MEGA INSTITUTE FOR ADVANCE RESEARCH AND DEVELOPMENT

Journal of Arts and Linguistics Studies
ISSN Print 2959-0957 ISSN Online 2959-0965
Volume 4 Issue 6 2026
Pages 489-521
Open Access

An Enhanced Ensemble Learning Framework for Stress Detection in Social Media Text Using Linguistic Analysis

Publication Details
Paper Received:
March 28, 2026
Paper Accepted:
June 28, 2026
Paper Published:
June 30, 2026

Abdul Mueed
Department of Computer Science & Information Technology,
University of Southern Punjab (USP) Multan, Punjab, Pakistan.
Email: honeyraajput646@gmail.com
Hamid Ghous
Department of Computer Science & Information Technology,
University of Southern Punjab (USP) Multan, Punjab, Pakistan.
Email: hamidghous@usp.edu.pk
Majid Khawar (Corresponding Author)
Department of Computer Science & Information Technology,
University of Southern Punjab (USP) Multan, Punjab, Pakistan.
Email: majidkhawar@usp.edu.pk

Article DOI: https://doi.org/10.71281/jals.v4i6.757

Abstract
Mental illness and depression are two of the apparently underlying causes of our stress: Stress has become a central mental health problem in the digital age. Individuals post their feelings, concerns and experiences they go through on social media more. This study presents a better ensemble-based learning stress detection system with respect to social media text. The framework combines linguistic analysis with the classical machine learning models and the transformer-based deep learning models. Contextual embeddings generated by BERT, DistilBERT and RoBERTa supplement linguistic features such as the emotional tone, sentiment polarity and the psycholinguistic features. Classical classifiers such as Naive Bayes, Support Vector Machine and Random Forest are used as well. The robustness and predictive accuracy are then boosted by using an ensemble approach or model (i.e. Random Forest feature selection and BERT embeddings). Experiments are run using public datasets of social media posts that contain stress-related postings. The findings indicate that transformer models that are based on deep learning are superior to conventional machine learning models. The performance of the proposed ensemble / hybrid model is overall the highest in terms of accuracy, precision, recall and F1-score. The results proved that the combination of linguistic features, ensemble, and transformer models greatly improve stress detection. The study is a valuable addition to computational mental health as it offers a scalable and dependable framework that may assist in aiding early stress detection and mental health awareness in the online communities. The proposed hybrid model had an accuracy of

95.80 and AUC-ROC value of 99.21 that is exceptionally high and indicates that the proposed model has exceptionally high stress-detecting capabilities.

Keywords: stress detection; social media; ensemble learning; transformer models; BERT; random forest; linguistic analysis; text classification; mental health.

Published by MIARD. Copyright: © the author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY NC) license (https:// creativecommons.org/licenses/by/ 4.0/).

1. Introduction

The issue of Mental Health has become one of the most significant topics of discussion among individuals, communities, and even the government of every country of the 21st century. Stress among others is one of the most common and devastating problems which have long term effects on both mental and physical status. Refractory stress is associated with anxiety conditions, depression, sleeping disorders, heart disorders, and low quality of life [1].

When written on social media, stress linguistically could be implicit, situational and culturally responsive. But it is generally manifested in some language habits such as the use of negative emotion words, words associated with first-person pronouns, cognitive distortions, and syntactic variants. These linguistic traits can be studied with very expressive psychological reflections. Machine learning models particularly ensemble models which integrate the advantages of several classifiers have been promising in representing such patterns so as to effectively classify them [2]. Nevertheless, due to the problem of generalization, the imbalance between the classes, the noisy data and the difference in the manifestation of people on different platforms, the existing models are problematic in some cases [3]. The framework will also seek to enhance the accuracy by combining machine learning and deep learning like the Random Forest and BERT, and resilience of detecting stress in short, noisy, and informal text.

1.1. Global Stress Trends (2019–2024)

The following figure is simply calculated using the actual data of years 2019-2024 in the real world which we can easily analyze to determine the stress level in the entire world. According to repost of the Gallup Global Emotions Reports regarding roughly 35 percent of adults said they are stressed on a daily basis or that they encounter stress. This number increased to 40% in the year 2020 which was the highest. In 2021, the stress level increased and the primary cause was COVID-19 and the level of stress was 41% in the given year. In COVID-19 the majority of the population lost their jobs which caused fear of the financial situation and health issues. The stress levels slowly reduced in the subsequent years, 39% in 2022, 38% in 2023 and 37% in 2024. However, they are still very high compared to the period of the pandemic which is the COVID-19.

As the statistics provided by the World Health Organization (WHO) in 2022 reveal, anxiety and depression have increased by a quarter during the pandemic COVID-19, which means that stress is an acute and deep-seated problem.

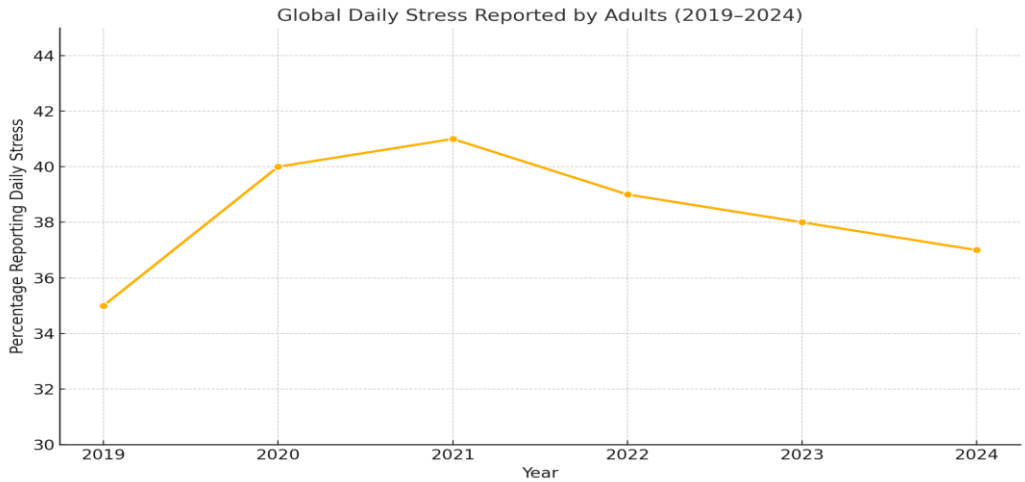

Figure 1. Graph showing Global Stress Report 2019-2024 Gallup Global Emotions 2024

2. Literature Review

This study reviews the complementary role of Large Language Models (LLMs), including Encoder-Decoder, Encoder-Only, Decoder-Only, and Mixture-of-Experts architectures, with physiological monitoring devices such as ECG, EEG, EMG, GSR, PPG, BVP, and SpO2 for real-time stress detection, highlighting challenges like data fusion, computational complexity, and privacy.[2].The study introduces *Calmify*, a web-based multimodal stress detection system that processes text, audio, video, and quizzes using NLP, CNNs, and audio features like MFCCs, pitch, and tone, achieving around 94% accuracy in text models and strong performance on FER2013 and CK+ datasets[3].This research presents a Persian psychological chatbot using the ArmanEmo dataset (7,000+ labeled samples) with models such as ParsBERT, XLM-RoBERTa, and GPT-4o Mini, achieving 75.39% accuracy in emotion detection and improving user stress levels measured through PHQ-9. [4].The study analyzes over 75,000 posts from datasets including MHB, CAMS, HelaDepDet, RMHD, DepressionEmo, and AITA, comparing zero-shot GPT-4o embeddings with MPNet-based classifiers and LIWC psycholinguistic features, achieving ~97% accuracy in binary depression detection[5].This work uses Student Stress Factors and Stress and Well-being datasets with preprocessing, SelectKBest, RFECV, PCA, and classifiers like SVM, Random Forest, Gradient Boosting, XGBoost, AdaBoost, and Bagging, where weighted voting achieved 93.09% and 99.53% accuracy.[6].The proposed Contextual Knowledge Attention GRU (CKA-GRU) model applies attention mechanisms to social media datasets like Dreaddit and Depression_Mixed, achieving 91.39% and 95.20% accuracy, outperforming models like BoW-LR and MentalBERT [7].This study uses the DAIC-WOZ dataset (275 participants) with VADER sentiment analysis and ensemble models including SVM, Logistic Regression, Random Forest, and Gradient Boosting for PTSD detection [8].The research evaluates ChatGPT for

Arabic emotion-sensitive Semantic Role Labeling (SRL) and Cross-Lingual Annotation Projection (CLAP), achieving high performance with BLEU score of 0.90 and emotion detection accuracy up to 96%.[9].A comparative study on Kaggle mental health sentiment data uses TF-IDF and word embeddings with ML models (LR, SVM, RF, LightGBM) and DL models (ALBERT, GRU), where ALBERT achieves approximately approx. 95% F1 score and AUC. It was observed that ALBERT was the only model which outperformed all other models which were being used in the study. If we talk about F1 score results of binary classification were highly superior as compared to the multiclass classification(Multi class classification showed poor results which were below 80 % [10].This study further compares ML (LR, RF, LightGBM) and DL (ALBERT, GRU) models using TF-IDF and embeddings, confirming that deep learning models achieve superior performance (96% F1 and 99.30% AUC) in mental health classification. In this research study the ALBERT outperformed all other models. The results of binary classification were highly superior as compared to the multiclass classification(Multi class classification showed poor results which were below 80 % [11].This review highlights AI-based stress detection using EEG, questionnaires, CNNs for images, and transformer-based models, achieving accuracy between 74% and 98%, while facing challenges like low interpretability and lack of standardized datasets. [12].The study uses ML models such as SVM, Random Forest, Naive Bayes, and Decision Trees with TF-IDF and Bag of Words, achieving 75–80% accuracy, with limitations including small dataset size and need for K-Fold validation. [13]This research proposes CNN-BiLSTM, BiLSTM-attention, CNN-BiGRU models for depression detection, achieving 96.71% accuracy and 0.89 F1 score, with future work suggested on CLEF-2018 and CLEF-2019 datasets. [14].

The Effective Stress Detection Framework (ESDF) uses hybrid ontology with keyword matching, GHS2 probabilistic models, and Tree Alignment techniques, but lacks multilingual support [15]. This study applies ensemble ML methods including Random Forest, Decision Trees, Naive Bayes, SVM, and Logistic Regression with boosting, bagging, and voting, achieving an F1 score of 76.60%, but lacks BERT and multi-source data[16].A neural network-based model is developed using preprocessing, training, and evaluation phases, achieving 95% accuracy and 94% recall with batch size 64 and 10 epochs. [17].This study uses KNN, SVM, Decision Trees, and Naive Bayes with TF-IDF and POS tagging on 1500 tweets, where ensemble methods achieve an F1 score of 64 [18]. Using DLATK with HappierFun tokenizer and LIWC dictionaries, this study analyzes stress-related language patterns but lacks causal insights and fails to detect stress in inactive users [19].

The study uses ML models including SVM, RF, DT, KNN, NB with CatBoost, achieving the best accuracy of 82.6% on a dataset split of 805 training and 205 testing samples.[20]. A no-lexicon-based framework using techniques such as CAE, LTP, MTL, and MAM achieves 0.685 accuracy and improves to 0.946 in more complex scenarios. [21].This study applies SVM, NB, DT, and RF on 2500 tweet-based sentences, identifying stress through positive and negative words, achieving accuracy above 90%.[22].Using Reddit and Twitter datasets with different splits (70/30, 80/20, 90/10), SVM achieves the highest accuracy of 0.81 in stress detection [23].This study uses the ERT dataset aligned with DASS-21 scale on 200 participants, applying DT, LSTM, RNN, and MLP, achieving R values of 0.7 (depression), 0.44 (anxiety), and 0.52 (stress), but LSTM struggles with learning. [24].Hybrid models combining LSTM, SVM, and VADER lexicon-based approaches show promising stress detection but lack real-time analysis and multimodal data [25].Ensemble learning using SVM, DT, KNN, and XGBoost on

a Kaggle dataset (2838 instances) achieves 73% accuracy, with slight improvements using neural networks [26].A Flask-based web system uses TF-IDF with an 80/20 split to detect stress levels (low, moderate, high), achieving 86% accuracy but limited to local deployment.[27] .Deep learning models (LSTM, RNN) trained on 90/10 data split achieve 99% accuracy in depression detection using text-based emotional patterns [28].

Using CLPsych 2015, eRisk, and Reddit datasets, ML models (LSTM, RNN, LR) achieve accuracies around 0.65–0.75, with ensemble + lexicon models performing best.[29].Bernoulli Naive Bayes applied on Reddit and GitHub datasets achieves 74% accuracy but lacks contextual understanding [30].The study uses GPT, Logistic Regression, ANN, and Doc2Vec features, where ANN achieves the highest accuracy of 80.34%, limited by dataset size[31].This study uses NB, SVM, KNN, and RF on tweet datasets labeled as stress/non-stress, where SVM achieves the highest accuracy of 71.21%, limited by single-source data[32].ML models (NB, SVM, RF, LR) applied to Reddit datasets detect depression and suicidal thoughts, achieving ~77% accuracy, but lack deep learning and real-time intervention strategies[33].A stacking ensemble model using TF-IDF, ELMo, and GloVe embeddings on datasets like Depression Reddit achieves 99.54% accuracy, but lacks multimodal data and raises privacy concerns.[34].A BERT-based model with CRF layer applied to Chinese social media data targeting postgraduate students achieves 92.55% accuracy.[35].CNN-based classification on Twitter data divides posts into positive, negative, and neutral classes, but is limited by single dataset usage.[36].This system uses NLP, LDA topic modeling, emotion analysis, and BERT to classify emotions into five categories, achieving 94% accuracy and 83% F1 score[37].A hybrid system combining CNN (feature extraction), SVM (classification), and KNN (hospital recommendation) achieves up to 84.2% accuracy using Facebook data[38].This study uses SVM, NB, DT

with BERT and Word2Vec on Reddit data, achieving the highest accuracy of 79.8% with Logistic Regression[39].A real-time system using Apache Kafka, PySpark, and Airflow on the Dreaddit dataset achieves 80.65% accuracy using XLNet.[40].The study compares SVM, LSTM, and BiLSTM, showing that BiLSTM provides better accuracy but requires more computation time [41].Using Tweepy API and models like SVM, NB, RF, BLSTM, and BERT, this study achieves 83.60% accuracy but is limited to English language and specific demographics.[42].Transformer models including BERT, RoBERTa, BART, and DeBERTa achieve around 84% accuracy across six stress categories[43].ML models (SVM, RF, NB, DT) applied on 2969 tweets using TF-IDF achieve up to 90% accuracy, though limited by small dataset size [44].

2.1. Research Gap

From the literature reviewed, it is clear that stress classification has seen notable advancements through both traditional machine learning and transformer-based models. Yet, there is a noticeable lack of studies exploring how contextual embeddings can be effectively combined with feature importance-based selection to boost both accuracy and interpretability. To bridge this gap, the following chapter introduces a hybrid framework that will serve as the foundation for the methodology adopted in this study.

3. Methodology

3.1. Dataset Description

The first dataset downloaded on the stress i.e. Kaggle, which is in the public domain, is called stress.csv (Rajani (2022), the Human Stress Prediction dataset available on Kaggle) and second is called combined dataset.csv (Sarkar Sentiment Analysis for Mental Health (2024), available on Kaggle). These two datasets were combined / merged to train the models. The second dataset was a multi-class data (the total number of classes was 7, used 4 classes) which was transformed into

binary data 1 and 0 following the matching of the type of labeling of both datasets and posts related to stress i.e. Anxiety, Stress, Suicidal were identified and combined and used as a training data. The better training process was also carried out through the normal posts. The data set has a total of 36398 instances.

	A	B
1	text	label
2	oh my gosh	1
3	trouble sleeping, confused mind, restless heart. All out of tune	1
4	All wrong, back off dear, forward doubt. Stay in a restless and restless place	1
5	I've shifted my focus to something else but I'm still worried	1
6	I'm restless and restless, it's been a month now, boy. What do you mean?	1
7	every break, you must be nervous, like something is wrong, but what the heck	1
8	I feel scared, anxious, what can I do? And may my family or us be protected :)	1
9	Have you ever felt nervous but didn't know why?	1
10	I haven't slept well for 2 days, it's like I'm restless. why huh :([].	1

Figure 2. First ten instances of the dataset

Figure 3. Frequently Used words in the dataset

3.2. Data Preprocessing

The following preprocessing techniques were used to clean the data:

Removal of HTML tags, Special characters.

Conversion of the text into lower case.

Token filtering is used for removal of stop words.

3.2.1. Naïve Bayes

Naive bayes is another probabilistic Classifier that is founded on the Bayes Theorem. The quoted classifier makes use of the decisions that are founded on the probabilities of an event that have been determined by the Bayes Theorem [45]. The Multinomial Naïve Bayes classifier is actually implemented by using a Scikit-learn pipeline along with TF-IDF feature extraction process. Predictions are made using "predict_proba" to support evaluation such as ROC-AUC etc.

The equation of Bayes theorem is given by:

$$\boldsymbol{P(A \mid B) = P(B \mid A) \cdot P(A)/P(B)} \qquad \textit{... Equation 1}$$

Where:

P(A|B): The probability of the occurrence of event A in terms of event B occurrence (it is also known as the posterior probability).

P (B|A): This is the probability that event B would occur because of the event A that occurred.

P(A): The probability of the occurrence of A only (this is known as prior probability).

P(B): The likelihood of event B to occur without any intervention (also referred to as the evidence or normalizing constant).

3.2.2. Supports Vector Machine

The Support Vector Machine (SVM) is an algorithm of supervised learning which can be effectively applied in high dimensional spaces. It works by determining the hyperplane which isolates the information into different categories in the best way.

[46]. It is trained on TF-IDF features by using a Scikit-learn pipeline for ensuring the consistent preprocessing as well as linear kernel is used to handle high sparse feature space. The probability estimation is also enabled to ensure the threshold predictions such as ROC-AUC.

3.2.3. Text Embedding and Model Training (BERT Family)

The BERT family is used to extract the deep meaning of words and sentences through contextual embeddings by using deep learning approach. The models which are implemented in this study are RF BERT (which is combination of the Random Forest importance based selected features which are concatenated with the BERT contextual embeddings), Distill BERT as well as RoBERTa. The techniques such as attention masking, tokenization, padding is used to prepare the input, drop out and dense layers used to control the output and sequence length as well as early stopping are applied to improve the training process of the model.

1) BERT Tokenization and Encoding

The BERT tokenizer was used to generate input id and attention mask. These are used to feed in the model i.e. BERT. The tokenizer basically converts the raw text into sequence of tokens which are further mapped into input IDs. These IDs are now representing the words or sub words that can be comprehended by a BERT model. The length of the sequence 128 is selected in order to have the same length of tokens and truncation to truncate the tokens that are beyond the 128 length.

2) RF Guided BERT Model Architecture

The RF- Guided BERT model is an integration of the more traditional feature selection and more contextual embeddings to assist in enhancing the performance of stress classification. During the first step, TF-IDF (unigrams and bigrams) features are filtered and a Random Forest classification model is trained that labels the top 500 most informative features (based on the feature importance scores). The chosen features are then joined to the CLS version of embedding a token

obtained by a fine-tuned BERT model to create a hybrid feature representation. The lexical and deep contextual semantics together are used to make decisions with greater accuracy by the combined feature vector that goes through dropout and dense layers to regularize the outputs of both features and classify them. In addition, the implementation of the RF-Guided BERT introduces sequence preprocessing, padding and truncation to make sure that all the inputs have the same length of 128 tokens to BERT. A hybrid embedding of CLS token embedding of BERT combined with top 5000 features is then subjected to a Random Forest and then through 2 dropout layers (dropout rate 0.3) and a dense layer of 64 units with a final binary classification output layer. The energy of model is training that uses Adam optimizer, binary cross-entropy loss, and early stopping using lose OnPlateau is used to ensure that there is no overfitting and improve stability in the convergence.

3) DistilBERT Model Implementation

DistilBERT model was applied as a lightweight transformer in reducing the computational based load in the contextual understanding of BERT. The tokenization was carried out with the help of the pre-trained tokenizer and the DistilBERT-base-uncased to transform the input text into both, token IDs and attention masks. In this model architecture, it has a custom DistilBERT layer, dropout layer to ensure regularization and dense layers to carry out classification, which is binary. The implementation of this model was carried out with the Adam optimizer and binary cross-entropy loss function and trained on the basis of several epochs with the validation monitoring to achieve stability. Its maximum sequence length of 128 is employed as well as the CLS token embedding is both retrieved out of the hidden layer and finally through the dropout and dense layer to perform the binary classification.

4) RoBERTa Model Implementation

The RoBERTa model had the same architecture as the DistilBERT but used the Roberta-base pre-trained model as the contextual based embedding. Dynamic padding and truncation were used to ensure that sequence lengths are of uniform length. The tokenization was undertaken through the quoted model tokenizer. The RoBERTa layer was responsible for the to extract the contextual embeddings from the output of pretrained model. The resultant element was a sigmoid passed through the output layer. The values that the sigmoid function gives out are 0 and 1 that is perfect when it comes to binary classification like stress and no stress. This architecture which combines RoBERTa's advanced contextual embeddings as well as padding which is dynamic, dense layers and the sigmoid output function, proved very effective solution for binary classification tasks such as stress detection. The custom RoBERTa layer extracted contextual embeddings which were passed through dense layers and a sigmoid output unit for classification such as binary classifications. The model was run and optimized by Adam and trained through early stopping and learning rate reduction callbacks to enhance the performance and avoidance of overfitting. To generate the input IDs and attention masks the RoBERTa-Base pretrained tokenizer is used and model optimization is ensured through using Adam optimizer with cross entropy loss.

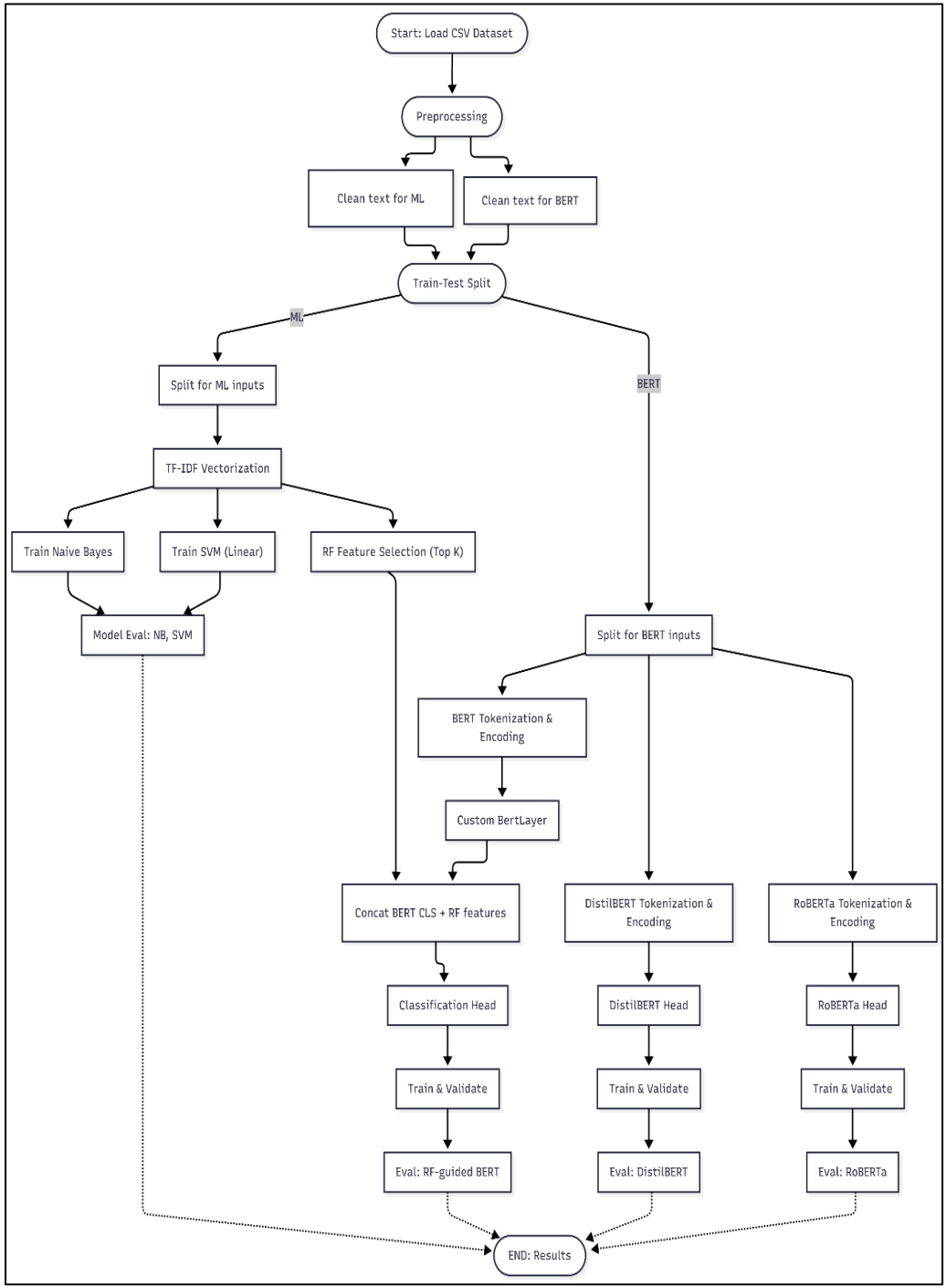

Figure 4. Methodology Diagram

The diagram, as illustrated above, is the diagrammatical representation of the entire methodology where it is illustrated that once the dataset is loaded it was cleaned by the implementation of the preprocessing techniques which include removal of stop words, lower casing, removal of HTML tags and readiness to proceed with the process of training the ML i.e. SVM, random forest and naive bayes and the DL techniques i.e. tokenization, padding and truncation were used. The ML models were trained with the help of TF-IDF vectorization and then RF features (top 5000) were approved to BERT to be concatenated to obtain improved final prediction. Model Evaluation strategy such as train-test split was used to assess the performance of our deep learning and machine learning models. The results will be talked about in the following chapter after the application of our specific methodology which we already discussed / explained in this chapter.

4. Results and Discussion

We will discuss the results obtained after the implementation of the purposed methodology. The results of the research are represented in this part with the help of a strong methodological framework. This part presents the results of using the techniques adopted, as it provides information on the efficiency and soundness of the proposed model.

4.1. Naïve Bayes

It can be helpful in the process of finding significant words and minimizing weight of frequent words. The Multinomial Naive Bayes is based on the assumption that the occurrence of any word in the text is distributed in a multinomial that makes it ideal and effective in the process of spam detection and sentiment analysis or any other text classification process like stress detection problem where the frequency of words is very crucial. [45]. To sum up the Multinomial Naïve Bayes as well as TF-IDF features proved to be very well performing and effective choice for the

task. The model depicted accuracy of 0.8622, recall of 0.9345 and F1 score of 0.8745 as well as precision of 0.8218 and ROC-AUC of 0.9590.

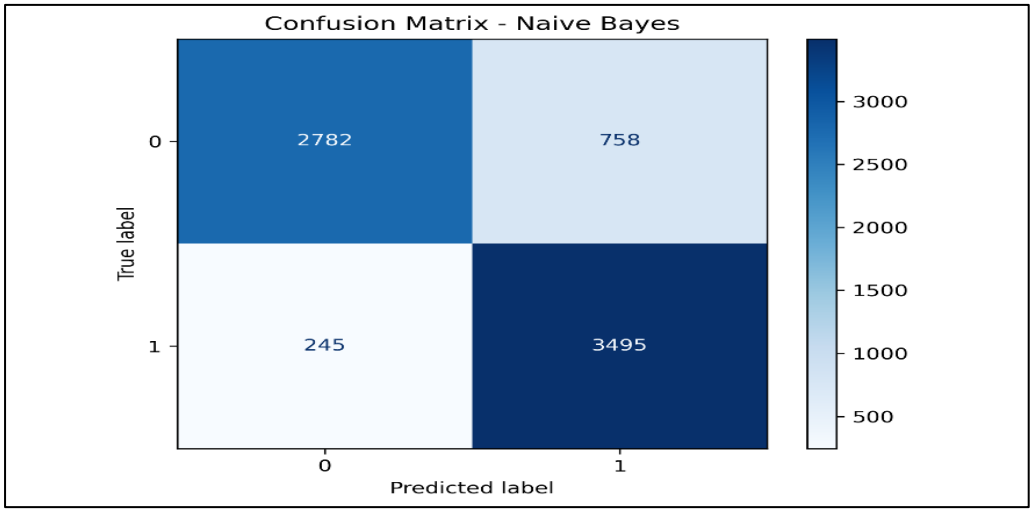

Figure 5. Confusion Matrix of Naive Bayes

The model achieved True Negative Rate TN=2782 and False Positive Rate FP=758 False Negative Rate FN=245 and True Positive Rate TP=3495 which depicts that how the model performed in comparing the actual value versus predicted values.

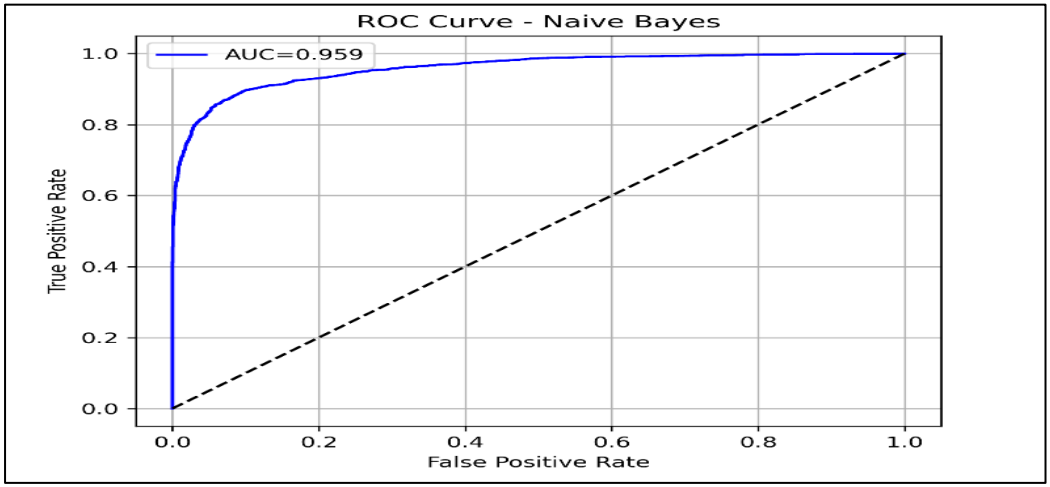

Figure 6. ROC Curve of Naive Bayes

The model achieved ROC-AUC of 0.959, curve illustrates that model has strong discriminative properties.

4.2. SVM (Supports Vector Machine)

It is a very powerful and quick model that is grounded on Supervised Machine Learning to conduct tasks such as classification. It merely works by finding hyperplane separating datapoints of different classes [46].SVM classifier provides the accuracy value of 0.9405, the recall rate of 0.9377 and F1 of 0.9419. It implies that the model was effective in categorizing around 93.18 percent of the cases and this is a definite indication that the model can be applied to make valid forecasts on the unknown data. The good compromise between precision and recall was 0.9419 and F1 score. It was noted that the model yielded fairly satisfactory outcomes in terms of precision that were 0.9460.

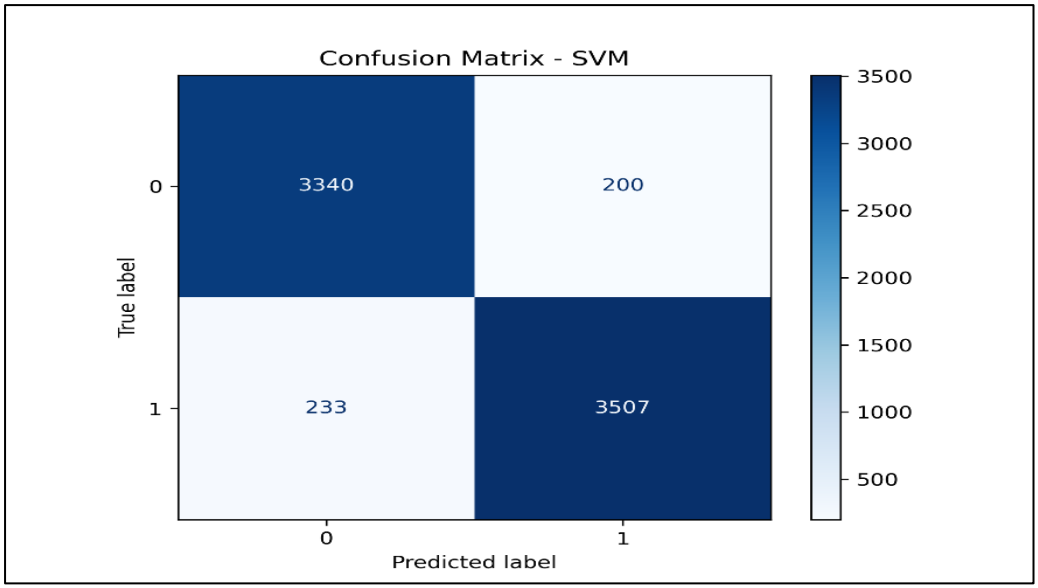

Figure 7. Confusion Metrics of SVM

The model achieved True Negative Rate TN=3340 and False Positive Rate FP=200 False Negative Rate FN=233 and True Positive Rate TP=3507 which

depicts that how the model performed in comparing the actual value versus predicted values.

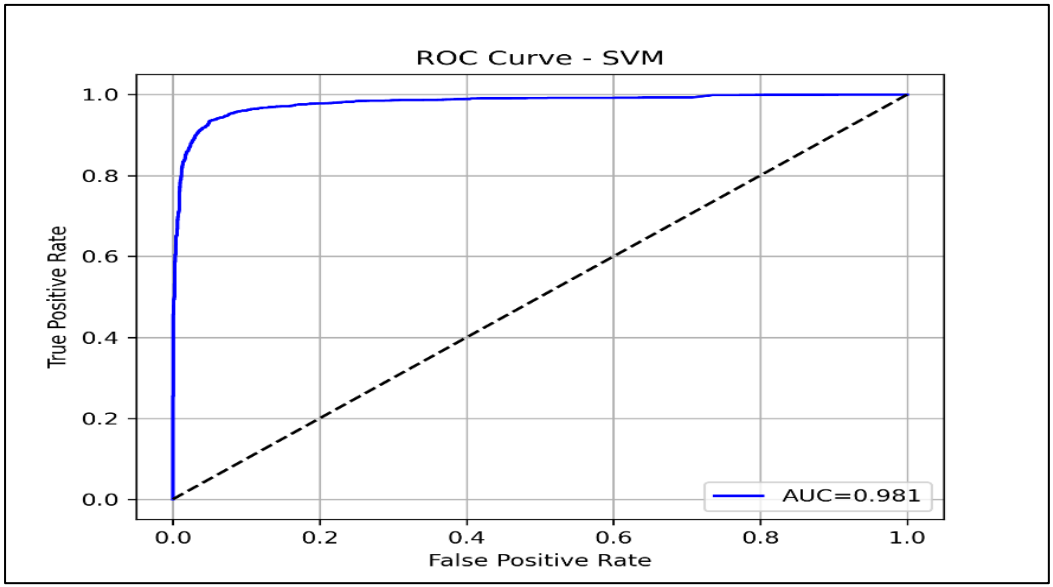

Figure 8. ROC Curve of SVM

The model achieved ROC-AUC of 0.981, curve illustrates that model has strong discriminative properties.

4.3. RF Guided BERT Classifier

Results are quite far-reaching to prove the thesis that this combined approach is effective. RF-Guided BERT has the highest accuracy (0.9580) and F1-score (0.9593), and a near-perfect AUC (0.9921), attribute it had superior performance in comparison with all the models that we tested. This implies that it was the least vulnerable to false-alarming as well as false-alarming stressed and non-stressed messages respectively. It even surpassed the norm, very potent BERT models (DistilBERT and RoBERTa). This RF-Guided BERT model is an excellent advancement in terms of the task of automatic stress detection, in which accuracy is a crucial factor in the potential usage of health issues.

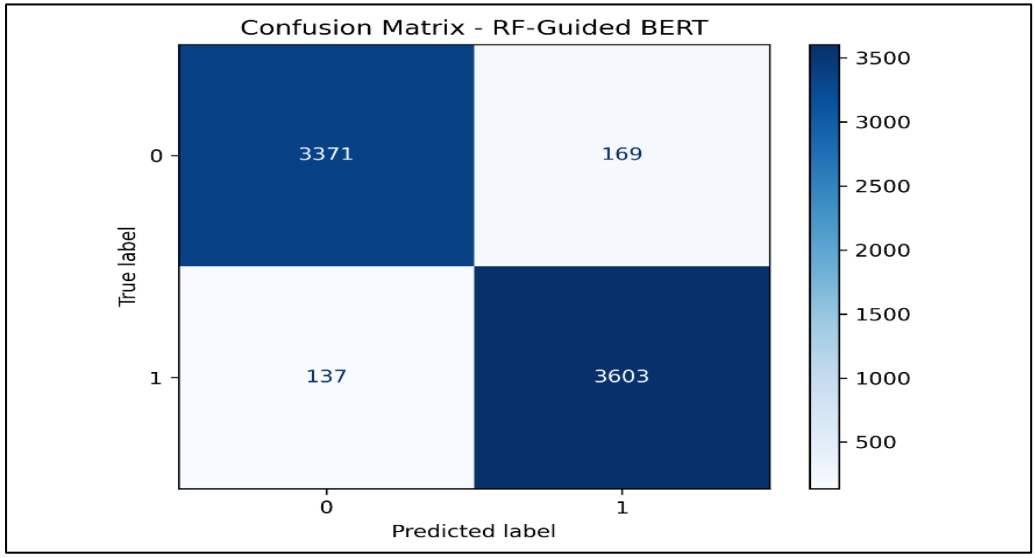

Figure 9. Confusion Metrics of BERT Classifier

The model achieved True Negative Rate TN=3371 and False Positive Rate FP=169 False Negative Rate FN=137 and True Positive Rate TP=3603 which depicts that how the model performed in comparing the actual value versus predicted values.

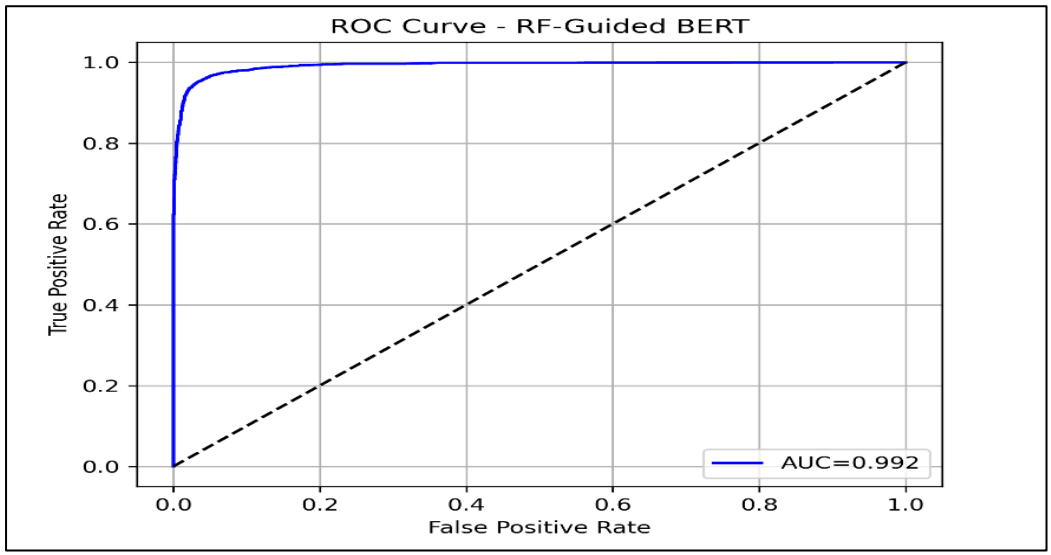

Figure 10. ROC Curve of BERT Classifier

The model achieved ROC-AUC of 0.9921, curve illustrates that model has exceptional discriminative performance.

4.4. DistilBERT Model

DistilBERT is just a small variant of the BERT model and is mostly destined to be capable of similar results with fewer training as compared to BERT. It has maintained the majority of the power of language understanding as that of BERT without being heavy and efficient. Regardless of the smaller size, DistilBERT is able to keep the majority of the BERTS language knowledge such as the ability to implicate complex relationships and contextual functions of the text [47]. This model was accurate in this study with the F1 score of 0.9035 and Recall 0.9503. The task of binary classification using DistilBERT was used whereby the aim was to produce text in either of two categories like stressed or not stressed. Nevertheless, being a light version, DistilBERT performed well with the accuracy of 0.8957 which implies that the model is right 89.57 out of 100 cases. DistilBERT recalls had a very high recall of 0.9503 which implies that it discriminated 95.03 percent of the actual positive cases or more than that in its category.

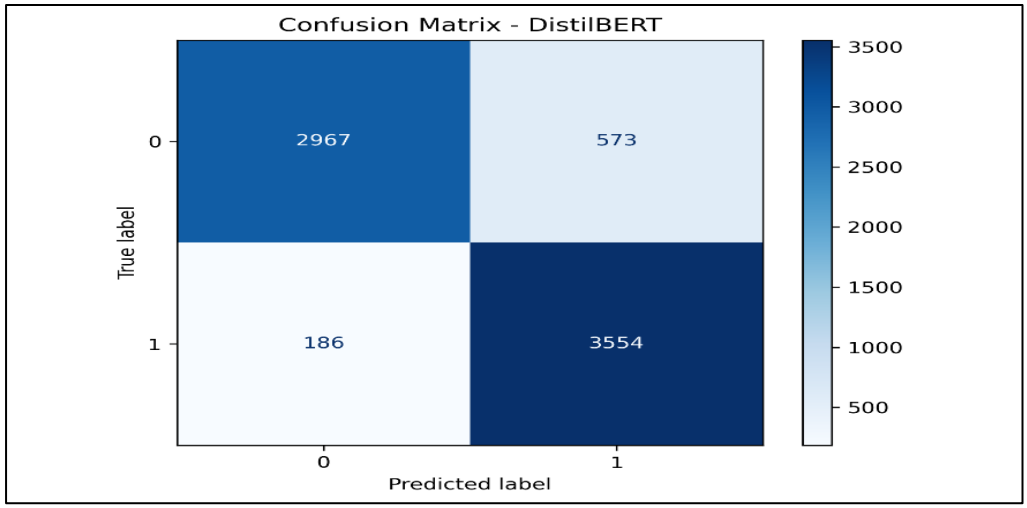

Figure 11. Confusion Metrics of DistilBERT

The model achieved True Negative Rate TN=2967 and False Positive Rate FP=573 False Negative Rate FN=186 and True Positive Rate TP=3554 which depicts that how the model performed in comparing the actual value versus predicted values.

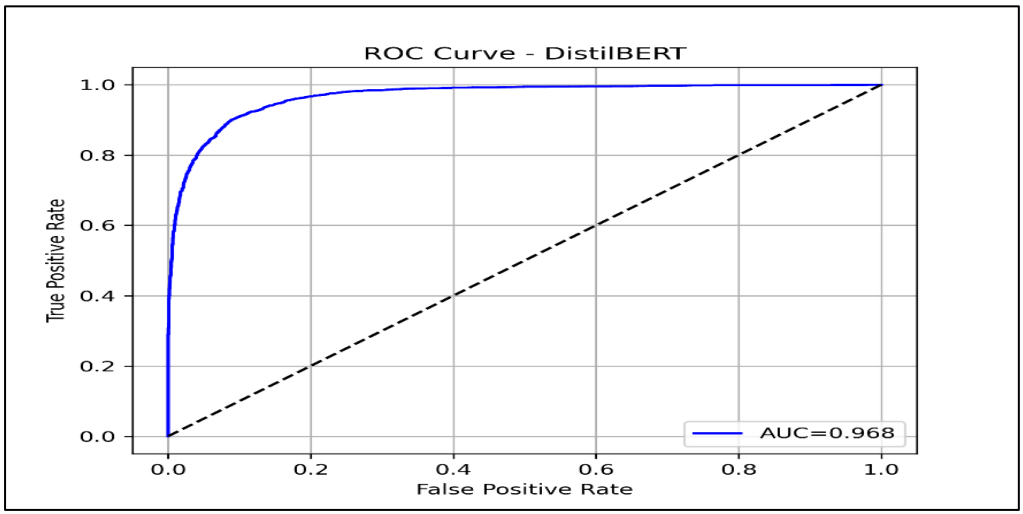

Figure 12. ROC Curve of DistilBERT

The model achieved ROC-AUC of 0.968, curve illustrates that model has exceptional discriminative properties.

4.5. RoBERTa Model

RoBERTa (which is basically Optimized Approach of BERT) is an improved version of BERT that refines its training process for achievement of better performance. It removes next sentence prediction, uses larger mini-batches, and trains on more data with masking which is dynamic in nature [47]. We recorded that this model during after implementation achieved accuracy rate of 0.9037, F1 score of 0.9089 as well as Recall of 0.9345. The F1 value is high which means that the model is efficient in both the detection of stressed texts and minimization of false alarms. Besides, RoBERTa reached a high recall of 0.9345, which implies that it recognized 93.45 percent of the real positive cases in the text classification.

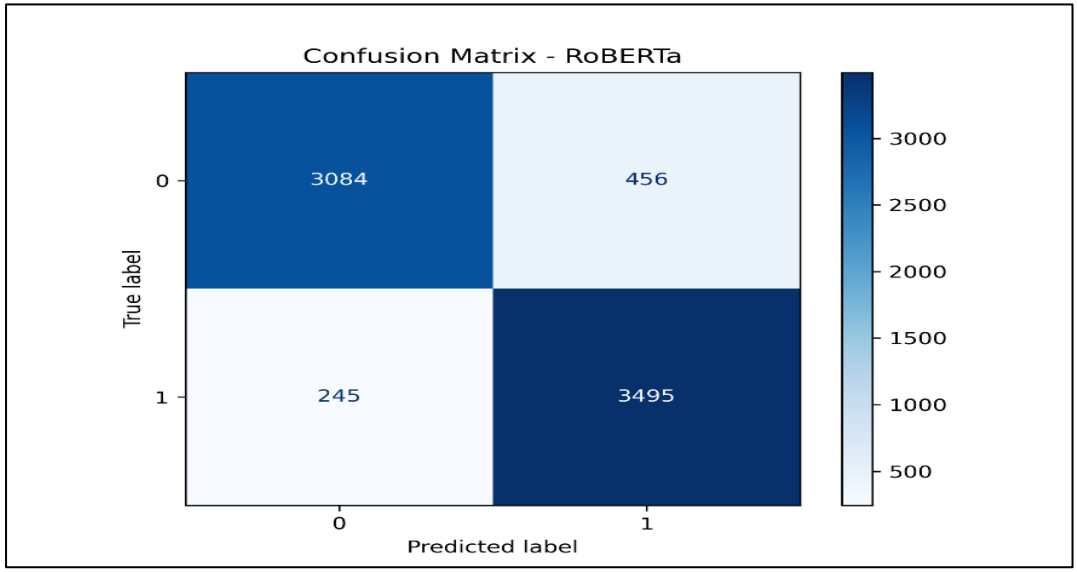

Figure 13. Confusion metrics of RoBERTa

The model achieved True Negative Rate TN=3084 and False Positive Rate FP=456 False Negative Rate FN=245 and True Positive Rate TP=3495 which depicts that how the model performed in comparing the actual value versus predicted values.

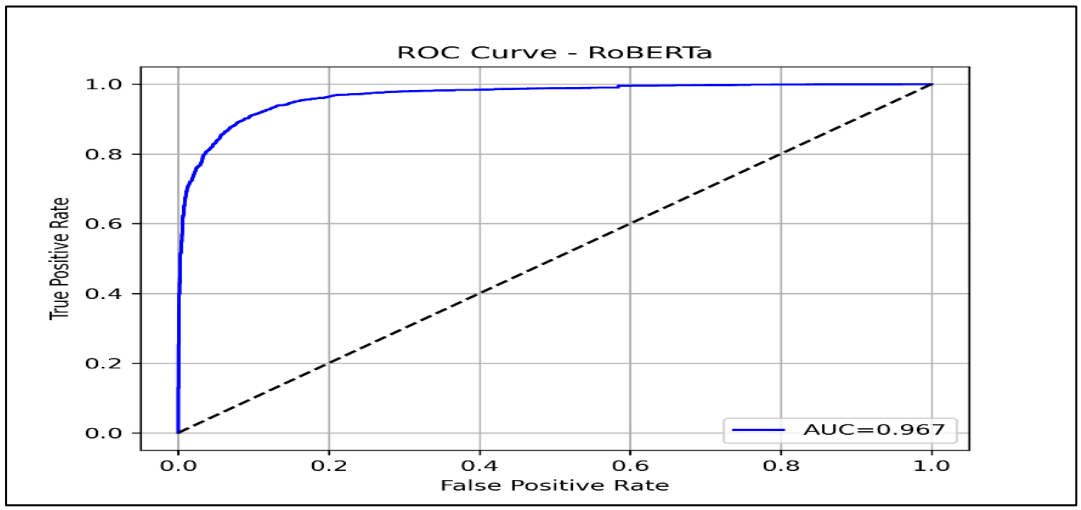

Figure 14. ROC curve of RoBERTa model

The model achieved ROC-AUC of 0.967, curve illustrates that model has exceptional discriminative properties.

Table 1. Comparison Table of All Models

Sr.	Model Name	Accuracy	F 1 score	Recall	Precision	AUC
1	Supports Vector Machine	0.9405	0.9419	0.9377	0.9460	0.9805
2	Naïve Bayes	0.8622	0.8745	0.9345	0.8218	0.9590
3	RF Guided BERT	0.9580	0.9593	0.9634	0.9552	0.9921
4	DistilBERT	0.8957	0.9035	0.9503	0.8612	0.9676
5	RoBERTa	0.9037	0.9089	0.9345	0.8846	0.9673

4.6. Discussion

As can be seen from table given below, the methodological and conceptual differences are also pronounced in this research and the initial work by Zhang et al. (2025) and Ding et al. (2025). Even though the two previous experiments focus on the general mental illness classification using the entire seven-class Kaggle dataset and are primarily based on the performance comparison between machine learning and deep learning models, the proposed study suggests reformulating tasks and redesigning the architectures. To be more precise, this study is less broad as it focuses not on general multi-classification of mental illness but on a geographically more specific topic of stress-spectrum detection (Normal, Anxiety, Stress, Suicidal) and thus focusing more on practically actionable problem. Proposed RF-Guided BERT will bring feature-level fusion, i.e. combining Random Forest-selected features with contextual BERT embeddings. The outcome of such an integration is a very discriminative model with an AUC of 0.9921 and a weighted F1-score of 96 percent based on deep semantic

representation, cumulative strength of a decision. Although the authors of the research by Ding et al. are reporting such a slightly higher F1-score (96.50%), AUC (99.30%) the best model is ALBERT and it took 354.18 minutes to be trained, however in the other side our hybrid model reported competitive results within 74.33 minutes (approx.). On the other hand, Zhang et al got weighted F1 score of 95.6% and RO-AUC of 95% which are lower than results of our study i.e. weighted F1 96% and ROC-AUC 99.12% In addition, the proposed model performs competitively but includes a feature of explainability based on the Random Forest feature importance, which provides only partial or no interpretation in the comparatively performed studies.

Table 2. Comparison with Previous Studies

Aspect	**This Study (Proposed Method)**	**Zhang et al., 2025 [10]**	**Ding et al., 2025 [11]**
Dataset Source	Combined dataset (Kaggle)	Kaggle – Sentiment Analysis for Mental Health	Kaggle – Sentiment Analysis for Mental Health
Dataset Usage	Subset of dataset (stress-related classes only) and Reddit Dataset (collectively)	Full dataset (Larger Dataset)	Full dataset (Larger Dataset)
Classes Used	Normal, Anxiety, Stress, Suicidal	7 classes	Same 7 classes
Feature Representation	TF-IDF, contextual embeddings (BERT), RF-selected features	TF-IDF, word embeddings	TF-IDF, word embeddings
Models Implemented	Naïve Bayes, SVM, DistilBERT, RoBERTa, RF-Guided BERT	Logistic Regression, SVM, Random Forest, LightGBM, ALBERT (Best Model), GRU	Logistic Regression, Random Forest, LightGBM, ALBERT (Best Model), GRU
Hybrid Learning	feature-level fusion (RF + BERT)	No hybrid model proposed	No hybrid model proposed

Best Accuracy	95.80% (4 classes)	Not reported	Not reported
Best F1-score	96%(weighted)	95.76%(weighted)	96.50%
AUC	0.9921	95%	99.28%
Evaluation Strategy	Train–test split	Cross-validation	Cross-validation
Training Time (Best Model)	Approx. 75 minutes	Not reported	354.07 minutes

Table 3. Comparison Table with previous studies

Aspect	This Study (Proposed Method)	Turukmane et al. (2025) [48]	Rahman et al. (2024) [26]
Dataset Source	Combined dataset (Reddit dataset and subset of mental health dataset from Kaggle)	Kaggle dataset (human stress prediction 2838 instances)	Kaggle dataset (human stress prediction 2838 instances)
Dataset Usage	Combined dataset (Reddit dataset and subset of mental health dataset from Kaggle)	Entire Kaggle dataset (all 2,838 instances)	Entire Kaggle dataset (all 2,838 instances)
Classes Used	Stress, No Stress	Stress, No Stress	Stress, No Stress
Research Objective	Enhanced ensemble learning for stress detection with linguistic analysis	Automated stress detection using ML models and ensemble	Comparative analysis of ML models + ensemble + sequential neural network for stress prediction
Feature Representation	TF-IDF, (BERT+RF Features)	TF-IDF, Bag-of-Words	Count Vectorizer (word counts), label encoding
Models Implemented	Naïve Bayes, SVM, Random Forest, DistilBERT, RoBERTa, RF-Guided BERT	Naive Bayes, SVM, Decision Tree, Random Forest, Weighted Ensemble	Naive Bayes, SVM, Decision Tree, KNN, XGBoost, Voting Classifier, Sequential Neural Network

Hybrid Learning	Yes – feature-level fusion (BERT + RF Features)	No –ensemble only	No-Simple ensemble + SNN but no hybrid embeddings
Best Accuracy	95.80%	Not exactly reported but graphs show low accuracy	76% (Sequential Neural Network)
Best F1-score	95.93%	Not reported	73%
Best AUC	0.9921(Hybrid model)	Not reported	Not reported
Training Time (Best Model)	Approx. 75 minutes	Not reported	Not reported
Evaluation Strategy	Train–test split	Train–test split	Train–test split

The comparison table helps clearly understand that the proposed study is superior to Turukmane et al. (2025) [48] and Rahman et al. (2024) [26] in several important aspects. Quantitatively, the proposed approach has a maximum accuracy of 95.80 and F1-score of 95.93, which is much larger than the 76% or worse results by Rahman et al. and Turukmane et al. as well. The above AUC of 0.9921 also substantiates the strength and reliability of the model. On the whole, the combination of deeply contextual embeddings, classical models of ML and ensemble with preprocessing and language analysis allows this study to provide a better, methodologically, and practically more effective solution to the detection of stress using social media textual data.

5. Conclusion and Future Work

5.1. Conclusion

As a conclusion, this study was able to develop and test proposed hybrid architecture, RF-Guided BERT, to the task of binary text classification on the combined dataset for the purpose of stress detection through linguistic analysis. The fundamental insight is that the performance of standalone models can be

improved by merging the deep contextual lexical features to a Transformer model. The RF-Guided BERT model demonstrated the best performance in all measurements such as accuracy of 0.9580 and AUC of 0.9921, setting an efficient record in this task. This result was not obtained in a vacuum, but it was the fruit of an intensive comparative design with robust benchmarks, such as the usual strength of Support Vector Machines, the efficiency of DistilBERT, the advanced pre-training of RoBERTa, and the simplicity of Naïve Bayes. The obvious ranking of the performance supports the architectural assumption and gives the practitioners the convenient guideline on what and how to use a model, depending on their particular level of accuracy and resource limitations.

5.2. Limitations of the Study

Although it has made the following contributions, this study has a number of limitations.

The study focused on the English language social media text, thus limiting extrapolation to the multilingual context.

Only the textual information was taken into consideration and other modalities like images audio or physiological indicators were not.

The data sets were mostly collected using Reddit that can create language bias in the platform. The assessment was done in an off-line experimental manner but not in live mode.

These shortcomings give a very strong guidance on how the research could be conducted in future.

5.3. Future Work

Future research can extend this work in many directions such as:

This work can be further expanded in a number of ways in future research.

The first one is that the framework may be extended to multilingual stress detection to enhance your applicability to a wider population. Second Multimodal

data (e.g. pictures audio or behavioral indications) may be included to increase the detection precision.

Third early intervention systems can be investigated with regard to real time deployment of the framework.

Fourth explainable artificial intelligence methods can be used to enhance transparency and credibility in stress forecasting.

Fifth Online stress detectors are only modifiable according to the linguistic patterns of an individual as the modules need to be tailored with the passage of time.

Lastly, with the assistance of mental health professionals, predictions can be validated and used ethically in the real world.

Reference

"Gallup Global Emotions," 2024.

O. Wahab and M. Adda, "ScienceDirect Comprehensive Literature Review on Large Language Models and Smart Monitoring Devices for Stress Management," *Procedia Comput. Sci.*, vol. 257, pp. 166–173, 2025, doi: 10.1016/j.procs.2025.03.024.

P. P. Patil, "A Multi-Modal Approach for Detecting Stress and Personalised Stress Reduction in Modern Life," vol. 18, no. 04, pp. 46–52, 2025.

S. Jafari *et al.*, "Psychological Health Chatbot , Detecting and Assisting Patients in their Path to Recovery," pp. 64–77, 2025.

S. Kim, "Interpretable Depression Detection from Social Media Text Using LLM-Derived Embeddings".

S. I. Ovi, J. Hossain, R. A. Rahi, and F. Akter, "Protecting Student Mental Health with a Context-Aware Machine Learning Framework for Stress Monitoring".

V. Devi, J. Rohini, T. Venkatesh, and S. D. Vaithilingam, "Stress and Depression Classification in Social Media using Contextual Knowledge Attention based Gated Recurrent Network," vol. 18, no. 3, pp. 420–431, 2025, doi: 10.22266/ijies2025.0430.29.

V. Prajapati, D. Sanghani, V. Undhad, B. Wadhvani, and B. Shah, "Predictive Model for Post-traumatic Stress Disorder," vol. 14, no. 81, pp. 66278–66285, 2023.

F. Senator, A. Lakhfif, and I. Zenbout, "Leveraging ChatGPT for Enhancing Arabic NLP : Application for Semantic Role Labeling and Cross-Lingual Annotation Projection," *IEEE Access*, vol. 13, no. December 2024, pp. 3707–3725, 2025, doi: 10.1109/ACCESS.2025.3525493.

Y. Zhang *et al.*, "Employing Machine Learning and Deep Learning Models for Mental Illness Detection," pp. 1–22, 2025.

Z. Ding, Z. Wang, Y. Zhang, Y. Cao, and Y. Liu, "Trade-offs between machine learning and deep learning for mental illness detection on social media," pp. 1–14, 2025.

A. Pandya, N. Singhal, P. Panchal, J. Patel, and L. Patel, "Evaluating current trends in stress and depression detection using artificial intelligence and machine learning," *Intell. Hosp.*, no. December 2025, p. 100047, 2026, doi: 10.1016/j.inhs.2025.100047.

M. R. Febriansyah, Nicholas, R. Yunanda, and D. Suhartono, "Stress detection system for social media users," *Procedia Comput. Sci.*, vol. 216, pp. 672–681, 2022, doi: 10.1016/j.procs.2022.12.183.

J. Philip, S. Yongchareon, and V. Liesaputra, "An attention-based CNN-BiLSTM model for depression detection on social media text," *Expert Syst. Appl.*, vol. 249, no. PC, p. 123834, 2024, doi: 10.1016/j.eswa.2024.123834.

M. Tajuddin and M. Misbahuddin, “Analysis of Social Media for Psychological Stress,” no. Icisc, pp. 181–185, 2020.

M. Illahi, I. F. Siddiqui, Q. Ali, and F. A. Alvi, “Ensemble Machine Learning Approach for Stress Detection in Social Media Texts,” vol. 20, no. 02, pp. 123–128, 2022.

M. O. Viktorovych, S. O. Vitalyivna, M. M. Oleksyivna, Z. O. Oleksandrivna, and Y. A. Volodymyrovych, “PRACTICAL IMPLEMENTATION OF NEURAL NETWORK METHOD FOR STRESS FEATURES DETECTION BY SOCIAL INTERNET NETWORKS,” pp. 160–167.

G. Tiwari and G. Das, “MACHINE LEARNING BASED ON APPROACH FOR DETECTION OF DEPRESSION USING SOCIAL MEDIA USING,” vol. 6, no. 3, pp. 131–135, 2021, doi: 10.51397/OAIJSE03.2021.00022.

S. C. Guntuku, A. Buffone, K. Jaidka, J. C. Eichstaedt, and L. H. Ungar, “Understanding and Measuring Psychological Stress Using Social Media RQ1 : Differential Language Analysis of,” no. Icwsm, 2019.

Z. N. Vasha, B. Sharma, I. J. Esha, J. Al Nahian, and J. A. Polin, “Depression detection in social media comments data using machine learning algorithms,” vol. 12, no. 2, 2023, doi: 10.11591/eei.v12i2.4182.

X. Wang, “A Meta-learning based Stress Category Detection Framework on Social Media,” pp. 2925–2935, doi: 10.1145/3485447.3512013.

S. Smys and J. S. Raj, “Analysis of Deep Learning Techniques for Early Detection of Depression on Social Media Network - A Comparative Study,” vol. 03, no. 01, pp. 24–39, 2021.

F. Kasmin, N. Afeeqah, I. Razali, S. Sakinah, S. Ahmad, and Z. Othman, “Stress Detection Through Text in social media Using Machine Learning Techniques,” vol. 2, no. 2, pp. 162–176, 2026.

A. Fatima, Y. Li, and T. T. Hills, "DASentimental : Detecting Depression , Anxiety , and Stress in Texts via Emotional Recall , Cognitive Networks , and Machine Learning," 2021.

B. Bharadwaj, S. Nayak, and P. K. Panigrahi, "Sentiment Analysis for Identifying Depression through Social Media Texts Using Machine Learning Technique," vol. 5, no. 2, pp. 102–118, 2025.

L. May and A. Ashik, "Comprehensive Analysis of Stress Levels Using Ensemble Learning and Neural Networks for Insightful Understanding," no. May, 2024, doi: 10.1109/ICEEICT62016.2024.10534534.

S. Gawade, R. Sawant, A. Rathod, and P. C. Dhavale, "No Title," no. April, 2022.

A. Amanat *et al.*, "Deep Learning for Depression Detection from Textual Data," pp. 1–13, 2022.

L. Ansari, S. Ji, Q. Chen, and E. Cambria, "Depression Detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 1, pp. 211–219, 2023, doi: 10.1109/TCSS.2022.3154442.

M. R. Chandra, B. N. Kumar, P. N. Reddy, and K. Kishore, "TEXT BASED STRESS DETECTION USING MACHINE LEARNING," no. 02, pp. 66–70, 2023.

Y. S. Taspinar and I. Cinar, "Stress Detection with Natural Language Processing Techniques from Social Media Articles," *Proc. Int. Conf. Intell. Syst. New Appl.*, pp. 70–74, 2024, doi: 10.58190/icisna.2024.93.

J. Priyanka, M. T. Student, and R. Forest, "STRESS DETECTION USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING," vol. 8, no. 9, pp. 720–727, 2023.

K. Rahu and A. Auvinen, "Depression and Suicide Analysis Using Machine Learning and NLP Depression and Suicide Analysis Using Machine Learning and," 2022, doi: 10.1088/1742-6596/2161/1/012034.

A. Gaius, R. W. Mwangi, and A. Ngunyi, "A Stacking-Based Ensemble Approach with Embeddings from Language Models for Depression Detection from Social Media Text," pp. 420–453, 2023, doi: 10.4236/jdaip.2023.114022.

M. Z. Id, D. Cheng, X. Lu, and X. T. Id, "Postgraduate psychological stress detection from social media using BERT-Fused model," pp. 1–21, 2024, doi: 10.1371/journal.pone.0312264.

"Detecting Psychological Stress using Machine Learning over Social Media Interaction 1," no. Icces, pp. 646–649, 2020.

T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions," *J. Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00575-6.

S. M. Chaware, C. Makashir, C. Athavale, M. Athavale, and T. Baraskar, "Stress Detection Methodology based on Social Media Network : A Proposed Design," no. 3, pp. 3489–3492, 2020, doi: 10.35940/ijitee.B7537.019320.

E. Turcan and K. Mckeown, "Dreaddit : A Reddit Dataset for Stress Analysis in Social Media," no. Louhi, pp. 97–107, 2019.

H. P. N. A, P. Ly, D. Le, and T. Do, "Real-time Stress Detection on Social Network Posts using Big Data Technology".

S. Jadhav, "Text Based Stress Detection Techniques Analysis Using Social Media," pp. 7–11.

[42] Y. Yang, A. Xie, S. Kim, J. Hair, M. Al-garadi, and A. Sarker, "Automatic Detection of Twitter Users Who Express," vol. 41, no. 9, pp. 717–724, doi: 10.1097/CIN.0000000000000985.

X. Wang, "Contrastive Learning of Stress-specific Word Embedding for Social Media based Stress Detection," pp. 5137–5149, doi: 10.1145/3580305.3599795.

K. Kumari and S. Das, “Stress Detection System using Natural Language Processing and Machine Learning Techniques,” 2022.

R. Fernando, Y. D. Proboningrum, and S. D. Supriati, “NLP Implementation For AI Generated Text Detection (ChatGPT) Using Naive Bayes Method,” no. 10, pp. 292–302, 2025.

J. Cervantes, F. Garcia-lamont, L. Rodríguez-mazahua, and A. Lopez, “Neurocomputing A comprehensive survey on support vector machine classification : Applications , challenges and trends,” vol. 408, pp. 189–215, 2020, doi: 10.1016/j.neucom.2019.10.118.

S. N. Zisad, “Comparative Analysis of Transformer Models in Disaster Tweet Classification for Public Safety”.

A. V. Turukmane, L. N. Avula, and C. Gunji, “Stress Detection Using NLP and DL Models,” no. Icsice 24, 2025.