evaluated using scores from the associated confusion matrices and evaluation measures, including accuracy, precision, recall, F1, and specificity. The metrics used for evaluation are based on TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). The confusion matrix is shown in the figure as follows

When it comes to accurately classified instances, each approach may perform differently. Accuracy is computed by dividing total amount of correctly predicted classes by the total amount of samples given in the model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6)$$

Overall, the outcomes were predicted through precision. This is defined as the total of individuals who are diagnosed with depression and who are actually expected to be depressed; it may be calculated using the equation shown in Equation 7.

$$Precision = \frac{TP}{TP+FP} \qquad (7)$$

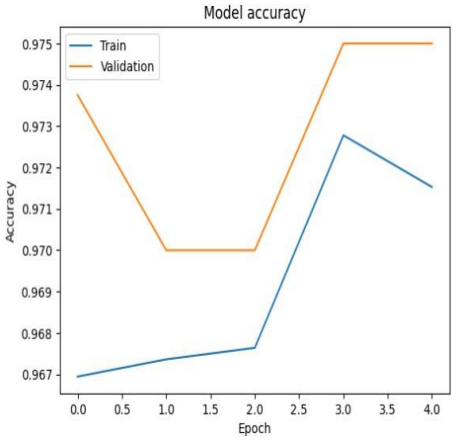
Out of all the positive instances, recall indicates how many positive depression ases the model correctly predicted. Recall is calculated as denoted in the

equation

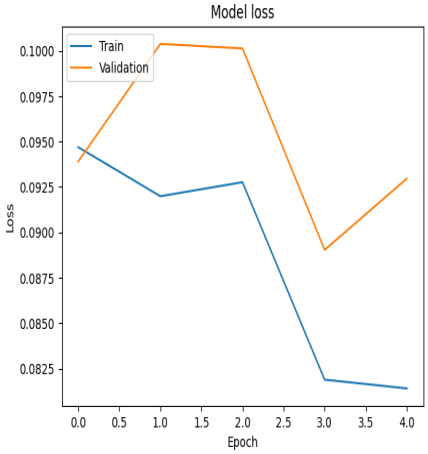
$$Recall = \frac{TP}{TP+FN} \qquad (8)$$

The F-Measure calculates different qualities indicated in Equation 9, its value assesses the harmonization of two components based on precision and recall.

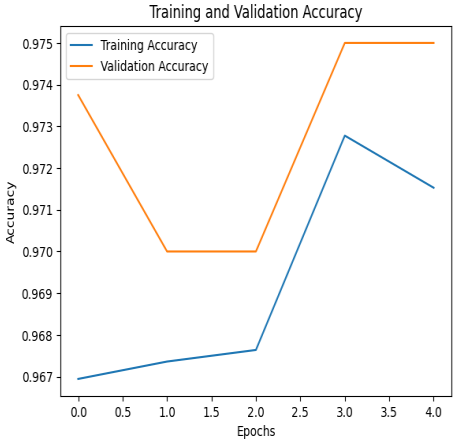
$$F1_{Score} = 2 * \frac{Precision*Recall}{Precision+Recall} \qquad (9)$$



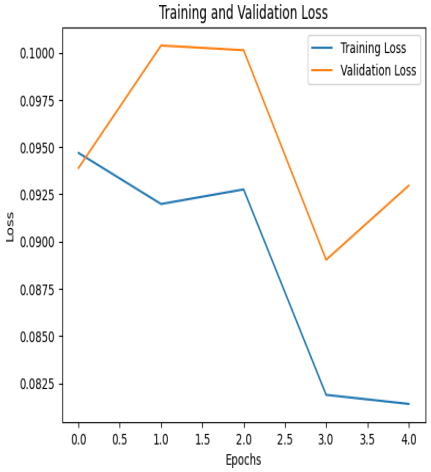
a) Accuracy of BERT



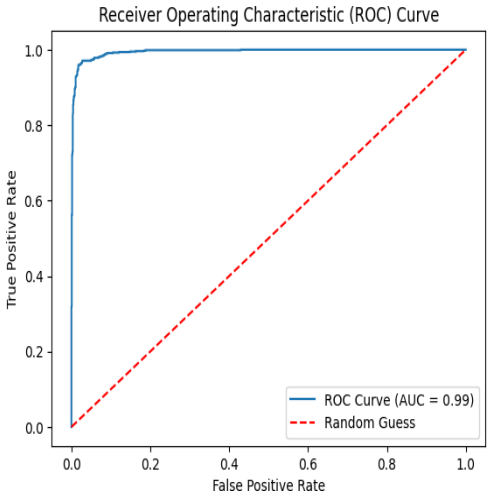
b) Loss of BERT



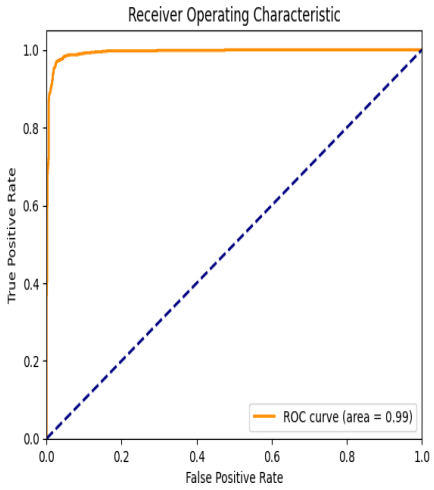
c) Validation Accuracy



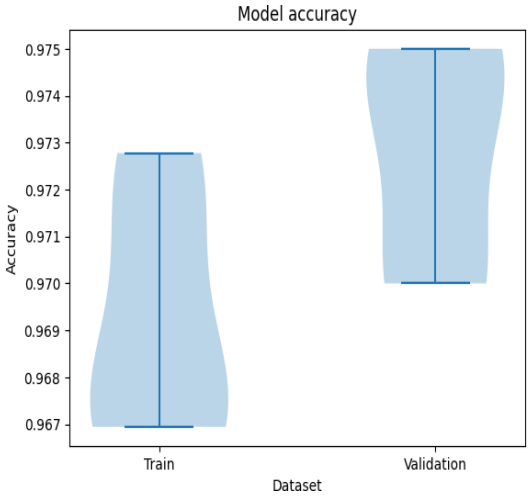
d) Validation Loss



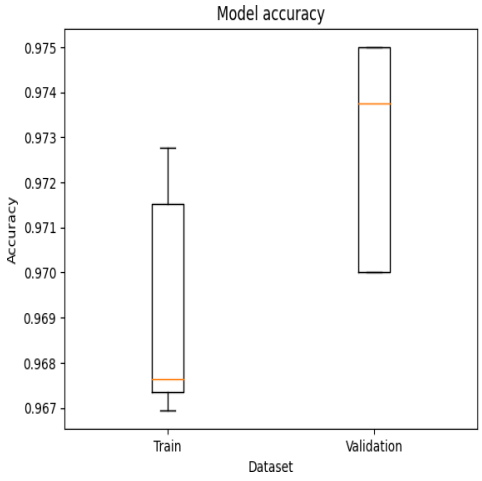
e) ROC of BERT



f) ROC of LSTM + RNN



g) Model Accuracy on BERT



h) Model accuracy on LSTM +RNN

Fig.7. Performance Analysis

4) Result analysis

Every encoder model undergoes training on a training dataset, and during the training process, validation is carried out on the validation dataset. Additionally, the optimal model weights are stored and utilized in addition to gather effectiveness samples for every model in each study by predicting the labels on test data. Figures 7 display the accuracy and loss graphs for training and testing at learning rates of 2e-5, 5e-5, and 8e-5, respectively. Training loss, test and validation accuracy, and rapid convergence in fewer training epochs are all considered in every encoder language model performs exceptionally well. When deploying a deep learning model on a device with limitations on its CPU capacity, memory use, or internet speed such as microcontroller devices, mobile devices, and Internet of Things devices optimization of the model is necessary. Implementing a model on hardware that

has been explicitly created is another use case for optimization. In current study, several kinds of optimization approaches have been developed. For instance, pruning strategies are employed to reduce the model's parameter count. Model pruning is a compression strategy that increases the sparsity of the model by setting model weights to zero during the training phase. It is easy to simplify a sparse model to reduce network latency even further. The other method is known as quantization, and it involves the model utilizing, during deployment, estimated lower precision cooking point weights. Quantization significantly reduces the deployment model size, making it appropriate for microcontrollers and smartphones. The result analysis of the proposed model among the four techniques such as LSTM, RNN+LSTM, BERT and CNN.

Table 1. Comparison of Accuracy among proposed models

Proposed Models	Accuracy	Precision	Recall	F1-Score
LSTM	0.972	0.971	0.969	**0.970**
RNN +LSTM	0.966	0.961	0.968	**0.964**
CNN	0.959	0.967	0.958	**0.962**
BERT	**0.965**	**0.958**	**0.958**	**0.957**

From the analysis it observed that the average high accuracy is obtained as 97.2% for LSTM and the minimal accuracy is obtained as 95.9% for CNN. The average high precision is obtained as 97.1% for LSTM and average low precision is obtained as 95.8% for BERT. Similarly, LSTM shows the high average recall of 96.9% and CNN, BERT model shows the low average recall of 95.8%. The high average F1 score is obtained for LSTM as 97% and low F1 score is obtained as 95.7% for BERT. The result analysis shows that the proposed methodology is efficient in the mental depression detection based on tweets. From the result analysis it is inferred that higher accuracy for tweet classification is obtained on LSTM model. This proposed methodology provided the efficient mental depressive disorder classification and suicidal thought identification.

5) Conclusion

The ability to automatically identify mental depression from text is essential for humans to assist them for their mental stability and to eradicate suicidal ideation. The suggested approach looked at using a multivariate technique to predict human sadness utilizing a one-hot technique for powerful characteristics to identify indicators of depression in the text information. The proposed work uses Twitter data to classify mental depression intensity through trials on four models such as BERT, LSTM, CNN, LSTM+RNN. Through the use of

downstream fine-tuning and transfer learning, a thorough assessment of these models is carried out for the binary classification of depression intensity. The resultant parameters used for the evaluation are accuracy, precision, recall and F1-Score. From the result analysis it is inferred that average high accuracy and precision is obtained as 97% for LSTM. Similarly average high recall and F1-Score is obtained as 96% and 97% respectively. The output of this proposed work can greatly aid machine learning and deep learning assessments and result in a useful user interface for enhanced services. Regarding the area of mental wellness, immediate forecasting analysis of moderate and severe mood disorder states may be facilitated by recognizing signs of anxiety and sadness from the text. In future we will evaluate how well our suggested model performs against a weighted ensemble of soft voting methods from several traditional deep learning approaches. More diverse data may be used to train the model in the future, allowing it to be more broadly used to depressed prognosis in both shorter and lengthier material.