Explainable Deep Learning for Mental Health Detection from English and Arabic Social Media Posts

ABHINAV KUMAR, Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology Allahabad, India
JYOTI KUMARI, Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be University), India
JIESTH PRADHAN*, Department of Computer Science and Engineering, National Institute of Technology Jamshedpur, India

The research communities have recently begun exploring the detection of depression through social media, making it a relatively new development in this field. Some research has been done to identify depressive signs from English social media postings, while little work has been done for the Arabic posts. This paper proposes a BERT and Bi-LSTM pipeline for the identification of depressive signs from Arabic social media posts. A fine-tuned RoBERTa-based model is also proposed for English social media posts for depressive state identification. Along with the proposed model, seven conventional machine learning and eight deep learning models are also explored for the identification of depressive signs from Arabic and English social media posts. The performance of the proposed model is validated on two Arabic datasets and one English dataset. The proposed BERT and Bi-LSTM pipeline achieved state-of-the-art performance with an F_1-score of 1.00 and 0.82 for two different Arabic datasets, whereas the proposed fine-tuned RoBERTa achieved a F_1-score of 0.60 which is comparable in identifying depressive sign from English social media posts. The majority of the suggested deep learning models are end-to-end, which necessitates a greater explanation for their success. An explainable AI-based model may enhance decision-making, transparency, and interpretability. Therefore, this research identifies where the suggested system learned well and where it failed in recognition of the depression signs that can help in future developments in the field of depression detection.

Additional Key Words and Phrases: Depression, Explainable AI, Deep Learning, Social media

1 INTRODUCTION

Many people all around the world struggle with some form of mental illness. In the United States alone, a large portion of adults is diagnosed with mental illness like depression (6.7%), anorexia and bulimia nervosa (1.6%), and bipolar disorder (2.6%) each year [26, 47]. In some cases, mental illness has been linked to tragic incidents such as mass shootings, which have resulted in the loss of many innocent lives [27]. Depression is a common mental health issue that is rampant compared to other forms of mental disease worldwide [46]. A thorough psychiatric evaluation by qualified psychiatrists is typically required for the diagnosis of depression, especially in the early stages [33]. Interviews, surveys, self-reports, and feedback from friends and family are frequently used in this

*Corresponding Author

Authors' addresses: Abhinav Kumar, Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology Allahabad, , Uttar Pradesh, India, abhinavanand05@gmail.com; Jyoti Kumari, Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India, j2kumari@gmail.com; Jiesth Pradhan, Department of Computer Science and Engineering, National Institute of Technology Jamshedpur, Jamshedpur, Jharkhand, India, jiteshpradhan.cse@nitjsr.ac.in.

ACM 2375-4699/2023/11-ART
https://doi.org/10.1145/3632949

process. Unfortunately, during the early stages of their troubles, people with depression avoid consulting doctors or going to clinics [48].

People and health organizations have embraced Internet platforms to create virtual communities that allow knowledge, counsel, and support to be shared more widely and swiftly. Studies have demonstrated that many people often debate and give advice on health-related subjects on social media, in addition to expressing their feelings and taking activities [32, 36]. These platforms offer a possible method of gaining access to information on mental health, including details about diagnoses, treatments, and drugs. Many people with mental health disorders openly or covertly share their emotions and everyday challenges on social media because they find consolation in doing so [5]. Hence, social media presents an opportunity for automatically identifying individuals who may be experiencing depression. However, manually sorting through countless social media posts and profiles to identify depressed individuals would be time-consuming. To address this, scalable computational methods can be employed to efficiently and promptly detect individuals who may be depressed. This approach has the potential to prevent major tragedies and provide timely assistance to those who genuinely require it. Users that are depressed, frequently display distinctive behaviors on social media, creating important behavioral data that may be utilized to extract various aspects. Not all of these behaviors, nevertheless, are distinctly related to depression.

Depression detection is important for several reasons: (i) Early detection of depression allows for timely intervention and treatment. Early identification can help prevent the condition from worsening and improve the chances of successful treatment outcomes. (ii) Suicide prevention: Depression is a major risk factor for suicide, and early detection can play a crucial role in preventing suicide. By identifying individuals who exhibit signs of depression, particularly those that express suicidal thoughts or behaviors, appropriate interventions and support systems can be activated to reduce the risk of self-harm and promote mental well-being. (iii) Improved Outcomes: When depression is identified early, individuals have a better chance of achieving positive outcomes. Early intervention can help elevate symptom identification, reduce the duration and severity of depressive episodes, and improve overall mental health and functioning. (iv) Reduced Stigma: Depression detection and awareness initiatives can help reduce the stigma surrounding mental health conditions. By promoting early detection and open discussions about depression, individuals may feel more comfortable seeking help and support, leading to a more compassionate and understanding society.

Detecting depression from online social media is an emerging field of research that utilizes natural language processing (NLP) and machine learning techniques to analyze textual data and identify potential signs of depression in individuals. While it is not a definitive diagnosis, it can provide valuable insights and help identify individuals who may benefit from further assessment or support [47]. Here are some common approaches used in depression detection from online social media: (i) Linguistic analysis: NLP techniques are employed to analyze the language used in social media posts, comments, or status updates. Researchers look for patterns and linguistic markers associated with depression, such as negative emotional expressions, self-deprecation, hopelessness, and frequent use of words related to sadness, isolation, or despair. (ii) Sentiment analysis: Sentiment analysis techniques determine the overall emotional tone of social media content. Algorithms classify posts as positive, negative, or neutral based on the words and phrases used. Users who consistently exhibit negative sentiment in their posts may be flagged as potentially at risk for depression. (iii) Topic modeling: Topic modeling algorithms, such as Latent Dirichlet Allocation (LDA), identify recurring themes or topics in social media posts. Researchers look for topics related to depression, mental health issues, or personal struggles. Users who frequently discuss these topics may be more likely to be experiencing depression. (iv) Network analysis: Social network analysis examines the structure and connections within an individual's online social network. Identifying individuals who are socially isolated, have limited social interactions, or exhibit reduced social support can be indicative of depression. (v) Machine learning models: Researchers train machine learning models on labeled datasets, where

posts or users are classified as depressed or non-depressed. These models learn patterns from the data and can then predict the likelihood of depression in new, unseen posts or users.

In recent years, deep learning has demonstrated successful applications in various domains, including fake news detection [22, 41], disaster management [19, 21, 23], hate and offensive language identification [6, 18, 35], etc. Specifically, deep learning has shown promising results in detecting depression from social media, surpassing the performance of traditional machine learning methods. Po'swiata and Perełkiewicz [31] developed a DepRoBERTa model, and Singh and Motlicek [42] proposed an ensemble of fine-tuned BERT, RoBERTa, and XLNet models for detecting depression in English language tweets, categorizing them as "not depressed", "moderately depressed" or "severely depressed". Similarly, Orabi et al. [30] experimented with Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) neural network for depression detection from English tweets. Irrespective of English language social media posts, Almars et al.[3] have proposed a Bi-LSTM-based model to classify Arabic tweets into depressed and not-depressed. Hassib et al. [12] experimented with several transformer-based models to classify Arabic tweets into three depression categories such as "depressed", "suicidal" and "neutral". Although deep learning methods have been shown to be helpful for depression identification, many presented deep learning algorithms lack explainability in their predictions, which might weaken faith in the models. To address this, an explainable deep learning-based model for depression sign detection from social media has been developed. This model can provide insights into how the deep learning model functions as well as opportunities for improvement, allowing for a better understanding of the decision-making process. The overall contribution of the paper can be summarized as follows:

- Introducing an interpretable model combining BERT + Bi-LSTM for detecting signs of depression in Arabic tweets and RoBERTa for English tweets.
- Evaluating the effectiveness of the suggested model by comparing its performance against current state-of-the-art models.
- Analyzing the predictions of the proposed model through a color-coded heatmap applied to the tweets. This visualization helps in identifying which parts of the text is emphasized by the model during prediction, as well as areas where it struggles to discern the accurate context.

The remainder of the study is organized as follows: Section 2 discusses the literature on depression detection, Section 3 discusses the suggested BERT and Bi-LSTM model. Results of the proposed framework are reported in Section 4, and the study is concluded with some future directions in Section 5.

2 RELATED WORKS

The online depression detection techniques can be broadly classified into two categories: (i) Manual feature extraction with Traditional Machine Learning (TML), and (ii) Automatic feature extraction using Deep Learning (DL) [45]. The former involves the process of manually analyzing and extracting elements from user-generated content like tweets, while the latter combines the user's social behavior with multimedia information. Further, there are some studies that combine the DL and TML techniques to enhance model performance.

2.1 Traditional Machine Learning-based Approaches

Initially, the features like sentiment analysis, word frequency, linguistic style, topic modeling, pitch, energy, spectral characteristics, etc., were manually extracted and thereafter the traditional ML techniques like SVM, Random Forest, Naive Bayes, etc. were used to detect the depression. Choudhury et al. [7] have examined various behavioral attributes like linguistic styles, language, emotion, social engagement, etc., of Twitter users diagnosed with clinical depression. They found that social signals like increased negative affect, highly clustered egonetworks, decrease in social involvement, heightened relational and medicinal concerns, etc., may characterize the onset of depression in individuals. Although, they did not achieve high performance, yet, their study contributed to

the field by providing a detailed analysis of feature engineering and modeling approaches. Further, Wang et al. [44] investigated Twitter and Weibo data by implementing a sentiment analysis approach to determine the depression inclination of each blog. The manually crafted rules using vocabulary have been used to measure the depression level in tweets. Simultaneously, their findings have emphasized text-based features in online data for depression detection. Thereafter, Deshpande et al. [9] have modeled the textual information from Twitter using natural language processing (NLP) and performed emotion analysis focusing on depression. Here, the tweet texts were represented using the Bag of Words (BOW) technique, allowing the classifier to automatically learn the hidden features. Following this, they achieved F1-Scores of 0.8329 and 0.7973 by using Naive Bayes (NB) and Support Vector Machine (SVM) respectively. Further, Shen et al. [39] have proposed an improved depression detection approach for identifying depressed users timely by harvesting social media data. So, in order to extract feature groups covering both online behaviors on social media and the clinical depression criteria, they have compiled a well-labeled dataset on depression and non-depression on Twitter. They have extracted six different depression-related feature groups including social behavior, posted pictures, and text. A multimodal depressive dictionary learning (MDL) model was proposed that learned the user's latent and sparse features. The experimental results demonstrated that the model achieved an F1-Score of 0.85 by highlighting the effectiveness of the dictionary learning strategy and the ensembling of multimodal data. Recently, TML-based research [1, 29, 43] has emerged in the field of depression detection. Notably, Mustafa et al. [29] have employed Term Frequency-Inverse Document Frequency (TF-IDF) algorithm for finding the 100 most frequently used words among depressive individuals. They used Linguistic Inquiry and Word Count (LIWC) to identify the emotions of the words. After that, weight was assigned to each word in the tweet and the authors claimed to achieve an impressive F1-Score of 0.89 based on a one-dimensional convolutional neural network (CNN-1D). This has marked the introduction of a neural network for depression detection in online social network (OSN) users. Similarly, Sharen and Rajalakshmi [38] have used XGBoost, Random Forest, and SVM to categorize the symptoms of depression out of which, the XGBoost has shown the accuracy of 61% and a macro F_1-score of 0.54.

2.2 Deep Learning-based Approaches

The DL-based approaches focus on combining the user's social behavior and multimedia information consisting of text, pictures, and videos. The modeling of textual information becomes a prominent research direction in such approaches. This alludes researchers to utilizing NLP techniques for embedding text into high-dimensional continuous vectors, enabling automatic mining of word features. Some studies have integrated manually extracted features as an input to the DNN or have combined the traditional and DNN classifiers to enhance the performance. Such approaches have performed well in various social network analysis tasks, including depression detection [14]. Regarding this, Orabi et al. [30] have evaluated several DNN classifiers known for their significant performance in NLP classification tasks by employing pre-trained Word2Vec model [28]. The experimental results showed that the CNN-1D architecture with a max-pooling structure yielded the highest result. Additionally, CNN-based models have outperformed the depression detection tasks in comparison to other recurrent structures and the LSTM neural network [13]. Sadeque et al. [34] introduced a Gated Recurrent Units (GRU) based sequential classifier and employed a weighted F1 metric in due course. Further, they adopted a "post-by-post" strategy by considering the tweet's text as individual documents and providing it as input to the classifier asynchronously. Instead of scanning a huge number of tweets from depressed users (e.g., users with 200 tweets recording their anti-depressant experience), this approach allowed the model to evaluate a user's depression inclination after each scanned tweet. Later on, Shen et al. [40] identified that using specific online social networks (OSNs) may not provide optimal global solutions across all other platforms for depression detection. To deal with this, the authors combined the DNN model with Feature Adaptive Transformation (DNN-FATC)

for the cross-domain result. This approach comprehensively considers features from various aspects and facilitates the transfer of important information across different domains. Recently, the use of DL techniques has increased greatly. Gui et al. [11] have explored the model under different proportions of normal and depressed users and concluded that the best performance is achieved for the balanced dataset. Further, they employed the reinforcement learning (RL) approach to enhance the model performance. Subsequently, Lin et al. [24] utilized the widely-used pre-trained model (BERT) [17] to embed word vectors. Thereafter, they extract the hidden layer output of BERT to fuse text and image features, enabling improved performance in downstream classification tasks. Po'swiata and Perełkiewicz [31] developed a DepRoBERTa model for detecting depression in English language tweets, categorizing them as "not depressed", "moderately depressed" or "severely depressed". They received a 0.583 macro-averaged F_1-score. Whereas, Singh and Motlicek [42] proposed an ensemble of fine-tuned BERT, RoBERTa, and XLNet models to classify tweets into "severely depressed", "moderately depressed" or "not depressed" classes. They achieved a macro F_1-score of 0.546.

The majority of works have focused on detecting depression from English posts. However, recently, some works have tried to identify depressive signs from Arabic social media posts also. Hassib et al. [12] created an Arabic dataset by annotating tweets into three depression categories namely "suicidal", "depressed", and "neutral" and experimented with different transformer-based models. They had the greatest results with MARBERT, with a F_1-score of 88.75%. Alghamdi et al. [2] gathered features from Arabic social media posts and tested them with different classifiers, including SVM, Random Forest, KNN, SGD, and Decision Tree, and found that the TF-IDF feature with SGD classifier achieved an accuracy of 0.73%. Almars et al. [3] proposed a Bi-LSTM-based model to classify Arabic tweets into depressed and not-depressed and achieved an accuracy of 0.83%.

3 METHODOLOGY

Figure 1 illustrates the comprehensive flowchart of our proposed methodology, which involves the development of multiple models for the detection of depressive indicators in both Arabic and English-language tweets. We employed various conventional machine learning techniques, including Random Forest (RF), K-Nearest Neighbour (KNN), Naive Bayes (NB), Decision Tree (DT), and Gradient Boosting (GB). We harnessed unigram, bigram, and trigram Term Frequency-Inverse Document Frequency (TF-IDF) features in all the conventional machine learning models for both Arabic and English tweets in the identification of depressive signs. Along with traditional machine learning models, several deep learning models such as (i) Arabic-camelBERT + Bi-LSTM (AC-BERT + Bi-LSTM), (ii) Arabic-camelBERT + LSTM (AC-BERT + LSTM), (iii) Arabic-camelBERT + GRU (AC-BERT + GRU), (iv) Arabic-camelBERT (AC-BERT) [15], and (v) BERT-base-Arabic (BERT-BA), and Multilingual BERT (mBERT) are implemented for Arabic tweets. For English tweets, several popular BERT variants such as (i) RoBERTa-base, (ii) BERT-base, (iii) RoBERTa + LSTM, and (iv) XLM-RoBERTa were implemented (as can be seen in Figure 1).

3.1 Data Description

To validate the proposed system, three sets of datasets were used. The first two datasets are in the Arabic language, and the third dataset is in the English language. This work uses the dataset published by [25] as the first dataset (D1), which contains ten depression classes of Arabic tweets. The second Arabic dataset (D2) [8] consists of depressive and not-depressive classes. To train models, the datasets were divided into the ratio of 80% and 20%. All the details of the number of samples in the training set and testing set can be seen in Table 1. The English language dataset (D3) shared in the Workshop [16] is used as the third data in this paper. This data contains a separate train and development set. Due to the unavailability of the public test set, this work uses development to validate the models. The details of data samples can be seen in Table 1. In the dataset $D1$, only one instance of *"feelings of worthlessness"* was there; therefore, it is removed from the dataset. In the preprocessing phase, we

Table 1. Data statistic used to validate proposed model

Datasets	Class	Training	Testing
Arabic (D1)[25]	Diminished ability to think or concentrate (DATC)	212	59
	Weight disorder (WD)	181	45
	Psychomotor agitation or retardation (PAR)	97	23
	Sleep disorder (SD)	93	24
	Low mood (LM)	87	22
	Loss of energy (LE)	84	17
	Suicidality (SU)	80	20
	Feelings of worthlessness (FW)	82	15
	Losing interest or pleasure in activities (LIPA)	60	20
	Total	976	245
Arabic (D2)[8]	Depressive	426	103
	Not-Depressive	420	109
	Total	846	212
English (D3)[16]	Moderate Depressive	6019	2306
	Not Depressive	1971	1830
	Severe Depressive	901	360
	Total	8891	4496

eliminated special characters, excess spaces, punctuation, and URLs from the tweet. Subsequently, all the words were transformed into lowercase.

3.2 Bidirectional Encoder Representations from Transformers (BERT)

The BERT model [17] uses a multi-layered bidirectional transformer encoder architecture. The encoder generates a sequence of contextualized token representations from a sequence of input tokens. The model's input is made up of a set of n input tokens, each of which is denoted by the formula $X = x_1, x_2, ..., x_n$, where x_i is an input token that has been embedded into a d-dimensional vector by means of an embedding matrix. The input tokens are subsequently sent via a series of transformer blocks with various levels. Each transformer block consists of two sub-layers: a position-wise feed-forward network and a self-attention mechanism. The feed-forward neural network employs non-linearity to further change the representations once the self-attention mechanism has recorded the relationships between the contexts of the words. A weighted sum of the input tokens is calculated by the self-attention mechanism, with the weights depending on how similar the tokens are. The following formula are used to calculate attention using equation 1:

$$\text{Attention } (\alpha, \beta, \gamma) = \text{softmax}(\frac{\alpha \times \beta^{\top}}{\sqrt{d_k}})\gamma \tag{1}$$

where α, β, and γ represent the query, key, and value embeddings for the input tokens, respectively, and d_k is the dimension of the key vectors [17]. The output of the self-attention layer is then processed by a position-wise feed-forward network, which applies a linear transformation and a non-linear activation function to each point in the sequence individually: $FFN(x) = max(0, x \times weight_1 + bias_1) \times weight_2 + bias_2$, where $weight_1$, $weight_2$, $bias_1$, and $bias_2$ are network learnable parameters. The position-wise feedforward network's input and output are added at the transformer block, and the resulting output is fed into the following transformer block in the chain. The last transformer block of the BERT model generates a series of contextualized token representations as its

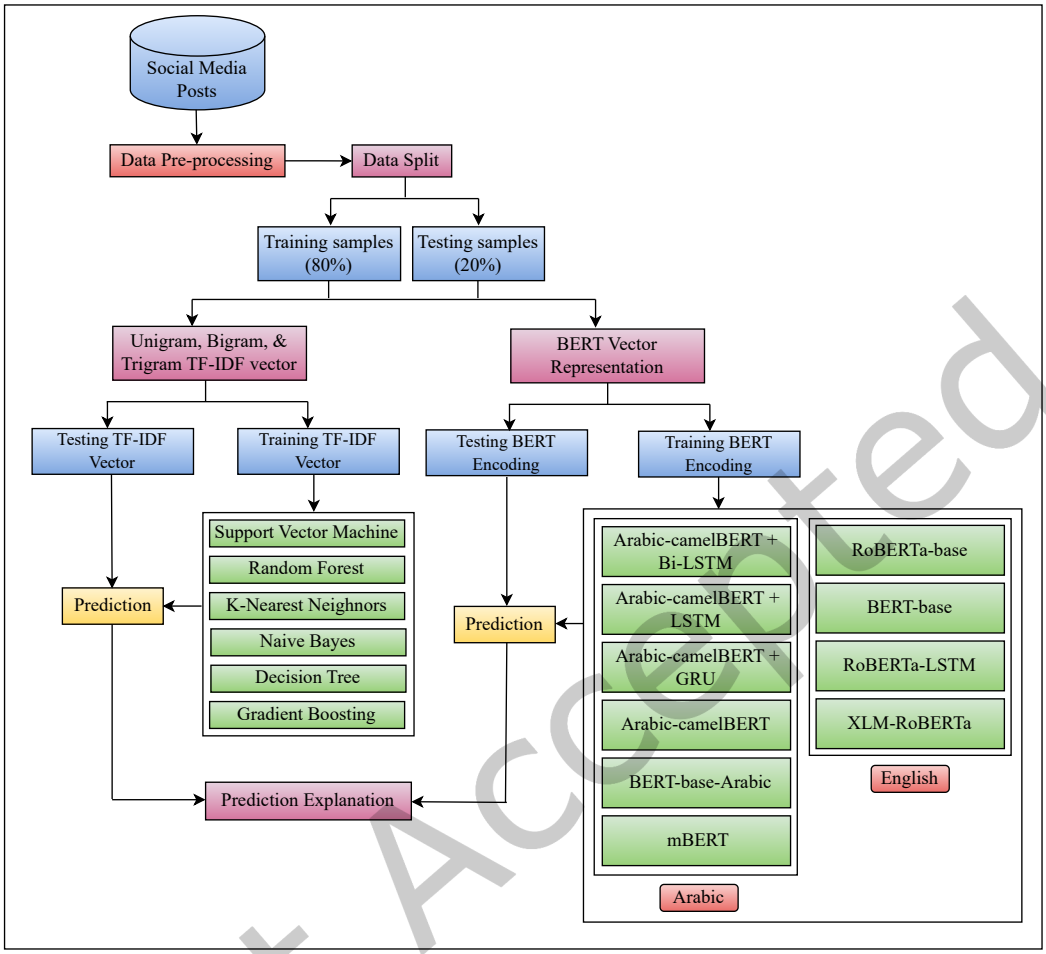

Fig. 1. Flow diagram of the proposed model for depression sign identification

output. Then the final encoded text vector is used by other deep learning models such as LSTM, and Bi-LSTM. To further understand how the encoded vector is subsequently processed for prediction, the following subsections provide an extensive overview of LSTM and Bi-LSTM networks.

3.3 Long-Short-Term-Memory (LSTM)

In order to manage sequence data, recurrent neural network topologies like the LSTM [13, 20] discover long-term associations. It addresses the vanishing gradient problem that plagues traditional RNNs by incorporating a memory cell and three gating mechanisms. The LSTM model consists of the following gates:

$$\text{Input gate: } i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{2}$$

where, i_t represents the input gate activation at time step t, x_t is the input at time step t, h_{t-1} is the hidden state from the previous time step $t-1$, c_{t-1} is the cell state from the previous time step $t-1$, W_{xi}, W_{hi}, and W_{ci} are

weight matrices, b_i is the bias term, σ is the sigmoid activation function.

$$\text{Forget gate: } f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{3}$$

where f_t represents the forget gate activation at time step t, W_{xf}, W_{hf}, and W_{cf} are weight matrices, b_f is the bias term.

$$\text{Cell state update: } g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \tag{4}$$

where g_t represents the candidate cell state at time step t, W_{xg} and W_{hg} are weight matrices, b_g is the bias term, tanh is the hyperbolic tangent activation function.

$$\text{Cell state: } c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

where c_t represents the updated cell state at time step t, $\odot$ denotes element-wise multiplication.

$$\text{Output gate: } o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{6}$$

where, o_t represents the output gate activation at time step t, W_{xo}, W_{ho}, and W_{co} are weight matrices, b_o is the bias term,

$$\text{Hidden state: } h_t = o_t \odot \tanh(c_t) \tag{7}$$

where, h_t represents the hidden output at time step t.

To summarise, the LSTM model updates the cell state based on the input, previous hidden state, and previous cell state using the input gate, forget gate, and candidate cell state. The output gate selects the elements of the cell state to be used in the calculation of the concealed state. To produce the final hidden state and output, the output gate is activated and the hyperbolic tangent function is then applied to the cell state. This process is performed for every step in the time series. Finally, the output of the LSTM hidden state is passed through the softmax layer to get probabilities for the output classes.

3.4 Bi-directional LSTM

The BiLSTM model [37], an advancement over the LSTM model, processes the input sequence both forward and backward to account for information from both the current and the future contexts. The system is composed of two separate LSTM layers, one used for forward processing and the other for backward processing. The output is created in its final form by concatenating the concealed states of both directions. The equations for the BiLSTM model can be described as follows:

- Forward LSTM equations:

$$\text{Input gate: } i_{ft} = \sigma(W_{xi}x_t + W_{hi}h_{(t-1)f} + W_{ci}c_{(t-1)f} + b_i) \tag{8}$$

$$\text{Forget gate: } f_{ft} = \sigma(W_{xf}x_t + W_{hf}h_{(t-1)f} + W_{cf}c_{(t-1)f} + b_f) \tag{9}$$

$$\text{Candidate cell state: } g_{ft} = \tanh(W_{xg}x_t + W_{hg}h_{(t-1)f} + b_g) \tag{10}$$

$$\text{Cell state: } c_{ft} = f_{ft} \odot c_{(t-1)f} + i_{ft} \odot g_{ft} \tag{11}$$

$$\text{Output gate: } o_{ft} = \sigma(W_{xo}x_t + W_{ho}h_{(t-1)f} + W_{co}c_{ft} + b_o) \tag{12}$$

$$\text{Hidden state/output: } h_{(t)f} = o_{ft} \odot \tanh(c_{ft}) \tag{13}$$

- Backward LSTM equations:

$$\text{Input gate: } i_{bt} = \sigma(W_{xi}x_t + W_{hi}h_{(t+1)b} + W_{ci}c_{(t+1)b} + b_i) \tag{14}$$

$$\text{Forget gate: } f_{bt} = \sigma(W_{xf}x_t + W_{hf}h_{(t+1)b} + W_{cf}c_{(t+1)b} + b_f) \tag{15}$$

$$\text{Candidate cell state: } g_{bt} = \tanh(W_{xg}x_t + W_{hg}h_{(t+1)b} + b_g) \tag{16}$$

$$\text{Cell state: } c_{bt} = f_{bt} \odot c_{(t+1)b} + i_{bt} \odot g_{bt} \tag{17}$$

$$\text{Output gate: } o_{bt} = \sigma(W_{xo}x_t + W_{ho}h_{(t+1)b} + W_{co}c_{bt} + b_o) \tag{18}$$

$$\text{Hidden state/output: } h_{(t)b} = o_{bt} \odot \tanh(c_{bt}) \tag{19}$$

- Final output:

$$\text{Final output: } h_t = [h_{(t)f}; h_{(t)b}] \tag{20}$$

where $h_{(t)f}$ represents the forward hidden state/output at time step t, $h_{(t)b}$ represents the backward hidden state/output at time step t, and ; denotes concatenation.

Each of the deep learning models underwent training for 100 epochs, utilizing a learning rate of $1 \times e^{-5}$ and a batch size of 32. In both the LSTM and Bi-LSTM layers, we employed 100-dimensional output hidden vectors. Subsequently, this output hidden vector was processed through the softmax layer to compute the probabilities associated with the output classes.

4 RESULTS

To evaluate the performance of the proposed model, Precision (P), Recall (R), F_1-score (F_1), confusion matrix, and AUC-ROC curve are used. The results of different conventional machine learning models for datasets *D1 (Arabic)*, *D2 (Arabic)*, and *D3 (English)* are listed in Tables 2, 3, and 4, respectively. As it can be seen in Tables 2, 3, and 4, Random Forest (RF) classifier performed best among all the implemented conventional machine learning classifiers. For dataset $D1$, the best-performed conventional machine learning classifier, Random Forest, achieved a precision, recall, and F_1-score of 0.98. The confusion matrix and ROC curve of the best-performed Random Forest classifier for dataset $D1$ can be seen in Figures 2 and 3, respectively. Similarly, for dataset $D2$, the Random Forest classifier performed best among all the implemented conventional machine learning classifiers by achieving a precision of 0.84, recall of 0.83, and F_1-score of 0.82. The confusion matrix and ROC curve for the best-performed Random Forest classifiers for dataset $D2$ can be seen in Figures 4 and 5, respectively. In the case of the English dataset (D3), the best-performed Random Forest classifier achieved a precision of 0.62, recall of 0.59, and F_1-score 0.52. The confusion matrix and ROC curve for the best-performed Random Forest for dataset $D3$ can be seen in Figures 6 and 7, respectively.

Next, the experimentation was performed with a number of deep learning models for all three datasets *D1*, *D2*, and *D3*. The results of the different deep learning models for *D1 (Arabic)* and *D2 (Arabic)* can be seen in Table 5 and Table 6, respectively. For dataset *D1*, proposed AC-BERT + BiLSTM (Arabic-camelBERT + Bi-LSTM) achieved the best performance among all the implemented conventional machine learning and deep learning models with precision, recall, and F_1-score of 1.00. The confusion matrix and ROC curve for the proposed AC-BERT + Bi-LSTM model for dataset $D1$ can be seen in Figures 8 and 9, respectively. Similarly, for the Arabic dataset $D2$, the proposed AC-BERT + Bi-LSTM model achieved the performance among all the implemented conventional machine learning and deep learning models. The proposed AC-BERT + Bi-LSTM achieved a precision, recall, and F_1-score of 0.82 for dataset $D2$. The confusion matrix and ROC curve for the proposed AC-BERT + Bi-LSTM model can be seen in Figures 12 and 13 respectively.

At last, experimentation was performed with several deep learning models for the English depressive dataset ($D3$) and the same can be seen in Table 7. As it can be observed in Table 7, fine-tuned RoBERTa model performed best among all the other implemented deep learning models with precision, recall, and F_1-score of 0.61, 0.61, and 0.60 respectively for identifying the depression state from English tweets. The confusion matrix and ROC curve for the fine-tuned RoBERTa can be seen in Figures 12 and 13, respectively. The performance of each of

Table 2. Results of the conventional machine learning models for the Arabic depressive tweet (D1) classification

Class	RF			KNN			NB			GB			DT		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
DATC	0.98	0.97	0.97	0.95	0.92	0.93	0.76	0.95	0.84	1.00	0.97	0.98	1.00	0.97	0.98
FW	1.00	0.93	0.97	0.93	0.93	0.93	0.80	0.80	0.80	1.00	0.93	0.97	0.93	0.93	0.93
LE	0.94	1.00	0.97	0.88	0.82	0.85	1.00	0.47	0.64	0.94	1.00	0.97	1.00	0.94	0.97
LIPA	0.95	1.00	0.98	0.86	0.95	0.90	0.91	0.50	0.65	1.00	1.00	1.00	1.00	0.95	0.97
LM	0.95	0.95	0.95	0.80	0.91	0.85	0.93	0.64	0.76	0.87	0.91	0.89	0.79	1.00	0.88
PAR	1.00	1.00	1.00	0.89	0.74	0.81	0.78	0.91	0.84	1.00	1.00	1.00	1.00	1.00	1.00
SD	1.00	0.96	0.98	0.96	0.96	0.96	0.76	0.79	0.78	0.92	0.96	0.94	0.95	0.88	0.91
SU	0.95	1.00	0.98	0.90	0.90	0.90	0.88	0.75	0.81	1.00	1.00	1.00	1.00	1.00	1.00
WD	1.00	1.00	1.00	0.96	1.00	0.98	0.85	1.00	0.92	1.00	1.00	1.00	1.00	1.00	1.00
Macro Avg.	0.98	0.98	0.98	0.90	0.90	0.90	0.85	0.76	0.78	0.97	0.97	0.97	0.96	0.96	0.96
Weighted Avg.	**0.98**	**0.98**	**0.98**	0.92	0.91	0.91	0.83	0.82	0.81	**0.98**	**0.98**	**0.98**	0.97	0.97	0.97

Table 3. Results of the conventional machine learning models for the Arabic depressive tweet (D2) classification

Class	RF			KNN			NB			GB			DT		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Depressive	0.92	0.70	0.80	0.74	0.79	0.76	0.59	0.77	0.66	0.89	0.69	0.78	0.76	0.76	0.76
Not-Depressive	0.77	0.94	0.85	0.79	0.74	0.76	0.69	0.49	0.57	0.76	0.92	0.83	0.77	0.77	0.77
Macro Avg.	0.85	0.82	0.82	0.76	0.76	0.76	0.64	0.63	0.62	0.82	0.80	0.80	0.76	0.76	0.76
Weighted Avg.	**0.84**	**0.83**	**0.82**	0.77	0.76	0.76	0.64	0.62	0.62	0.82	0.81	0.80	0.76	0.76	0.76

Table 4. Results of the conventional machine learning models for the English depressive tweet (D3) classification

Class	RF			KNN			NB			GB			DT		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Moderate Depressive	0.56	0.93	0.70	0.14	0.00	0.00	0.56	0.91	0.69	0.54	0.92	0.68	0.55	0.54	0.54
Not Depressive	0.76	0.26	0.39	0.41	1.00	0.58	0.75	0.16	0.27	0.70	0.18	0.28	0.50	0.44	0.47
Severe Depressive	0.36	0.06	0.10	0.00	0.00	0.00	0.10	0.09	0.09	0.36	0.15	0.21	0.20	0.33	0.25
Macro Avg.	0.56	0.42	0.40	0.18	0.33	0.19	0.47	0.39	0.35	0.53	0.41	0.39	0.42	0.44	0.42
Weighted Avg.	**0.62**	**0.59**	**0.52**	0.24	0.41	0.24	0.60	0.54	0.47	0.59	0.55	0.48	0.50	0.48	0.49

the implemented conventional machine learning and deep learning models in terms of F_1-scores for datasets $D1$, $D2$, and $D3$ can be seen in the Figures 14, 15, and 16, respectively.

Several examples of tweets with color heat maps on each word are included to explain how the proposed models behave when making decisions. These examples illustrate which word models are given greater importance during decision-making and which word-restricted models in making correct decision-making. The Arabic tweet depicted in Figure 17 originally belongs to *DATC* class, and the proposed model has also predicted in the *DATC* class. Color heat-map on the word having a darker shade has more weight in deciding the class value. A similar kind of observation can be seen for Figure 18 also, where, the proposed model predicted it in the correct *WD* class. The proposed AC-BERT + Bi-LSTM prediction for Arabic dataset $D2$, is represented in Figures 19 and 20. In both examples, Figures 19 and 20 predicted the correct class of depression and color heat-map, showing

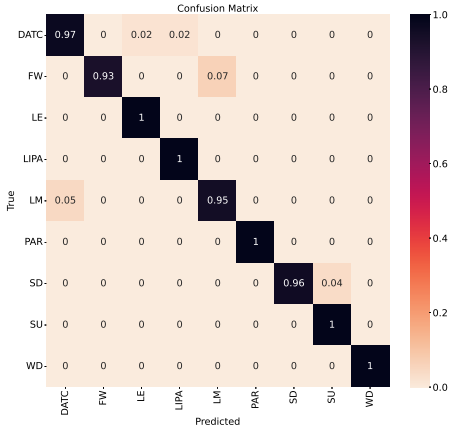

Fig. 2. Confusion matrix for the Random Forest classifier for dataset $D1$

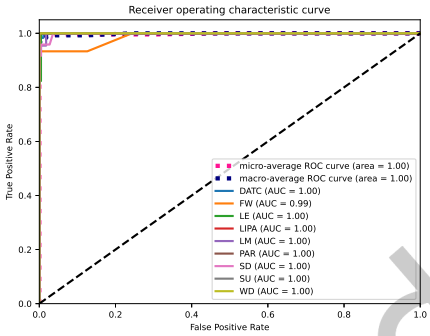

Fig. 3. ROC curve for the Random Forest classifier for dataset $D1$

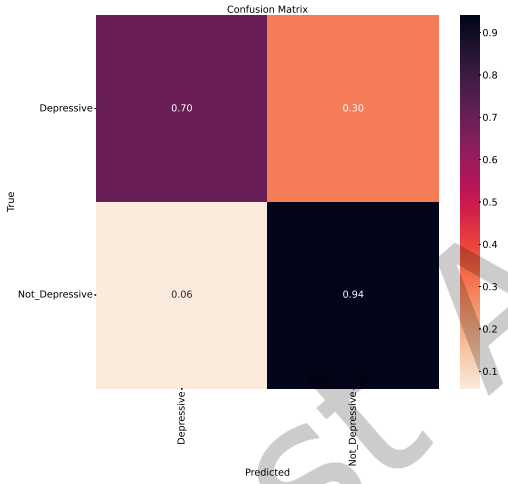

Fig. 4. Confusion matrix for the Random Forest classifier for dataset $D2$

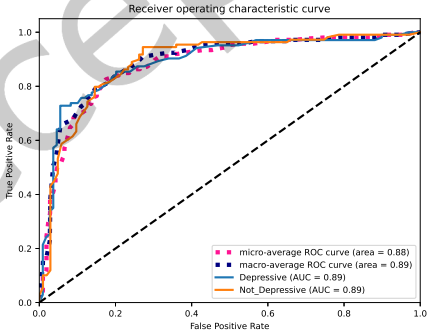

Fig. 5. ROC curve for the Random Forest classifier for dataset $D2$

the contribution of each word in the decision-making. Similarly, the prediction of fine-tuned RoBERTa model for the English tweet of dataset $D3$ is shown in Figures 21 and 22. The tweet sample shown in Figure 21 is of the originally severe depressive class, and the fine-tuned RoBERTa also predicted it in the severe depressive class. Tweet shown in Figure 22 also originally belonged to the severe depressive class, but the fine-tuned RoBERTa predicted it as a moderate depressive class. As the length of text shown in the Figure 22 is more compared to the text shown in Figure 21, the model failed to identify it in the correct class. Therefore, these observations can be taken into consideration for future developments in the identification of the depressive behavior of social media users. The comparison of the proposed models with the baseline and existing work is listed in Table 8. As it can be seen in Table 8, the proposed model performs well in comparison to baseline and existing works. The

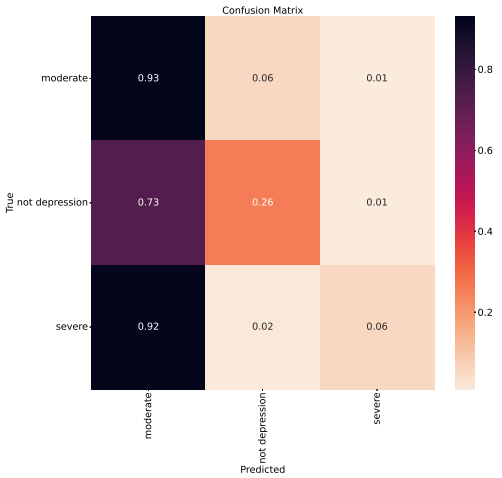

Fig. 6. Confusion matrix for the Random Forest classifier for dataset $D3$

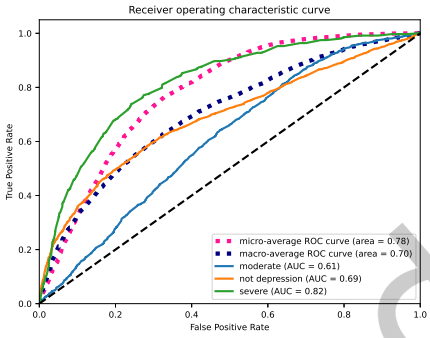

Fig. 7. ROC curve for the Random Forest classifier for dataset $D3$

Table 5. Results of the deep learning models for the Arabic depressive tweet (D1) classification

Class	Arabic-camelBERT + Bi-LSTM			Arabic-camelBERT + LSTM			Arabic-camelBERT + GRU			Arabic-camelBERT			Bert-Base-Arabic			mBERT		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
DATC	0.98	1.00	0.99	0.95	0.98	0.97	0.89	0.98	0.94	1.00	0.98	0.99	1.00	1.00	1.00	0.98	0.92	0.95
FW	1.00	1.00	1.00	1.00	0.87	0.93	1.00	0.93	0.97	1.00	0.93	0.97	1.00	1.00	1.00	0.93	0.93	0.93
LE	1.00	1.00	1.00	1.00	0.94	0.97	1.00	0.94	0.97	0.94	1.00	0.97	1.00	1.00	1.00	1.00	0.88	0.94
LIPA	1.00	0.95	0.97	1.00	0.95	0.97	1.00	0.90	0.95	0.95	1.00	0.98	1.00	0.95	0.97	0.95	0.95	0.95
LM	1.00	1.00	1.00	1.00	0.91	0.95	1.00	0.95	0.98	1.00	1.00	1.00	0.96	1.00	0.98	0.84	0.95	0.89
PAR	1.00	1.00	1.00	0.88	1.00	0.94	1.00	0.96	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SD	1.00	1.00	1.00	0.92	0.92	0.92	0.92	1.00	0.96	1.00	1.00	1.00	0.96	0.96	0.96	0.88	0.92	0.90
SU	1.00	1.00	1.00	0.95	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
WD	1.00	1.00	1.00	0.98	0.98	0.98	0.95	0.91	0.93	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.98
Macro Avg.	1.00	0.99	1.00	0.96	0.95	0.96	0.97	0.95	0.96	0.99	0.99	0.99	0.99	0.99	0.99	0.95	0.95	0.95
Weighted Avg.	**1.00**	**1.00**	**1.00**	0.96	0.96	0.96	0.96	0.96	0.96	0.99	0.99	0.99	0.99	0.99	0.99	0.95	0.95	0.95

Table 6. Results of the deep learning models for the Arabic depressive tweet (D2) classification

Class	Arabic-camelBERT + Bi-LSTM			Arabic-camelBERT + LSTM			Arabic-camelBERT + GRU			Arabic-camelBERT			Bert-Base-Arabic			mBERT		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Depressive	0.81	0.83	0.82	0.83	0.74	0.78	0.83	0.72	0.77	0.75	0.85	0.80	0.76	0.86	0.81	0.82	0.77	0.79
Not-Depressive	0.83	0.82	0.82	0.78	0.85	0.81	0.76	0.86	0.81	0.84	0.72	0.78	0.85	0.74	0.79	0.79	0.84	0.82
Macro Avg.	0.82	0.82	0.82	0.80	0.80	0.80	0.80	0.79	0.79	0.79	0.79	0.79	0.81	0.80	0.80	0.81	0.81	0.81
Weighted Avg.	**0.82**	**0.82**	**0.82**	0.80	0.80	0.80	0.80	0.79	0.79	0.79	0.79	0.79	0.81	0.80	0.80	0.81	0.81	0.81

addition of BiLSTM with the BERT model outperformed BERT by effectively capturing sequential dependencies and enhancing contextual understanding. While BERT excels at contextual word embedding, it processes text in parallel, potentially missing subtle sequential nuances. The BiLSTM, with its sequential processing capability, is

Table 7. Result for deep learning models for English depression tweet (D3) identification

Class	RoBERTa-Base			BERT-Base			RoBERTa-LSTM			XLM-RoBERTa		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Moderate Depressive	0.62	0.73	0.67	0.59	0.80	0.68	0.56	0.88	0.69	0.59	0.77	0.67
Not Depressive	0.65	0.48	0.55	0.65	0.39	0.49	0.66	0.31	0.42	0.64	0.39	0.48
Severe Depressive	0.40	0.46	0.43	0.43	0.36	0.39	0.42	0.03	0.06	0.41	0.42	0.42
Macro Avg.	0.56	0.56	0.55	0.56	0.52	0.52	0.55	0.41	0.39	0.55	0.53	0.52
Weighted Avg.	**0.61**	**0.61**	**0.60**	0.60	0.60	0.58	0.59	0.58	0.53	0.59	0.59	0.57

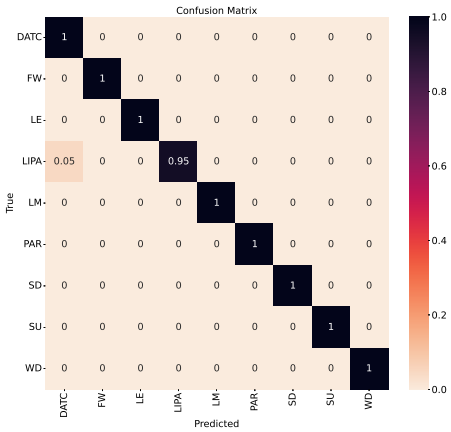

Fig. 8. Confusion matrix for the AC-BERT + BiLSTM model for dataset $D1$

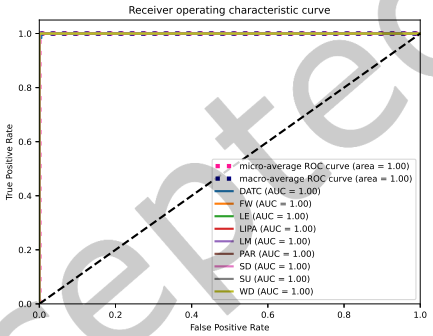

Fig. 9. ROC curve for the AC-BERT + BiLSTM model for dataset $D1$

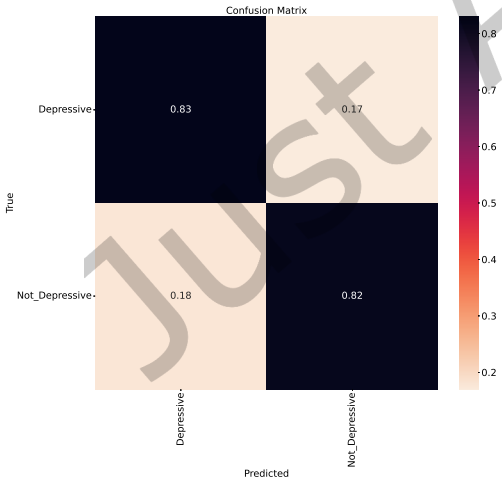

Fig. 10. Confusion matrix for the AC-BERT + BiLSTM model for dataset $D2$

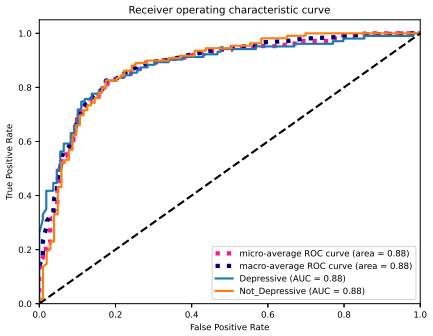

Fig. 11. ROC curve for the AC-BERT + BiLSTM model for dataset $D2$

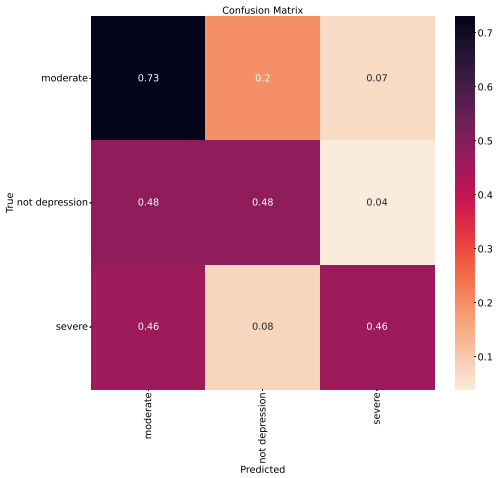

Fig. 12. Confusion matrix for the RoBERTa-base model for dataset *D*3

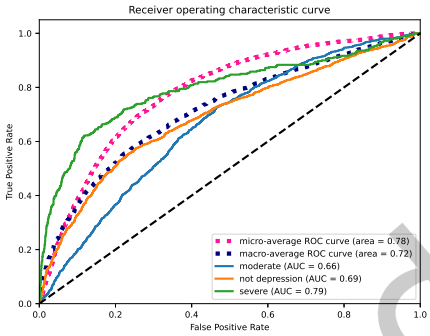

Fig. 13. ROC curve for the RoBERTa-base model for dataset *D*3

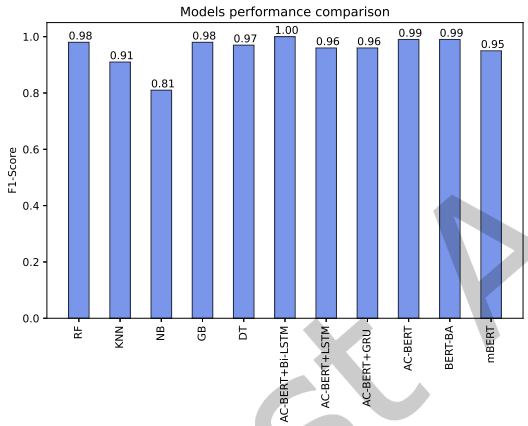

Fig. 14. Models performance comparison for depressive sign classification (*D*1)

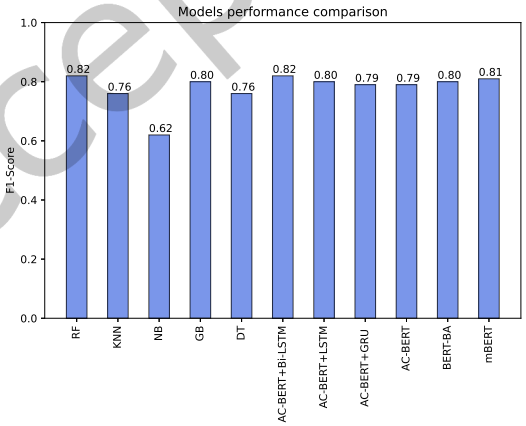

Fig. 15. Models performance comparison for depressive sign classification (*D*2)

particularly advantageous in tasks where the order of words is crucial. This combination of BERT and BiLSTM provides a richer, more nuanced representation of the tweets in identifying depressive signs from tweets.

5 CONCLUSION

Depression detection is critical because it helps those suffering from this mental health issue to obtain prompt assistance and treatment. Depression is a common and debilitating disorder that interferes with a person's thoughts, emotions, and overall well-being. It can cause a wide range of symptoms, including chronic melancholy, lack of interest in activities, changes in eating and sleep habits, difficulties focusing, and even suicidal or self-harming ideas. This paper proposed an explainable BERT and Bi-LSTM pipeline for identifying depressive

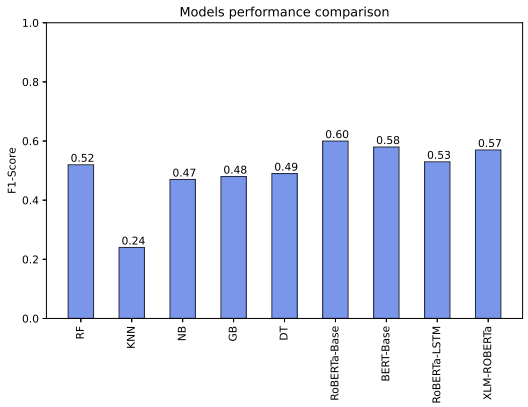

Fig. 16. Models performance comparison for depressive sign classification (*D*3)

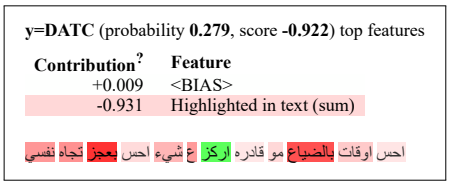

Fig. 17. Example 1: Tweets with color heat maps on each word for AC-BERT + Bi-LSTM model (D1)

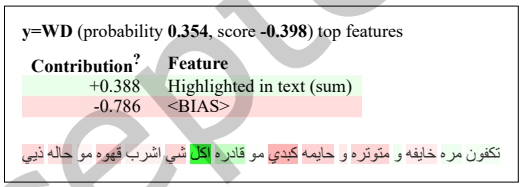

Fig. 18. Example 2: Tweets with color heat maps on each word for AC-BERT + Bi-LSTM model (D1)

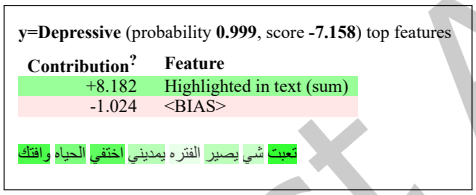

Fig. 19. Prediction explanation for the proposed AC-BERT + Bi-LSTM model (*D*2)

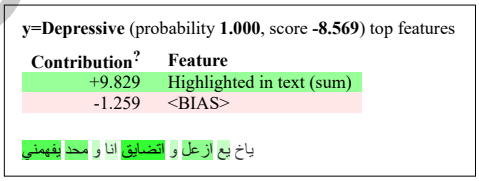

Fig. 20. Prediction explanation for the proposed AC-BERT + Bi-LSTM model (*D*2)

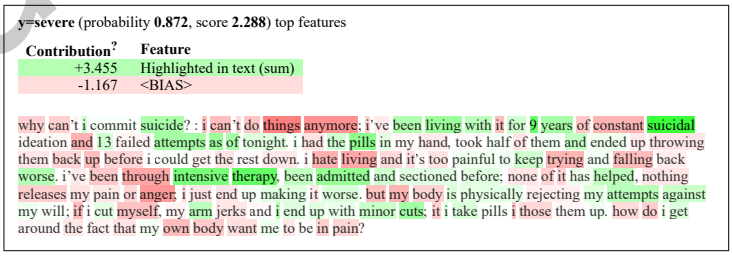

Fig. 21. Explanation of the model in the prediction for RoBERTa-base model (*D*3)

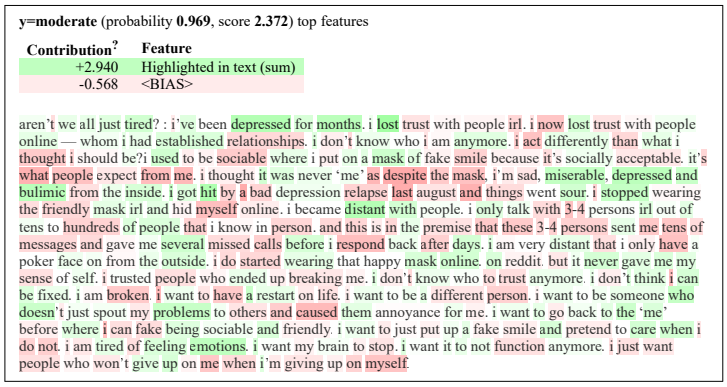

Fig. 22. Explanation of the model in the prediction for RoBERTa-base model (*D3*)

Table 8. Comparison of the proposed model with the existing models in the identification of depressive sign detection (All baseline and existing models are implemented on same dataset)

Dataset	Authors	Model	Features	Performance
D1	Baseline	Random Forest	TF-IDF	F_1-score = 0.98
	Baseline	mBER	BERT embedding	F_1-score = 0.95
	Baseline	BERT-Base-Arabic	BERT embedding	F_1-score = 0.99
	Proposed	AC-BERT + Bi-LSTM	BERT embedding	F_1-score = 1.00
D2	Baseline	mBER	BERT embedding	F_1-score = 0.81
	Baseline	BERT-Base-Arabic	BERT embedding	F_1-score = 0.80
	Almars [3]	Bi-LSTM	Word embedding	F_1-score = 0.78
	Proposed	AC-BERT + Bi-LSTM	BERT embedding	F_1-score = 0.82
D3	Dowlagar & Mamidi.[10]	SVM	SMOTE and Under Sampling	F_1-score = 0.60
	Anantharaman, et al. [4]	Fine-tuned BERT	BERT embedding	F_1-score = 0.58
	Proposed	Fine-tuned RoBERTa	BERT embedding	F_1-score = 0.60

signs from Arabic social media posts and achieved state-of-the-art performance with F_1-scores of 1.00 and 0.82 for two different Arabic datasets. Similarly, an explainable fine-tuned RoBERTa is developed for English social media posts that achieved a comparable F_1-score of 0.60. Identifying depressive behavior from social media posts is challenging as social media posts have several grammatical mistakes, non-standard abbreviations, emoticons, and code-mixed text. Therefore, more robust feature engineering can be done in future, which can be embedded with deep learning models for better performance. In future, a sequence of conversations can also be considered for better predicting the depressive behavior of the people.

DECLARATION OF COMPETING INTEREST

There is no Conflict of Interest.

REFERENCES

[1] Nafiz Al Asad, Md Appel Mahmud Pranto, Sadia Afreen, and Md Maynul Islam. 2019. Depression detection by analyzing social media posts of user. In *2019 IEEE international conference on signal processing, information, communication & systems (SPICSCON)*. IEEE, 13–17.

[2] Norah Saleh Alghamdi, Hanan A Hosni Mahmoud, Ajith Abraham, Samar Awadh Alanazi, and Laura García-Hernández. 2020. Predicting depression symptoms in an Arabic psychological forum. *IEEE Access* 8 (2020), 57317–57334.

[3] Abdulqader M Almars. 2022. Attention-based Bi-LSTM model for Arabic depression classification. *CMC-COMPUTERS MATERIALS & CONTINUA* 71, 2 (2022), 3091–106.

[4] Karun Anantharaman, S Angel, Rajalakshmi Sivanaiah, Saritha Madhavan, and Sakaya Milton Rajendram. 2022. SSN_MLRG1@ LT-EDI-ACL2022: Multi-Class Classification using BERT models for Detecting Depression Signs from Social Media Text. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 296–300.

[5] Krishna C Bathina, Marijn Ten Thij, Lorenzo Lorenzo-Luaces, Lauren A Rutter, and Johan Bollen. 2021. Individuals with depression express more distorted thinking on social media. *Nature human behaviour* 5, 4 (2021), 458–466.

[6] Shankar Biradar, Sunil Saumya, Abhinav Kumar, and Ashish Singh. 2022. Pradvis Vac: A Socio-Demographic Dataset for Determining the Level of Hatred Severity in a Low-Resource Hinglish Language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* (dec 2022). https://doi.org/10.1145/3573199

[7] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, Vol. 7. 128–137.

[8] Depression-detector. 2023. . https://github.com/alhuri/Depression-detector/tree/main/depressionDetector/data [Online; accessed 16-October-2023].

[9] Mandar Deshpande and Vignesh Rao. 2017. Depression detection using emotion artificial intelligence. In *2017 international conference on intelligent sustainable systems (iciss)*. IEEE, 858–862.

[10] Suman Dowlagar and Radhika Mamidi. 2022. DepressionOne@ LT-EDI-ACL2022: Using Machine Learning with SMOTE and Random UnderSampling to Detect Signs of Depression on Social Media Text.. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 301–305.

[11] Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 110–117.

[12] Mariam Hassib, Nancy Hossam, Jolie Sameh, and Marwan Torki. 2022. AraDepSu: Detecting Depression and Suicidal Ideation in Arabic Tweets Using Transformers. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*. 302–311.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Feiran Huang, Xiaoming Zhang, Jie Xu, Zhonghua Zhao, and Zhoujun Li. 2019. Multimodal learning of social image representation by exploiting social relations. *IEEE transactions on cybernetics* 51, 3 (2019), 1506–1518.

[15] Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. 92–104.

[16] S Kayalvizhi, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, et al. 2022. Findings of the Shared Task on Detecting Signs of Depression from Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 331–338.

[17] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

[18] Abhinav Kumar, Sunil Saumya, and Ashish Singh. 2023. Detecting Dravidian Offensive Posts in MIoT: A Hybrid Deep Learning Framework. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* (apr 2023). https://doi.org/10.1145/3592602

[19] Abhinav Kumar and Jyoti Prakash Singh. 2022. Deep Neural Networks for Location Reference Identification From Bilingual Disaster-Related Tweets. *IEEE Transactions on Computational Social Systems* (2022), 1–12. https://doi.org/10.1109/TCSS.2022.3213702

[20] Abhinav Kumar, Jyoti Prakash Singh, Yogesh K Dwivedi, and Nripendra P Rana. 2022. A deep multi-modal neural network for informative Twitter content classification during emergencies. *Annals of Operations Research* 319 (2022), 791–822. https://doi.org/10.1007/s10479-020-03514-x

[21] Abhinav Kumar, Jyoti Prakash Singh, Nripendra P Rana, and Yogesh K Dwivedi. 2023. Multi-Channel Convolutional Neural Network for the Identification of Eyewitness Tweets of Disaster. *Information Systems Frontiers* 25, 4 (2023), 1589–1604.

[22] Abhinav Kumar, Jyoti Prakash Singh, and Amit Kumar Singh. 2022. COVID-19 Fake News Detection Using Ensemble-Based Deep Learning Model. *IT Professional* 24, 2 (2022), 32–37.

[23] Abhinav Kumar, Jyoti Prakash Singh, and Amit Kumar Singh. 2022. Randomized Convolutional Neural Network Architecture for Eyewitness Tweet Identification During Disaster. *Journal of Grid Computing* 20, 3 (2022), 20.

[24] Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. Sensemood: depression detection on social media. In *Proceedings of the 2020 international conference on multimedia retrieval*. 407–411.

[25] Ashwag Maghraby and Hosnia Ali. 2022. Modern Standard Arabic mood changing and depression dataset. *Data in Brief* 41 (2022), 107999.

[26] Kathleen Ries Merikangas, Jian-ping He, Marcy Burstein, Sonja A Swanson, Shelli Avenevoli, Lihong Cui, Corina Benjet, Katholiki Georgiades, and Joel Swendsen. 2010. Lifetime prevalence of mental disorders in US adolescents: results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry* 49, 10 (2010), 980–989.

[27] Jonathan M Metzl and Kenneth T MacLeish. 2015. Mental illness, mass shootings, and the politics of American firearms. *American journal of public health* 105, 2 (2015), 240–249.

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013), 1–9.

[29] Raza Ul Mustafa, Noman Ashraf, Fahad Shabbir Ahmed, Javed Ferzund, Basit Shahzad, and Alexander Gelbukh. 2020. A multiclass depression detection in social media based on sentiment analysis. In *17th International Conference on Information Technology–New Generations (ITNG 2020)*. Springer, 659–662.

[30] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*. 88–97.

[31] Rafał Poświata and Michał Perełkiewicz. 2022. OPI@ LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text using RoBERTa Pre-trained Language Models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 276–282.

[32] Kyle W Prier, Matthew S Smith, Christophe Giraud-Carrier, and Carl L Hanson. 2011. Identifying health-related topics on Twitter: An exploration of tobacco-related tweets as a test topic. In *Social Computing, Behavioral-Cultural Modeling and Prediction: 4th International Conference, SBP 2011, College Park, MD, USA, March 29-31, 2011*. Springer, 18–25.

[33] Esteban Andrés Ríssola, Mohammad Aliannejadi, and Fabio Crestani. 2020. Beyond modelling: Understanding mental disorders in online social media. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42*. Springer, 296–310.

[34] Farig Sadeque, Dongfang Xu, and Steven Bethard. 2018. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 495–503.

[35] Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2021. Offensive language identification in Dravidian code mixed social media text. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*. 36–45.

[36] Daniel Scanfeld, Vanessa Scanfeld, and Elaine L Larson. 2010. Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control* 38, 3 (2010), 182–188.

[37] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[38] Herbert Sharen and Ratnavel Rajalakshmi. 2022. DLRG@ LT-EDI-ACL2022: Detecting signs of Depression from Social Media using XGBoost Method. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 346–349.

[39] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, et al. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution.. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. 3838–3844.

[40] Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat-Seng Chua, and Wendy Hall. 2018. Cross-domain depression detection via harvesting social media. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 1611–1617.

[41] Jyoti Prakash Singh, Abhinav Kumar, Nripendra P Rana, and Yogesh K Dwivedi. 2020. Attention-based LSTM network for rumor veracity estimation of tweets. *Information Systems Frontiers* 24 (2020), 459–474. https://doi.org/10.1007/s10796-020-10040-5

[42] Muskaan Singh and Petr Motlicek. 2022. IDIAP Submission@ LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 362–368.

[43] Maxim Stankevich, Vadim Isakov, Dmitry Devyatkin, and Ivan V Smirnov. 2018. Feature Engineering for Depression Detection in Social Media.. In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods*. 426–431.

[44] Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013*. Springer, 201–213.

[45] Yiding Wang, Zhenyi Wang, Chenghao Li, Yilin Zhang, and Haizhou Wang. 2022. Online social network individual depression detection using a multitask heterogenous modality fusion approach. *Information Sciences* 609 (2022), 727–749.

[46] Aqsa Zafar and Sanjay Chitnis. 2020. Survey of depression detection using social networking sites via data mining. In *2020 10th international conference on cloud computing, data science & engineering (confluence)*. IEEE, 88–93.

[47] Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web* 25, 1 (2022), 281–304.

[48] Maria Li Zou, Mandy Xiaoyang Li, and Vincent Cho. 2020. Depression and disclosure behavior via social media: A study of university students in China. *Heliyon* 6, 2 (2020), e03368.