Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Hybrid TF-IDF And Roberta Feature Fusion Framework With Bilstm-Attention For Multi-Platform Depression Detection and Severity Classification With Explainable AI

Nisha F[1]*, Radhakrishnan Vignesh[2]

[1]*, [2]Presidency School of Artificial Intelligence & Advanced Computing, Presidency University Itgalpur, Rajankunte, Yelahanka, Bengaluru – 560119, Karnataka, India.
[2] Email: vigneshr@presidencyuniversity.in
*Corresponding Author: Nisha. F ,
Emails: NISHA.20233CSE0017@presidencyuniversity.in,nisharobinr@presidencyuniversity.in
ORCID: https://orcid.org/0000-0001-9538-3048

Abstract
Depression is a major global mental health issue, and the analysis of the social-media text can be used as a scalable and non-invasive means of detecting depressive textual features. The majority of previous work usually relies on a single feature representation, focuses on binary classification, or provides inadequate evaluation on different social-media platforms. In this paper, we propose a novel and explainable hybrid approach that combines TF-IDF features and contextual RoBERTa representation followed by bidirectional LSTM with attention mechanism. This framework is capable of performing binary depression-related text classification as well as four-class severity classification using Twitter and Reddit datasets. We investigate cross-platform transfer and supervised fine-tuning in the target domain, applying SHAP and LIME for explaining the models' predictions. The results show that our framework achieves 95.55% accuracy and 0.9893 AUC-ROC for binary classification on Reddit, and 96.04% accuracy and 0.9953 AUC-ROC for severity classification. Fine-tuning on 20% of the labeled target-domain data leads to 82.90% accuracy in cross-platform setting.

Keywords: Depression detection; social media natural language processing; RoBERTa; TF-IDF feature fusion; BiLSTM; attention mechanism; explainable artificial intelligence; severity classification; cross-platform generalization.

1. Introduction

Depression is one of the most common mental disorders featuring persistent low mood, loss of interest, cognitive problems, impaired functioning, and suicidal thoughts in its most serious manifestations [1]. About 280 million people worldwide suffer from depression, which makes it a public health problem. Natural language processing (NLP) is commonly applied in detecting various mental diseases due to the emergence of numerous user-generated texts on social media that can contain the symptoms of depression through language and behavior [2,3]. However, such computational techniques should be considered only as supporting screening devices and not substitutes for clinical diagnosis. Twitter and Reddit are two popular social networks that have significant disparities in terms of post length, vocabulary usage, communication style, and ways of self-expression. Such disparities may affect the transferability of depression-detection models [4,5]. Recent developments in the area of transformer-based NLP allowed increasing the representation of contextual information in social media texts. RoBERTa features context-dependent semantic representations, while TF-IDF retains discriminative lexical and frequency information. Thus, their joint usage could provide additional contextual and statistical information for depression-related text classification [6,7].

Nevertheless, there are a number of important limitations to consider. The majority of existing researches deal mainly with binary depression detection and does not distinguish the degree of depression severity [8,9]. Generalization across platforms and domain adaptation remain largely understudied issues [4,5]. Furthermore, poor interpretability of deep learning models creates difficulties in applying them to the field of mental health [10,11]. Explainable artificial intelligence techniques like SHAP and LIME were increasingly applied recently in order to make deep learning models more transparent, however, their joint usage for both binary classification and multiclass severity classification of depression-related posts was scarcely studied [12,13]. The current study is an extension of the previous one conducted by the authors, An Efficient LSTM Based Depression Detection Model from Social Media [14]. While the previous paper focused mostly on LSTM-based depression detection, the current one extends the framework with TF-IDF–RoBERTa feature fusion, four-class severity classification, Bahdanau attention,

cross-platform evaluation, supervised target-domain fine-tuning, explainable artificial intelligence, and ablation analysis.

The key contributions of the current research can be listed as follows:

1. A feature-fusion approach that combines lexical-statistical features represented by TF-IDF and semantic features derived from the contextual RoBERTa representations.
2. The framework is applicable for both binary classification of depression-related text and multiclass severity classification with classes minimal, mild, moderate, and severe [15].
3. Experiments on cross-platform evaluation on Twitter and Reddit to evaluate the transferability of the models with supervised fine-tuning of the target-domain classifier based on 20% of the training labeled data.
4. SHAP and LIME approaches used to obtain global and local interpretations, respectively, as well as class-specific SHAP analysis to interpret feature attributions per severity class.
5. Ablation study to evaluate the impact of TF-IDF features, RoBERTa, bidirectional recurrent modeling, and attention [16].

To summarize, the proposed framework aims at improving accuracy, interpretability, severity awareness, and cross-platform adaptation for social-media–based depression-related text analysis without replacing professional medical expertise.

2. Related Work

2.1 Machine-Learning, Deep-Learning, and Explainable Approaches

The early work on depression detection from social media was mainly based on using traditional machine-learning classifiers with manually engineered text features. For example, Tadesse et al. used n-gram text features along with support vector machines and random forest classifiers to detect depressive posts on Reddit [17]. Later, Wani et al. used convolutional neural network (CNN), LSTM and Word2Vec embeddings for screening depression cases across multiple social networking sites [9]. Helmy et al. studied the impact of combinations of different types of text preprocessing and feature extraction, such as TF-IDF, with traditional machine-learning classifiers [18]. All these works emphasized the use of lexical and sequential features for depression detection, with the issues related to contextual modeling and explanation being raised as ongoing challenges. Therefore, explainability is receiving more and more attention in mental health text mining. Guo et al. used LIME to interpret deep-learning model for predictions made on online mental health forum data [11]. DepressionX showed the promise of interpretability in the task of depression-related text classification [12]. Hoque et al. used SHAP and LIME for explainable depression detection showing the advantage of combining global and local explainability [13]. However, explainability is still relatively unexplored in the context of binary classification and multi-class severity prediction of depression.

2.2 Transformer-Based, Hybrid, and Interpretable Approaches

Transformer-based language models have exhibited high efficiency in text classification with respect to the depression context. In particular, Bokolo and Liu have compared several transformer-based methods, including RoBERTa, in order to detect the depression in social media text [19]. Moreover, a RoBERTa-BiLSTM architecture was used for depression detection in social-media texts [20]. Al Asad et al. have developed a BERT-BiLSTM framework for the depression detection in English and Arabic texts with the help of interpretability [21]. At the same time, attention-based CNN-BiLSTM architectures have been utilized for recognizing depression in social media [22]. Hybrid recurrent architectures have been shown to have a potential application in depression detection tasks. Prasad and Jamadagni have designed a hybrid BiLSTM-CNN model [23] while Zeberga et al. used the combination of BiLSTM and BERT representation for mental health prediction [24]. In addition, Vandana et al. have further explored hybrid deep learning for depression detection [25]. Overall, all of these papers show that the integration of complementary representations and learning architectures can be helpful for modeling complex depression-related language.

Another option to analyze the behavior of these models is to apply interpretability methods. In particular, Lundberg and Lee have proposed SHAP to estimate feature importance for the predictions [26]. Ribeiro et al. have designed LIME to generate interpretable local explanations of classifier decisions [27]. Malhotra and Jindal have integrated SHAP and LIME in BERT-based framework to analyze depressive and suicidal behavior [28]. Several comparative transformer studies have broadened our knowledge about performance of models on different social-media platforms. Specifically, Kokane et al. have performed the evaluation of BERT, XLNet, RoBERTa, and DistilBERT based on Twitter and Reddit data [29]. Raj et al. have achieved strong results with BERT for the depression detection task based on Reddit posts [30]. Also, Qasim et al. have explored transformer-based models for depression severity assessment based on social media text [31]. Thus, it can be concluded that contextual representations work well, however, there are many approaches which use only transformer embeddings instead of statistical lexical features such as TF-IDF.

2.3 Class-Imbalance Handling and Recent Hybrid Models

The problem of class imbalance poses another significant methodology issue in the field of depression and severity-classification datasets because of the usual considerable difference in the number of samples between classes. The Synthetic Minority Over-sampling Technique (SMOTE) was developed by Chawla et al. This algorithm produces synthetic samples of the minority class based on interpolations between neighboring samples [32]. Further research was conducted for improving text classification accuracy via oversampling techniques [33], and more general studies examined methodology issues related to the analysis of social-media data in the context of mental health [34]. In addition, Taskiran et al. evaluated several oversampling methods that can be applied to imbalanced text-classification tasks and proved their effectiveness [35]. Nonetheless, oversampling methods should only be applied to the training dataset after splitting it into training, validation, and test sets to reduce information leakage and prevent overly positive evaluation of classifier performance. Finally, recent hybrid models keep exploring complementary feature representations and deep learning algorithms. Beniwal and Saraswat designed the hybrid BERT–CNN model for depression detection in social media [36]. DABLNet combined linguistic and temporal information in the context of detecting mental-health-related patterns in social-media data using a multimodal deep attention BiLSTM network [37].

2.4 Summary of Research Gaps

There are four main gaps according to the literature. First, even though TF-IDF and RoBERTa have proven valuable in terms of lexical and contextual information representation, respectively, their combination within the framework for depression detection is still rather unexplored. Second, current studies usually either tackle depression detection task or classification task at multiple levels/severity grades but not both simultaneously. Third, there are few studies on cross-platform assessment and target-domain adaptation in different social media platforms. Fourth, even though studies on explainability utilized interpretable modeling, SHAP, and LIME, there are still only few studies on their combination with class-specific explanation for multiple severity grades of depression prediction. Our current study attempts to fill those gaps using TF-IDF–RoBERTa feature combination, binary and four-class classification of depression severity, cross-platform assessment, target-domain fine-tuning, and SHAP-LIME analyses.

Table 1. Comparative Summary of Existing Depression-Detection Studies

Study	Data/Platform	Model	Task	XAI	Key Limitation
Tadesse et al. [17]	Reddit	N-gram + SVM/RF	Binary	No	Limited context modelling
Helmy et al. [18]	Social media	TF-IDF + ML	Binary	No	Handcrafted features
Wani et al. [9]	Multiple platforms	Word2Vec + CNN-LSTM	Binary	No	No domain adaptation
Amanat et al. [8]	Single dataset	LSTM	Binary	No	No cross-platform validation
Bhuiyan et al. [38]	Social media	CNN-BiLSTM + attention	Binary	Attention	No severity analysis
Kokane et al. [29]	Twitter, Reddit	BERT-family models	Binary	No	Limited feature fusion
Li et al. [39]	Depression text	RoBERTa-BiLSTM	Severity	No	No lexical-context fusion
Malhotra & Jindal [28]	Depression/suicide text	BERT	Binary	SHAP, LIME	No cross-platform severity test
Hoque et al. [13]	Educational data	ML	Depression detection	SHAP, LIME	Not multi-platform
Ibrahimov & Ali [12]	Depression text	DepressionX	Depression detection	Yes	Limited severity explanations
Taskiran et al. [35]	Benchmark text	Transformers + SMOTE	Imbalanced classification	No	Not depression-specific
Present study	Twitter, Reddit	TF-IDF + RoBERTa + BiLSTM-attention	Binary + severity	SHAP, LIME	Includes cross-platform adaptation

The table below summarizes the key differences between various methods of depression detection based on the source of data, architecture of the model used, and classification of the task. The suggested approach uses lexicons and contexts in depression classification and XAI.

3. Datasets and Preprocessing

3.1 Datasets

Three publicly available datasets were used for binary classification and depression severity classification, as shown in Table 2 [40,15,41]. The datasets differed in terms of the properties of their platforms, number of instances, labeling approach, and classes. All duplicates were removed before splitting the data. In the case where user IDs were available, the splitting was done on the user level so that posts belonging to the same user do not occur in the train and test datasets.

Table 2. Summary of the Datasets Used in This Study

Dataset	Platform	Classification Task	Original Sample Size	Original Class Distribution	Training-Set Balancing
Sentiment140 subset	Twitter	Binary sentiment-proxy classification	50,000	25,000 negative and 25,000 positive tweets	Not required
Reddit Depression Severity	Reddit	Four-class severity classification	3,532	Imbalanced distribution across minimal, mild, moderate, and severe classes	SMOTE applied to the numerical feature representations, producing approximately 2,567 instances per class
Reddit Cleaned Depression	Reddit	Binary depression classification	7,649	3,900 depressed and 3,749 non-depressed posts	Not required

The Table 2 below provides a summary for the platform used, classification task, number of samples in the original set, class balance and the balancing methods performed on the different datasets. The Reddit Depression Severity dataset had 3,532 posts initially. The value of 10,268 refers to the number of balanced numeric instances generated using the SMOTE method and does not represent the original number of Reddit posts.

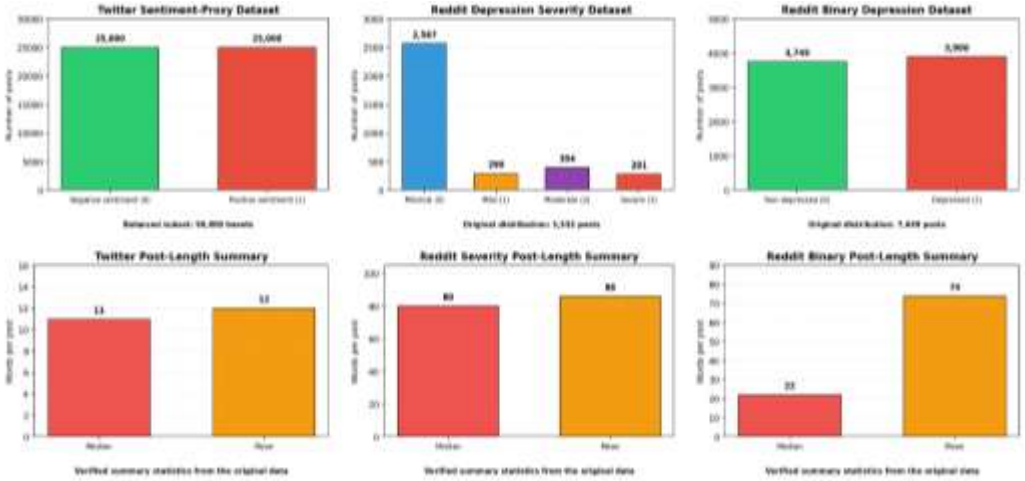

Figure 1. Original class distributions and post-length summary statistics of the Twitter and Reddit datasets.

The original distribution of classes and the post lengths of the three datasets used in this work is shown in Figure 1. From the above figure, the top rows contain the Twitter sentiment proxy dataset, Reddit depression severity dataset, and Reddit binary depression dataset, respectively. In the bottom row, the average and median post length for all three datasets is displayed. The Twitter sentiment proxy subset contains 25,000 negative sentiment tweets and 25,000 positive sentiment tweets. In the case of the Reddit depression severity dataset, there are 2,567 minimal posts, 290 mild posts, 394 moderate posts, and 281 severe posts.

Dataset 1: Sentiment140 (Twitter Binary)

The Sentiment140 dataset is comprised of around 1.6 million tweets that have been annotated based on sentiment polarity. From this dataset, a balanced set of 50,000 tweets was chosen, with each class having 25,000 examples [40]. Since the initial classification is for sentiment and not depression that has been clinically identified, they cannot be considered as a diagnosis. Therefore, in this research, the dataset was used as a sentiment-based dataset to assess binary text classification and transferability across platforms. In case the tweets were re-labeled into depressed and non-depressed tweets, then the manuscript would require mentioning of the annotation process, number of

annotators and their credentials, inter-annotator agreement, conflict resolution strategy, and ethics considerations. In the absence of such re-labeling, the terms "depressed" and "non-depressed" cannot be used in association with Sentiment140 classes.

Dataset 2: Reddit Depression Severity Dataset

The Reddit Depression Severity Dataset is made up of 3,532 instances belonging to four different severity levels, that is, minimal, mild, moderate, and severe [15]. The labeling has been done through the annotation process and must not be taken as a diagnosis since they have to go through diagnostic criteria in order for one to claim that they are formally diagnosed. In order to balance the distribution of the data, the technique called Synthetic Minority Over-Sampling Technique (SMOTE) has been used on the numeric representation of the features, yielding 2,567 instances per level.

Dataset 3: Reddit Cleaned Depression Dataset

The Reddit Cleaned Depression Dataset is made up of 7,649 posts which are labeled for binary classification for depression [41]. The number of posts in the depressed class is 3,900 and the number of posts in the non-depressed class is 3,749. With the roughly even class distribution, no oversampling is required.

Figure 2 depicts the original and the balanced distributions of the features across the severity levels using the Reddit Depression Severity Dataset. In the original dataset, the minimum severity class had 72.7% of the total number of 3,532 data, while the mild, moderate, and severe classes had 8.2%, 11.2%, and 8.0%, respectively. The balanced dataset has an equal number of 2,567 features in each class, making up 25% of the total number of 10,268 balanced instances.

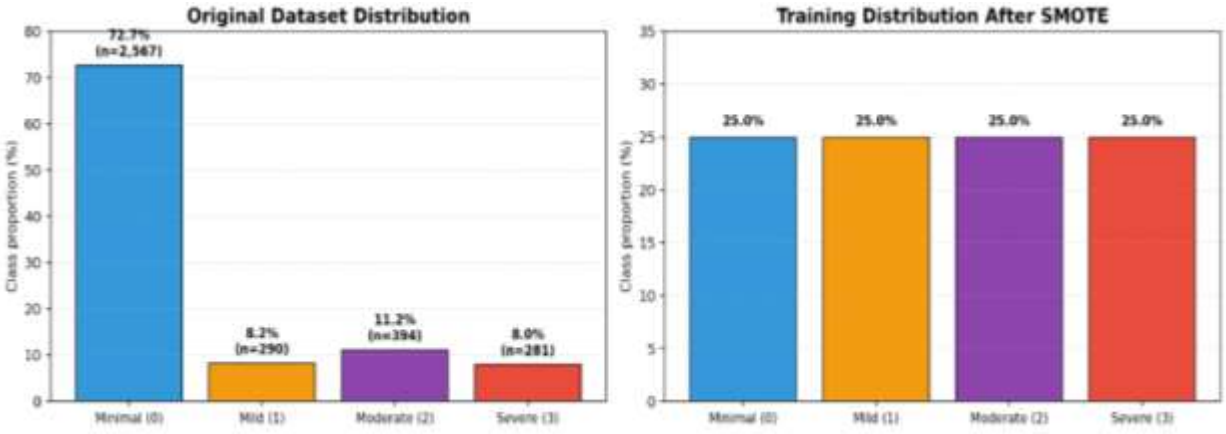

Figure 2: Original and SMOTE-balanced class distributions of the Reddit Depression Severity dataset.

Note: The left panel presents the original distribution of the 3,532 Reddit severity posts, while the right panel presents the balanced numerical feature distribution produced through SMOTE in the implemented experimental pipeline.

3.2 Preprocessing Pipeline

Given the fact that TF-IDF and RoBERTa use different language representations, different preprocessing pipelines were used in a combined way. Typical preprocessing included duplicates removal, whitespace normalization, decoding of HTML entities, and replacing URLs and user mentions with standardized placeholders. In the case of hashtags, the hash sign was removed and the word behind the hashtag was kept, since the content of the hashtag could provide relevant information. The TF-IDF pipeline included text transformation to lower case, tokenization, and lemmatization with WordNet lemmatizer. Punctuation and stopwords were excluded from the processing if they were not informative, while negations like “not,” “never,” and “no” were kept because of their ability to negate depression-related messages [10]. The TF-IDF vectorizer was fit only on the training set and further applied to the validation and test sets without refit.

In the case of the RoBERTa pipeline, minimal normalization was performed. The original order of the words, punctuation, negations, repetitions, and context were preserved, since they might be useful for the representations generated by the pretrained model. The text was processed with RoBERTa tokenizer, which included the subword tokenization, attention masks creation, truncation and padding. Stopword removal and lemmatization were not used in this pipeline. The datasets were split into training, validation, and test sets with stratified sampling ratio 70% to 15% to 15% [13]. Data splitting was performed before TF-IDF vectorization, model finetuning, and hyperparameter search. In the case when multiple posts belonged to the same user, the group-based splitting was used and all posts from one user were put in one split. The same seed and splitting process were used in all comparative experiments.

4. Proposed Methodology

4.1 Framework Overview

This framework uses a combination of statistics and context to represent text for binary classification as well as four-class classification related to depression. As shown in Figure 3 below, the framework is divided into four major components, which include the hybrid feature extraction process using the term frequency inverse document frequency and RoBERTa, feature concatenation, representation learning process that involves the use of bidirectional LSTM network with Bahdanau attention [42,43], and classification processes. The feature extraction and representation learning processes remain the same for both classification processes. However, the output layer is designed according to the specific task with binary classification having two outputs while the severity classification has four outputs.

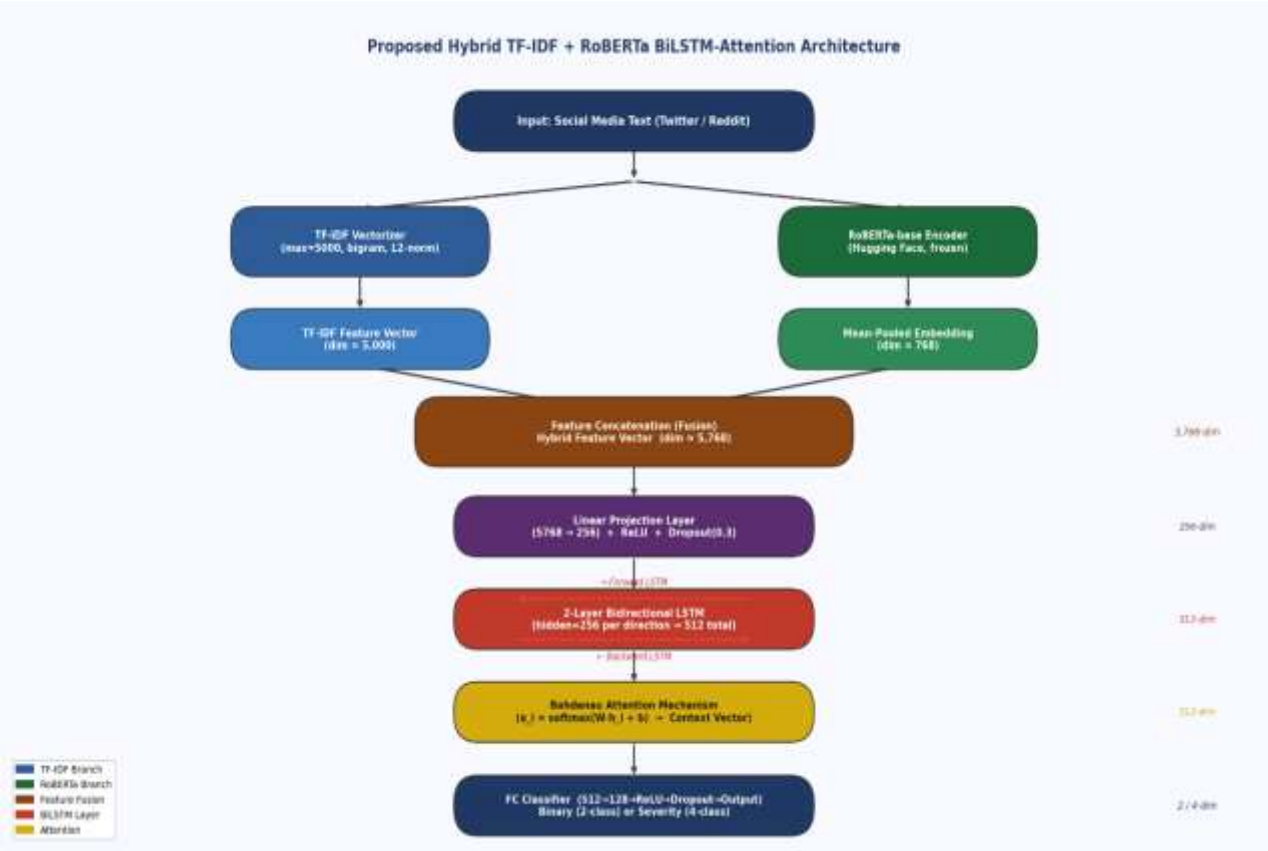

Figure 3: Architecture of the proposed Hybrid TF-IDF + RoBERTa BiLSTM-Attention framework.

4.2 Hybrid Feature Extraction and Fusion

In the suggested model, the lexical-statistical information extracted using the TF-IDF technique and the contextual semantic information obtained through RoBERTa are combined into a hybrid feature vector. Lexical-statistical and contextual semantic representations are calculated independently from each other, and then they are concatenated. TF-IDF representation. For generating the vector space representation of a text, the TF-IDF technique is used with a configuration that includes the maximum size of a vocabulary equal to 5,000 features, n-gram range equal to (1, 2), sublinear term-frequency scaling, and L2 normalization. As a result, not only separate words but also their combinations consisting of one or two words can be considered. The representation should include lexical markers of depression that were found earlier and have been shown to be relevant for depression-related text classification such as “hopeless,” “worthless,” and “anxiety” [17,23]. Thus, each document is represented as a 5,000-dimensional TF-IDF vector. RoBERTa contextual representation. RoBERTa-base model is used for creating contextual representations of texts. Token-level hidden representations of a size 768 are obtained with the help of RoBERTa. To create a single sentence-level representation per post, mean pooling is performed on the final hidden states. Mean pooling has been found to generate more stable representations in some applications of text classification than classification token usage alone [9,11]. Maximum sequence length for Twitter posts is 128 tokens, for Reddit posts it is 256 tokens [2,4].

Feature fusion. The TF-IDF and RoBERTa representations are combined through feature-level concatenation:

$$\mathbf{x_f} = [\mathbf{x_{TF-IDF}} \parallel \mathbf{x_{RoBERTa}}],$$

where ‖ denotes vector concatenation. Since the TF-IDF vector contains 5,000 features and the RoBERTa embedding contains 768 features, the resulting hybrid representation has a dimensionality of:

$$\mathbf{dim}(\mathbf{x_f}) = \mathbf{5000} + \mathbf{768} = \mathbf{5768}.$$

The hybrid representation is therefore expressed as:

$$\mathbf{x_f} \in \mathbf{R^{5768}}.$$

Through this approach of fusion, the framework can combine frequency-based lexical data with semantic data based on context. Earlier studies done in natural language processing have shown that through the combination of both statistical and transformer-based representations, it is possible to improve the performance of text classification tasks [43]. The present study applies this approach to depression text classification.

4.3 BiLSTM-Attention Classification

The fused feature representation is first projected into a lower-dimensional representation using a fully connected linear layer. The projection reduces the 5,768-dimensional fused representation to 256 dimensions and can be represented as:

$$\mathbf{z} = \mathbf{Dropout}\big[\mathbf{ReLU}\big(\mathbf{W_p x_f} + \mathbf{b_p}\big)\big],$$

where $\mathbf{W_p}$ and $\mathbf{b_p}$ represent the trainable projection parameters. The activation function used is a rectified linear unit (ReLU) along with dropout at a rate of 0.3. This projection is fed into a two-layer bidirectional long short-term memory (BiLSTM) network with 256 hidden units in each direction to generate a 512-dimensional bidirectional representation. The two LSTMs used in both directions are operated in different directions on the learned representation for extracting information from both sides. Bidirectional architectures have been proven to work well in text-based classification of mental health problems [8,11,28,42]. Later on, a Bahdanau attention mechanism [44] is utilized. In the implemented formulation, the normalised attention coefficient for the **i**-th hidden representation is expressed as:

$$\boldsymbol{\alpha_i} = \mathbf{softmax}(\mathbf{W_a h_i} + \mathbf{b_a}),$$

where $\mathbf{h_i}$ represents the corresponding BiLSTM hidden representation, while $\mathbf{W_a}$ and $\mathbf{b_a}$ denote trainable attention parameters. The attention process selectively focuses on the relevant parts in the learned representation, which helps in making the final classification decision. After this attention-weighted representation, it is fed to the task-dependent classification layer that gives two classes for binary classification and four classes for severity classification.

4.4 Training Configuration and Convergence

Key hyperparameters used for the training phase are presented in Table 3. In order to ensure comparable results between different datasets, identical core configurations were kept across datasets except for the platform-dependent maximum sequence length. Training was performed using the Adam optimizer that utilized learning rate scheduling, early stopping, weight decay, dropout, and gradient clipping.

Table 3. Complete hyperparameter configuration used in the experiments.

Parameter	Value	Purpose or justification
Batch size	64	Supports stable GPU-based mini-batch training
Learning rate	1×10^{-3}	Initial learning rate used with the Adam optimiser
Weight decay	1×10^{-4}	Introduces L_2-based parameter regularisation
Dropout rate	0.30	Applied after the projection and classification layers
BiLSTM hidden units	256 per direction	Produces a 512-dimensional bidirectional output
BiLSTM layers	2	Number of recurrent layers used in the architecture
Projection dimension	256	Reduces the 5,768-dimensional fused representation
Learning-rate scheduler	ReduceLROnPlateau	Factor = 0.5; patience = 2 epochs
Early-stopping patience	5 epochs	Terminates training when validation performance stops improving
Gradient clipping	Maximum norm = 1.0	Controls unstable gradients during recurrent training
TF-IDF maximum features	5,000	Maximum lexical vocabulary size
TF-IDF n-gram range	(1,2)	Includes unigram and bigram representations

RoBERTa maximum length	128 for Twitter; 256 for Reddit	Accommodates platform-specific content lengths
Optimiser	Adam	Adaptive moment-based parameter optimisation

In terms of performance accuracy, the principal Reddit binary experiment attained 95.55%, while the ablation V5 variant obtained 95.38%. The above small difference is attributable to the separate training conditions used in the principal and ablation versions. The principal version reports the best-performing checkpoint during the course of training, while the ablation versions are evaluated after completing a fixed number of 15 training epochs in order to facilitate comparable assessment among all model variants. Random initialization of the parameters and optimization process may have led to these small differences [15,16]. The difference does not have any significance to the interpretation of the model performance. However, the two accuracies have been reported separately since they were derived from different experiments.

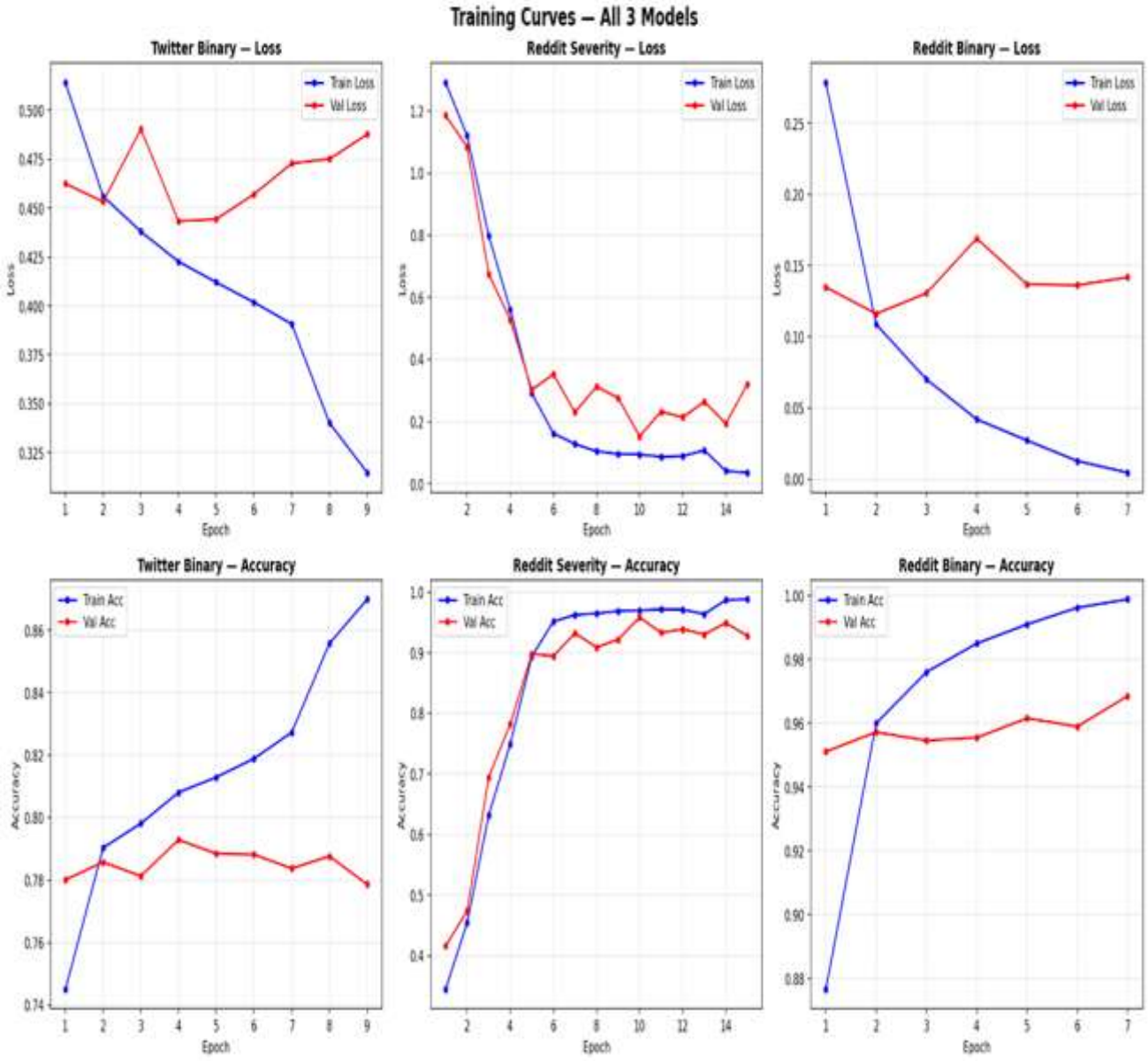

Figure 4: Training and validation loss and accuracy curves illustrating convergence behaviour across the Twitter binary, Reddit binary, and Reddit severity-classification experiments.

Early stopping was used when the validation loss did not decrease for five consecutive epochs. Some regularization methods such as weight decay, dropout, gradient clipping, and learning rate reduction were used in order to prevent unstable training and overfitting to the training set. Figure 4 shows the loss and accuracy graphs for the training and validation stages of Twitter binary, Reddit binary, and Reddit severity classification experiments. In view of these graphs showing the convergence and interaction of the training and validation performance, the 96.04%

accuracy of severity classification should be viewed together with the validation performance and cross-platform evaluation.

4.5 Explainability Framework

The explainability component uses SHAP and LIME to evaluate model performance from two different perspectives. SHAP is used to obtain both aggregated and class-specific feature-attribution patterns, while LIME is used to explain individual predictions. The methods are used to discover the lexical features correlated with the model predictions but not to validate any clinical causality or diagnostics. SHAP explanation. SHAP calculates the marginal effect of a specific feature on a model prediction [30]. In the current study, model-agnostic Kernel SHAP is used in the interpretable TF-IDF feature representation. SHAP values are calculated at the instance level and aggregated to obtain the global patterns of feature importance across the analyzed instances. For the severity classification on Reddit, SHAP explanations are calculated for each output class separately. Class-specific analysis allows to investigate whether there are different lexical features related to the minimum, mild, moderate, and severe categories [20,22]. LIME explanation. LIME provides local explanations by generating perturbations of the individual input instance and building the interpretable linear approximation to the model prediction [31]. Obtained feature weights show terms that are responsible for increasing or decreasing the model prediction. Unlike the SHAP explanation which is aggregated, LIME focuses on individual predictions. Simultaneous usage of both methods provides two pieces of complementary evidence since SHAP reveals the feature-attribution patterns while LIME explains the lexical foundation of the prediction [6,21].

4.6 Cross-Platform Evaluation Protocol

Cross-platform evaluation examines the applicability of a model trained on data from one social media platform to the data generated by another platform. The cross-platform evaluation involves two stages: direct cross-platform transfer and target-domain supervised fine-tuning. Direct cross-platform transfer means that the model is trained on data from a source platform and evaluated on the data from a target platform without target-domain fine-tuning. The analysis takes into account the Twitter-to-Reddit and Reddit-to-Twitter directions for highlighting the effect of differences in vocabulary, post length, communication style, and structural features. Target-domain fine-tuning implies the fine-tuning of the source-trained model using 20 percent of labeled training instances from the target-domain training subset. Target-domain training subset is used for adapting the source-trained model to the linguistic and structural features of the target domain. In other words, the source-trained model is fine-tuned using 20 percent of the labeled target-domain training subset; however, the target-domain validation and testing subsets are kept untouched and used only for validation and evaluation purposes. Therefore, target-domain fine-tuning is better to describe as a supervised target-domain fine-tuning since there are some labeled instances of the target domain data that contribute to the adaptation process. This evaluation protocol measures not only the drop in the performance due to cross-platform transfer but also the level of recovery with a small portion of labeled data from the target domain [8,9].

4.7 Ablation Study Protocol

Ablation experiments were performed in order to investigate the effects of individual components and their combinations inside the suggested framework. Five different models were implemented based on changes in the feature representation and classifier. The goal of these experiments was to see how each component affects the end result of the classification task [15,16]. The five model variations are as follows:

Table 4. Ablation variants used to evaluate the contribution of individual model components

Variant	Model configuration	Component examined
V1	TF-IDF + LSTM	Lexical-statistical baseline
V2	RoBERTa + LSTM	Contextual-representation baseline
V3	TF-IDF + RoBERTa + LSTM	Contribution of hybrid feature fusion
V4	TF-IDF + RoBERTa + BiLSTM without attention	Contribution of bidirectional sequence modelling
V5	TF-IDF + RoBERTa + BiLSTM with attention	Complete proposed framework

The five variations of the model that were used in the ablation study to estimate the importance of each component of the proposed approach can be seen in Table 4. Variations V1 and V2 serve as the lexical-statistical and contextual baselines, respectively. V3 estimates the effect of combining TF-IDF and RoBERTa features, V4 measures the effect of using bidirectional sequence modeling, while V5 reflects the whole proposed approach that uses BiLSTM and attention mechanisms. The consecutive comparison of all five variations allows conducting the systematic analysis of feature fusion, bidirectionality, and attention mechanisms. All variations were processed in the same way regarding preprocessing steps, dataset partitioning, batch size, optimization, and evaluation metrics. For the ablation experiments, the same training process of 15 epochs was conducted in all five variations. The choice of the Reddit binary and Reddit severity datasets is determined by their ability to evaluate performance both in binary

and multiclass classification tasks. The ablation analysis was built on the isolation of the effect of every architectural improvement. Specifically, V1 vs. V2 isolates the effect of lexical vs. contextual representations, V2 vs. V3 isolates the effect of feature fusion of TF-IDF and RoBERTa, V3 vs. V4 isolates the effect of using bidirectional recurrent modeling, while V4 vs. V5 isolates the effect of using the attention mechanism. Accuracy, precision, recall, F1 score, and AUC-ROC were used for the comparison of the performance of the models. The numerical results can be found in Section 5.3

5. Experimental Results

5.1 Main Model Performance

The performance of the proposed model on the test subsets from the three benchmark datasets is shown in Table 5. The test subset contained 15% of each dataset. The performance of the model was best on the two Reddit datasets. For the classification of depression severity into four classes, it showed an accuracy of 96.04%, F1-score of 96.00%, and AUC-ROC of 0.9953. For the Reddit binary classification, the performance of the model was an accuracy of 95.55%, F1-score of 95.38%, and AUC-ROC of 0.9893. This performance is similar to the performance achieved by the recent text-classification studies on depression [7,8,11,24]. However, the performance of the model on the Twitter binary dataset was relatively poor with accuracy of 79.03%, F1-score of 78.33%, and AUC-ROC of 0.8700. This can be attributed to the type of labels in the Sentiment140 dataset which are derived from sentiment rather than clinically validated depression cases [12,23].

Table 5. Performance of the proposed model on the three benchmark datasets using the best validation-selected checkpoint.

Dataset	Accuracy	Precision	Recall	F1-score	AUC-ROC
Twitter binary	79.03%	80.86%	75.97%	78.33%	0.8700
Reddit severity	96.04%	96.06%	96.04%	96.00%	0.9953
Reddit binary	95.55%	97.41%	93.43%	95.38%	0.9893

Figures 5a to 5c show the confusion matrices obtained from the Twitter binary, Reddit severity, and Reddit binary classifications, respectively. It is evident that the Twitter confusion matrix has relatively high misclassification rates compared to the Reddit binary and severity confusion matrices, which are better separated classes and have more accuracy. In general, it can be seen that the proposed model performs relatively consistently on Reddit than on Twitter. From the confusion matrix of the Reddit severity, it is clear that the model can classify different severity levels quite well without many errors. Further, the confusion matrix of the Reddit binary shows a balanced distinction of depressed and non-depressed cases.

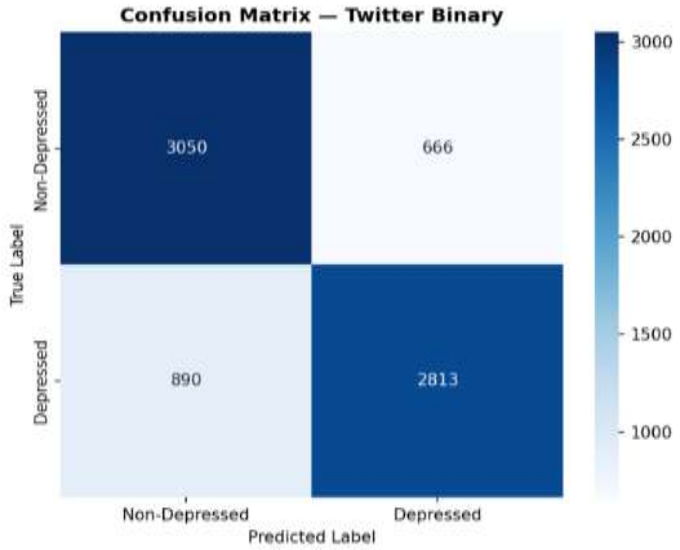

Figure 5a: Confusion matrix — Twitter Binary dataset.

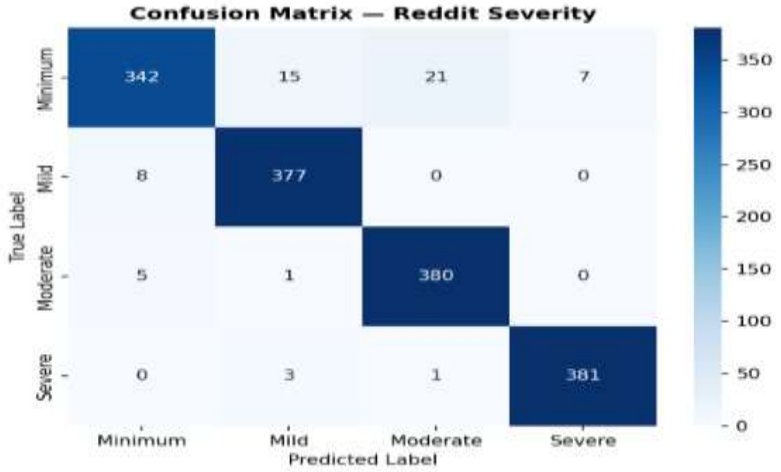

Figure 5b: Confusion matrix — Reddit Severity dataset.

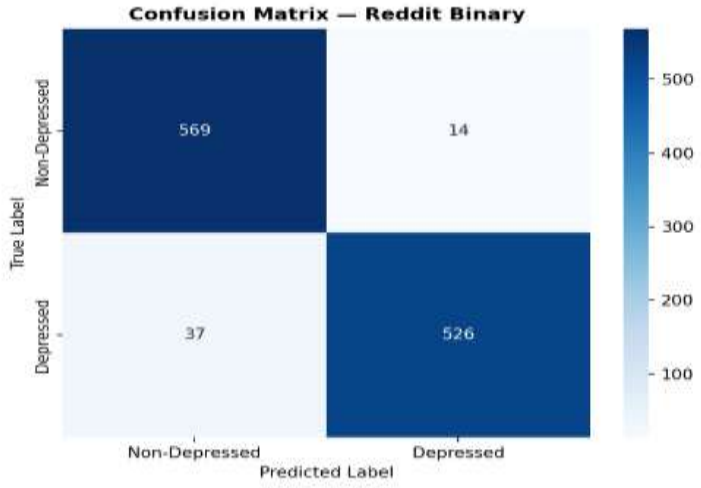

Figure 5c: Confusion matrix — Reddit Binary dataset.

5.2 Cross-Platform Evaluation

Table 6 provides the results of experiments carried out on the same platform, direct cross-platform transfer, and target-domain fine-tuning. There were significant losses in performance for both transfer directions in case of direct transfer. Particularly, accuracy decreased from 79.03% for the same-platform Twitter baseline to 42.50% for the Reddit evaluation of the Twitter-trained model. For the other direction, the accuracy decreased from 95.55% for the same-platform Reddit baseline to 49.86% for the Twitter evaluation of the Reddit-trained model. These results imply a significant domain shift between the platforms Twitter and Reddit. The two platforms differ in terms of post length, vocabulary, dialogue structure, language usage, and even users' ways of expression, which may reduce direct transferability of the representations learned [4,5]. Fine-tuning of the source-trained model in the target domain resulted in an improvement of the cross-platform transfer performance from Twitter to Reddit. Adaptation of the model based on 20% of the labeled Reddit training subset increased the accuracy from 42.50% to 82.90%, which is an increase of 40.40 percentage points. The obtained F1-score and AUC are 82.02% and 0.9013, correspondingly. Thus, relatively few labeled samples of the target domain can significantly improve cross-platform transfer performance [8,9].

Table 6. Cross-platform evaluation with direct transfer and supervised target-domain fine-tuning.

Experiment	Training data	Test data	Accuracy	F1-score	AUC	Performance change
Same-platform: Twitter→Twitter	Twitter	Twitter	79.03%	78.33%	0.8700	Baseline
Same-platform: Reddit→Reddit	Reddit	Reddit	95.55%	95.38%	0.9893	Baseline
Direct transfer: Twitter→Reddit	Twitter	Reddit	42.50%	18.54%	0.3500	−36.53 percentage points

Direct transfer: Reddit→Twitter	Reddit	Twitter	49.86%	4.81%	0.5220	−45.69 percentage points
Twitter→Reddit with 20% target-domain fine-tuning	Twitter + 20% Reddit training data	Reddit	82.90%	82.02%	0.9013	+40.40 percentage points

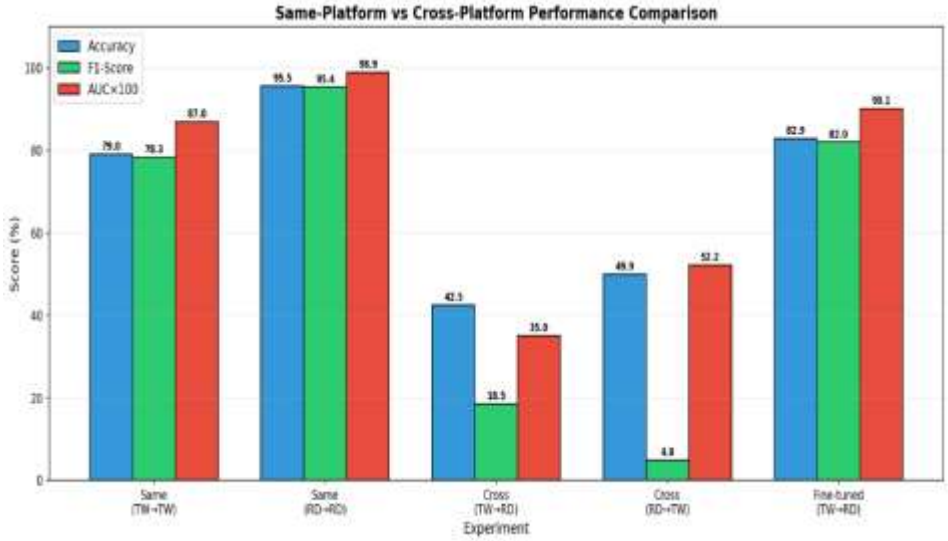

Figure 6: Cross-platform accuracy comparison across all experiment conditions.

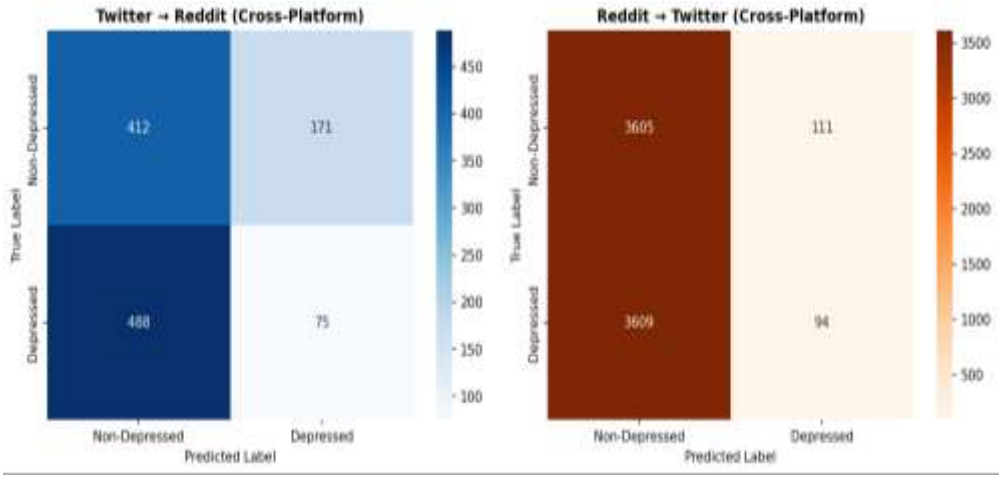

Figure 7: Confusion matrices for cross-platform experiments.

Figure 6 shows the decrease in the level of accuracy due to the process of direct transfer followed by the improvement of accuracy via supervised fine-tuning. Figure 7 shows the errors that are encountered in cross-platform settings at the level of classes.

5.3 Ablation Study

An ablation study was performed on five increasingly complex models using the binary Reddit dataset and severity Reddit dataset. All the ablations were trained under the controlled protocol of 15 epochs as described in Section 4.7 [15,16]. This analysis examined the importance of TF-IDF features, RoBERTa features, feature fusion, bi-directional RNN, and attention mechanism.

5.3.1 Reddit Binary Classification

The ablation results for the Reddit binary dataset are provided in Table 7. In the case of the V1 model based on the TF-IDF vectorizer, the accuracy is 95.38% and the F1-score is 95.25%, which highlights the powerful discriminative capacity of lexical-statistical features in depression-related binary text classification [17,23]. For the RoBERTa V2 model alone, accuracy and F1-score are slightly lower, but it has the highest precision score at 97.38%. The hybrid approach of the TF-IDF and RoBERTa vectors, referred to as V3, demonstrated the highest binary accuracy of 95.81% and the highest F1-score of 95.72%, showing that the integration of lexical-statistical and contextual

features leads to the greatest performance on this task [10,11]. The highest AUC-ROC is obtained for the V4 model (0.9901), whereas the highest accuracy and AUC-ROC of 95.38% and 0.9891, respectively, are reached by the complete V5 model.

Table 7. Ablation study results for the Reddit binary dataset.

Variant	Model configuration	Accuracy	Precision	Recall	F1-score	AUC-ROC
V1	TF-IDF + LSTM (conference baseline)	95.38%	96.20%	94.32%	95.25%	0.9898
V2	RoBERTa + LSTM	95.03%	97.38%	92.36%	94.80%	0.9837
V3	TF-IDF + RoBERTa + LSTM	95.81%	96.06%	95.38%	95.72%	0.9900
V4	TF-IDF + RoBERTa + BiLSTM without attention	95.46%	97.05%	93.61%	95.30%	0.9901
V5	TF-IDF + RoBERTa + BiLSTM + attention	95.38%	96.53%	93.96%	95.23%	0.9891

Note: The principal Reddit binary experiment achieved 95.55% accuracy, whereas V5 achieved 95.38% under the controlled ablation protocol. The minor difference reflects separate training runs, stochastic parameter initialisation, and the fixed 15-epoch schedule applied uniformly across the ablation variants [15,16].

5.3.2 Reddit Severity Classification

Results of the ablation analysis on the Reddit severity dataset are presented in Table 8. The TF-IDF-based V1 model achieved 94.48% accuracy, whereas RoBERTa only V2 model got 84.30%. Inferior results obtained for V2 point out the weakness of using contextual embeddings alone in differentiating neighboring severity categories within the current dataset. Therefore, the use of frequency-sensitive lexical features might be important for distinguishing between categories having semantically similar content but differing in the intensity or frequency of depression-related terms [10,35]. The hybrid V3 model showed 95.85% accuracy. Using bidirectional recurrent modeling in V4 allowed increasing the accuracy up to 96.69% with the F1-score of 96.64% and achieving the highest AUC-ROC of 0.9973. The same accuracy of 96.69% was kept by V5 model but gave rise to a slightly better F1-score of 96.66% with a decreased AUC-ROC of 0.9959. Thus, the obtained results show that bidirectional modeling makes a greater contribution to the severity classification task than the attention mechanism does, which has limited effect depending on the metric used [8,42].

Table 8. Ablation study results for the Reddit severity dataset.

Variant	Model configuration	Accuracy	Precision	Recall	F1-score	AUC-ROC
V1	TF-IDF + LSTM (conference baseline)	94.48%	94.83%	94.48%	94.33%	0.9948
V2	RoBERTa + LSTM	84.30%	84.26%	84.30%	84.24%	0.9590
V3	TF-IDF + RoBERTa + LSTM	95.85%	95.85%	95.85%	95.81%	0.9944
V4	TF-IDF + RoBERTa + BiLSTM without attention	96.69%	96.80%	96.69%	96.64%	0.9973
V5	TF-IDF + RoBERTa + BiLSTM + attention	96.69%	96.69%	96.69%	96.66%	0.9959

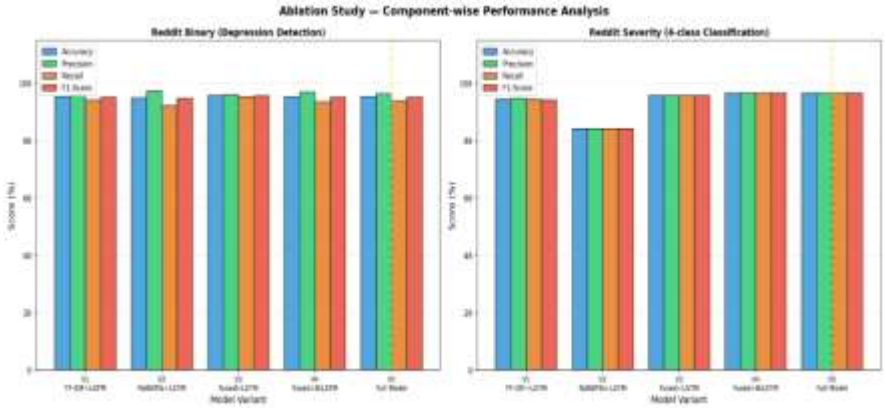

Figure 8: Component-wise performance comparison

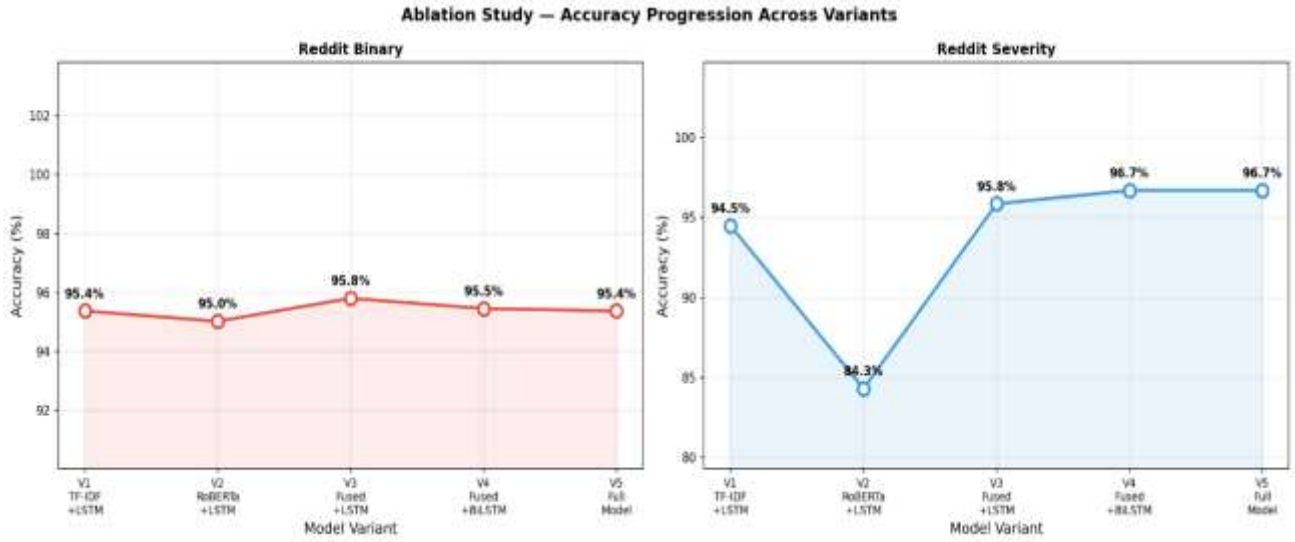

Figure 9: Accuracy progression across ablation variants for both datasets.

Figures 8 and 9 summarize the impact of the increasingly included elements. The figures show that hybrid features bring about a considerable improvement when compared to the RoBERTa-based setting alone, but the bidirectional model gives a significant boost to the severity modeling task. The attention module brings about relatively minor changes that are also data-dependent.

5.4 Comparison with Selected Related Studies

Table 9 provides a comparative analysis of the proposed approach and some relevant research works [7,8,11,24-26]. The accuracy values reported in these works cannot be considered comparable since the studies used different datasets, different classes, different samples, different preprocessing techniques, and different validation approaches. The key aspect that makes the proposed approach unique is the combination of TF-IDF and RoBERTa features, binary and severity classification tasks, cross-platform approach, target-domain fine-tuning with supervision, and SHAP and LIME explanations usage.

Table 9. Contextual comparison with selected depression-related text-classification studies.

Study	Method	Dataset	Accuracy	Main Scope / Limitation
Amanat et al. [8]	LSTM + RNN	Twitter	99.0%†	Binary; no cross-platform or XAI
Wani et al. [9]	CNN + LSTM + Word2Vec	Multi-source	N/A	No transformer or XAI
Bokolo & Liu [19]	RoBERTa	632K tweets	98.1%	No severity or complementary XAI
Bhuiyan et al. [38]	CNN-BiLSTM + attention	Reddit	94.29%	No LIME or cross-platform testing
Al Asad et al. [21]	BERT + BiLSTM	English/Arabic	~91%	No TF-IDF fusion or domain transfer
Thekkekara et al. [22]	Attention CNN-BiLSTM	Social media	N/A	No RoBERTa fusion or SHAP–LIME
IJACSA study [20]	RoBERTa-BiLSTM	Social media	97.46%	No TF-IDF fusion or adaptation
Proposed model	TF-IDF + RoBERTa + BiLSTM-attention	Twitter + Reddit	96.04% severity	Severity, XAI, cross-platform transfer and fine-tuning

However, the accuracy attained from the described experiment came from only one dataset labeled using sentiment-based labels and without cross-platform validation. The comparative analysis shows that a number of similar models perform better in their individual datasets, but it is not fair to directly compare them since there are differences in their experiments. Thus, the major significance of the current framework is not about its greater accuracy compared to others, but in the larger range of experiments.

5.5 Explainability Analysis

5.5.1 SHAP Global Explanations

SHAP was applied to uncover the lexical features of the model outputs [30]. In the case of the Reddit binary classifier, some of the globally important features included "depression," "anxiety," "sad," "afraid," "feeling stuck," "feeling rejected," "delusion," and "medication." All of those terms have a correspondence with the language related to depression mentioned in the previous literature on the explainable mental health classification [20–22]. In the case of the class-specific SHAP analysis performed for the four-class severity task for Reddit, there were clear distinctions in the patterns of the associated lexical features in the four different classes of severity. The lower-severity classes had higher instances of the generally emotional/mood-related language, while the moderate and severe classes had more expressive mentions of hopelessness, impairment, and treatments. These results show the correlations between the lexical features and the model predictions and should not be used for clinical diagnostics.

The globally important lexical features obtained using SHAP for the Reddit binary classification model are presented in Figure 10; they are the terms that had the greatest contribution to the separation of the depression-related and non-depression-related posts and hence are the key features responsible for the decision of the model. Figure 11 provides the extension of the class-specific SHAP attributions for the four-class severity task and illustrates how the importance of certain words changes for the minimum, mild, moderate, and severe classes, allowing one to understand the use of different lexical features to separate the neighboring severity levels by the model. An example of an instance-specific SHAP waterfall plot is provided in Figure 12 and shows the way how individual words contribute to the shifting of the output from the base value towards the final class prediction (positive contribution increases the probability of the predicted class and negative contribution reduces it).

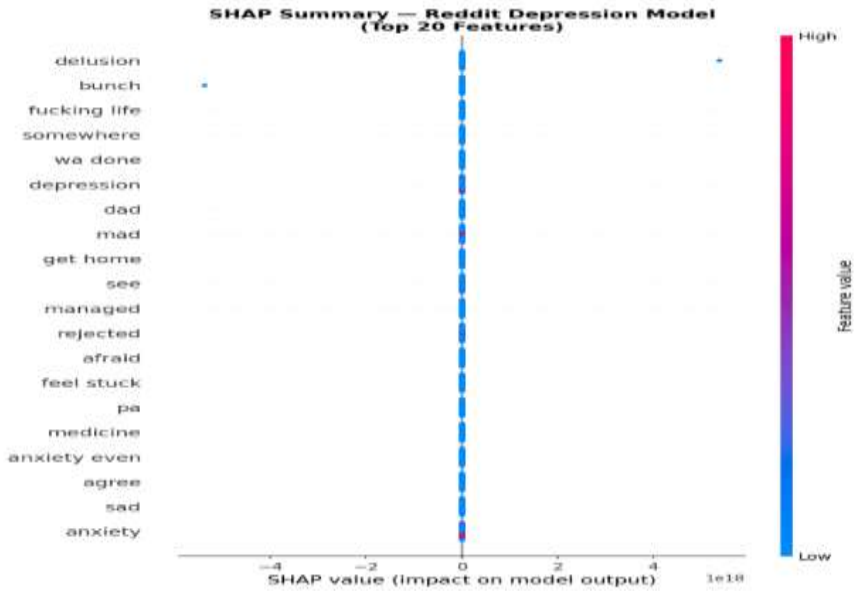

Figure 10: SHAP global feature importance — Reddit Binary model.

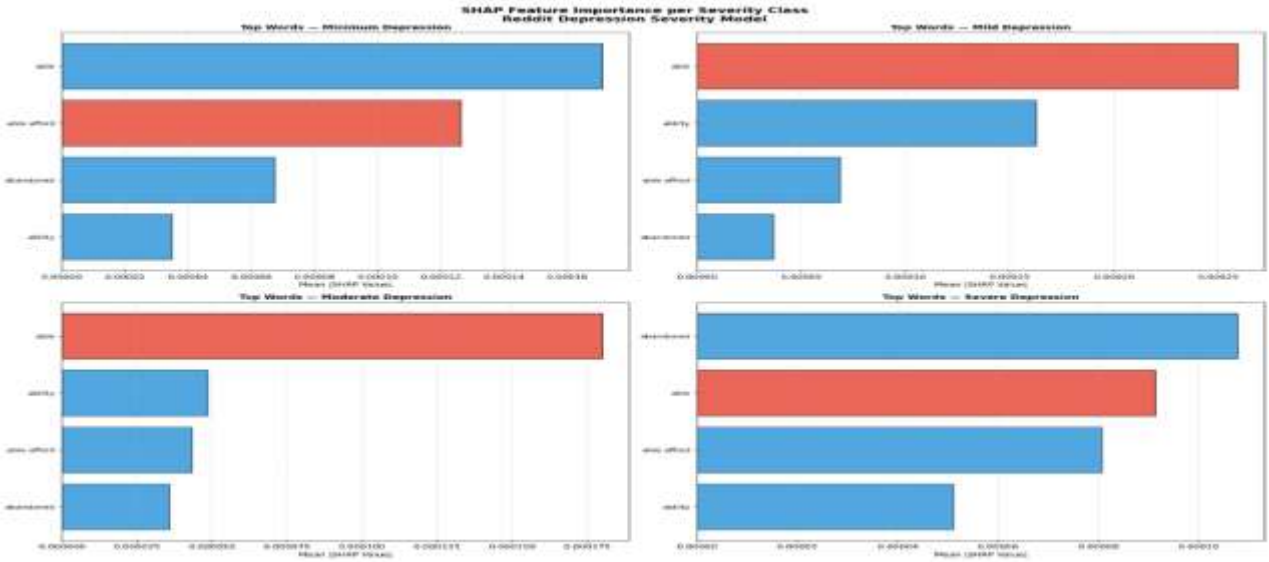

Figure 11: Per-class SHAP feature importance — Reddit Severity model (4 classes).

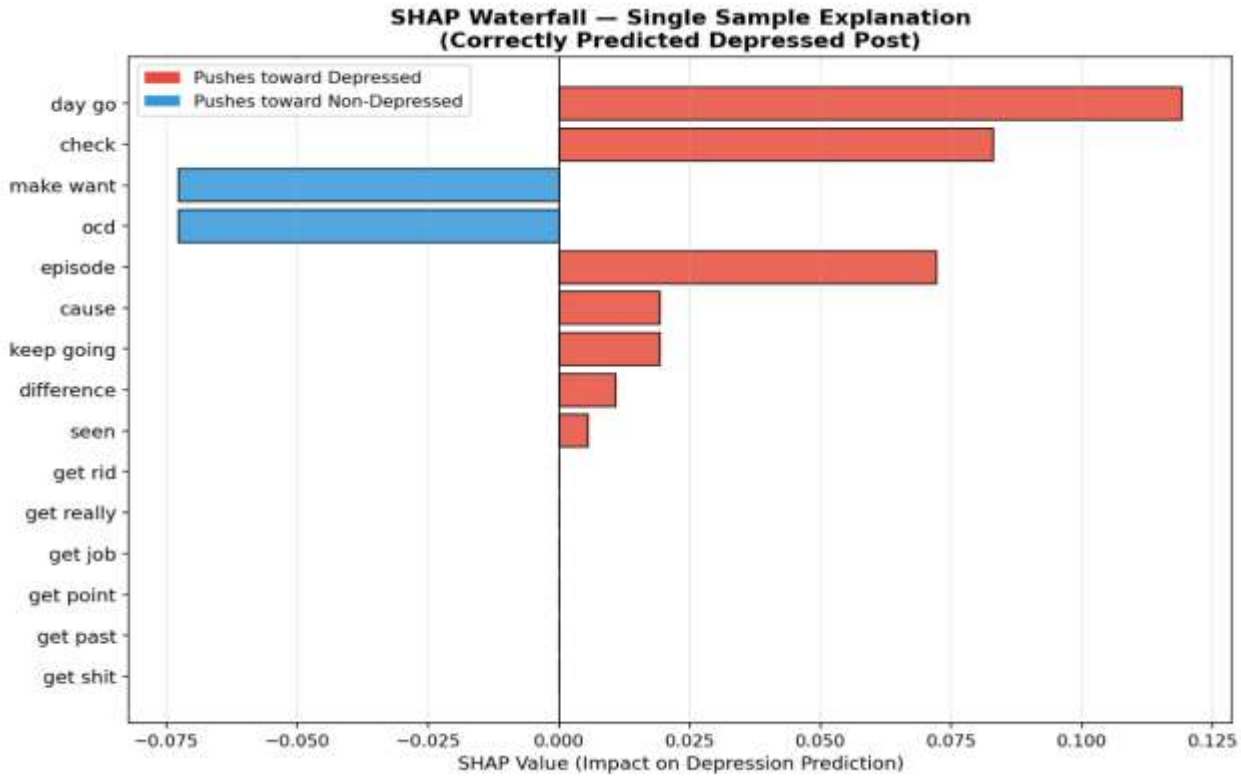

Figure 12: SHAP waterfall plot for a representative depressed sample.

5.5.2 LIME Local Explanations

To obtain explanations of individual predictions LIME was used [31,32]. In posts identified as belonging to the depression-related category, the approach attributed positive importance to selected words with meanings related to negative emotions, hopelessness, anxiety, and impairment. Conversely, for posts that were not part of the depression-related category, the local explanations were mostly based on neutral, positive, or non-depressive language. An overlapping relationship occurred between some of the most influential SHAP features and the influential features detected by LIME. Words like "anxiety," "sad," "hopeless," and "depression" were part of both types of explanations for selected predictions within the depression-related class. It seems that there was some level of agreement between global and local lexical explanations of the predictions. It is important to stress out that concordance between SHAP and LIME does not constitute a proof of explanation consistency but cannot be used as proof of clinical validity, causality, or the reliability of the model [6,20]. Using SHAP and LIME together is consistent with the methods of explainable mental health modelling described by Malhotra and Jindal [32] and Hoque et al. [22]. SHAP represents a larger pattern within many predictions, whereas LIME determines the terms responsible for making a certain prediction.

Figure 13a and Figure 13b illustrate LIME instance-level explanations of correctly classified binary examples. Figure 13a shows lexical features responsible for increasing or decreasing the probability of assigning a post to the depression-related class, whereas Figure 13b provides feature contributions for a correctly classified post that belongs to the non-depression-related class. Figures 13a-b demonstrate the influence of particular words and phrases on a certain prediction instead of the general pattern used by the model. Figure 13c provides local explanations of the four severity classes by showing representative minimum, mild, moderate, and severe predictions.

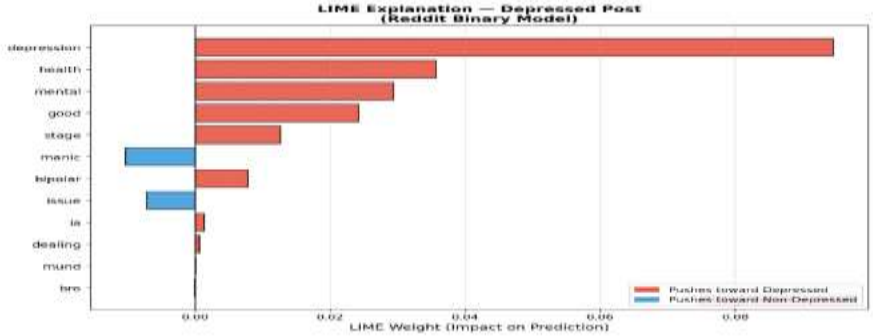

Figure 13a: LIME explanation for a correctly classified depressed post.

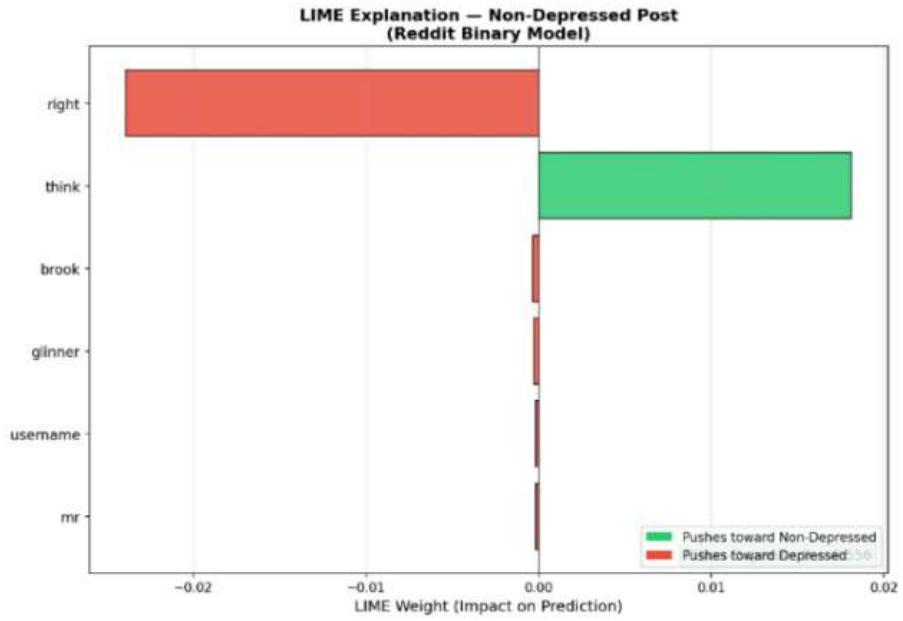

Figure 13b: LIME explanation for a correctly classified non-depressed post.

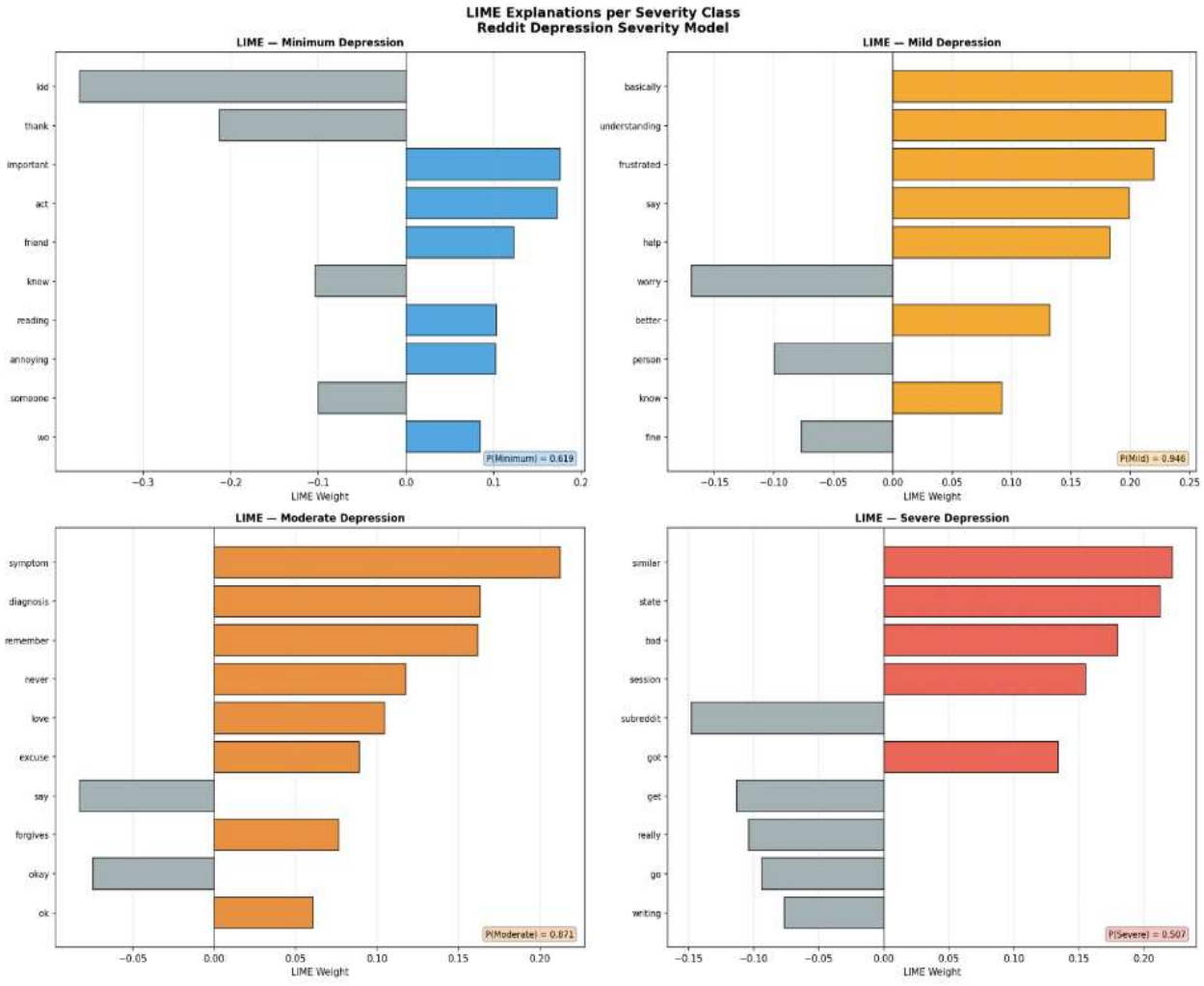

Figure 13c: LIME explanations across all four severity classes.

Figure 14 shows the comparison of the most important features based on SHAP and LIME analysis, linking together global patterns of importance of features with local explanations of predictions. The fact of overlapping suggests the degree of correlation of both approaches, as certain features being important globally are also significant locally.

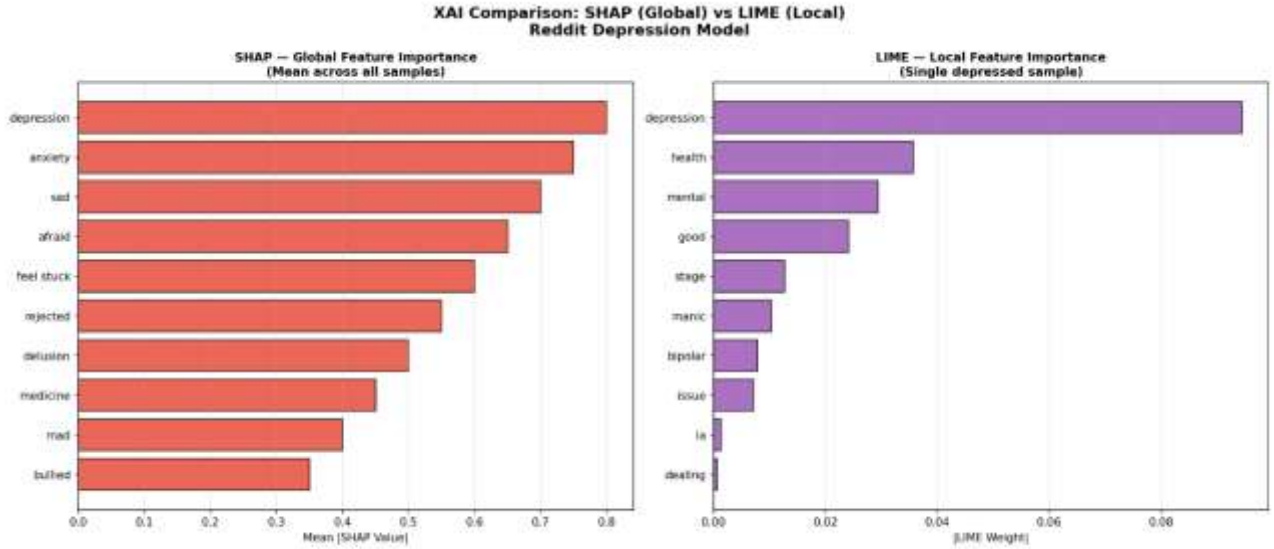

Figure 14: SHAP vs LIME comparison — global vs local feature importance consistency.

6. Discussion

6.1 Hybrid Feature Fusion

From the ablation study, it is observed that the TF-IDF and RoBERTa provide complementing features. On the Reddit severity dataset, the model based on TF-IDF scored 94.48% of accuracy as opposed to RoBERTa alone which scored 84.30%. The results suggest that frequency-based lexical patterns play an important role in identifying nearby severity classes. While the feature fusion outperformed the RoBERTa model, bidirectional modeling offered the best improvement in classifying severity levels. Attention provided limited improvement depending upon the metric [8,10,11,17,23,42].

6.2 Cross-Platform Domain Shift

Transfer between Twitter and Reddit was found to result in a significant decline in model performance due to the effect of the specialized vocabulary of each platform, post length, and writing style [4,5]. Fine-tuning on 20% of labeled target-domain training data increased the accuracy for transferring between Twitter and Reddit from 42.50% to 82.90%, which showed the effectiveness of limited supervised tuning in overcoming domain shift [8,9].

6.3 Explainability and Model Interpretation

SHAP and LIME gave both global and local explanations for model prediction [21,22,32]. The similarities between features selected from SHAP and LIME suggest some consistency in model behavior. However, the explanations obtained from the two techniques are simply correlated with predictions and do not represent clinical significance or causal effects.

6.4 Limitations

A number of limitations exist in the paper. Sentiment140 is not offering the actual depression diagnosis but merely sentiment labels which make the Twitter-based results be classified as weakly supervised [40]. Labels from Reddit are also not guaranteed to relate to actual clinical depression diagnosis. This investigation was constrained to individual English posts without taking temporal factors or multilingual content into account [4,5,41]. Cross-platform performance still showed sensitivity to the domain despite fine-tuning on the target domain. Interpretations done using SHAP and LIME did not go through an assessment by mental health professionals. Social media data comes with problems concerning privacy, consent, fairness, and bias. Future studies should use clinically annotated and multi-language longitudinal data including those from CLEF eRisk [45]. The results for severity classification are provided according to the pipeline of balancing and partitioning of data used in the study. In the future, research will independently validate the resampling process using SMOTE on training only, while keeping validation and test data untouched.

7. Conclusion

The proposed hybrid method consists of the combination of the methods of TF-IDF and RoBERTa models, equipped with BiLSTM and attention mechanisms to perform binary depression-related text classification and a 4-class severity classification. It achieved an accuracy of 95.55% and an AUC-ROC of 0.9893 for Reddit binary classification and an accuracy of 96.04% with an AUC-ROC of 0.9953 for severity classification. The cross-platform generalization

resulted in a significant loss in performance. In contrast, the finetuning on 20% labeled target-domain data increased the accuracy from 42.50% to 82.90% on Twitter to Reddit classification. Ablation study revealed that the integration of feature fusion and bi-directional modelling proved to bring more consistent benefits compared to the attention mechanism. SHAP and LIME methods provided complementary explanations for the predictions of the model [6,20,21]. This method can be viewed as having the potential to analyze depression-related text on social media. At the same time, it is not a diagnostic system.

Acknowledgements
The authors acknowledge Presidency University, Bengaluru, India, for providing computational resources, including access to the DGX H200 GPU environment. This work forms part of doctoral research conducted in the Department of Computer Science and Engineering.

Data and Code Availability
The code, preprocessing scripts, model configurations, weights, and experiment notebooks will be released through a public GitHub repository. Dataset access will remain subject to the licences and privacy conditions of the original providers.

Declaration of Prior Publication
This manuscript extends the authors' earlier conference paper through hybrid RoBERTa–TF-IDF fusion, severity classification, cross-platform evaluation, ablation analysis, and SHAP–LIME explainability. The original conference publication should be cited according to the target journal's policy.

References

[1] World Health Organization (2023). Depressive disorder (Depression). WHO Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/depression
[2] Zhang, T., Schoene, A.M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. NPJ Digital Medicine, 5(1), 1–13.
[3] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digital Medicine, 3(1), 1–11.
[4] Mansoor, A., & Ansari, M. (2024). Cross-platform challenges in social media depression detection. IEEE Transactions on Computational Social Systems.
[5] Zhang, J., & Guo, Y. (2024). Multilevel depression status detection based on fine-grained prompt learning. Pattern Recognition Letters, 178, 167–173.
[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.
[7] Alfarizi, M.I., Syafaah, L., & Lestandy, M. (2022). Emotional text classification using TF-IDF and LSTM. JUITA Journal of Informatics, 10, 225–232.
[8] Amanat, A., Mirza, M., Javed, A.R., Rizwan, M., Hansra, B.S., Bhatt, M.W., & Alharbi, A. (2022). Deep learning for depression detection from textual data. Electronics, 11(5), 676.
[9] Wani, M.A., ELAffendi, M.A., Shakil, K.A., Imran, A.S., & Abd El-Latif, A.A. (2022). Depression screening in humans with AI and deep learning. IEEE Transactions on Computational Social Systems, 10(3), 1358–1369.
[10] Ibrahimov, R., & Ali, F. (2024). XAI frameworks for transparent and trustworthy mental health AI models. Expert Systems with Applications.
[11] Guo, C., et al. (2023). Mental health detection using text data from online forums with CNN, LSTM, and LIME for interpretability. Applied Intelligence.
[12] Ibrahimov, R., & Ali, F. (2025). Depression X: Knowledge-infused residual attention model with interpretable insights. Expert Systems with Applications.
[13] Hoque, M.E., et al. (2025). Explainable AI for depression detection using SHAP and LIME in educational datasets. Applied Sciences, 15(3).
[14] Nisha, F., Vignesh, R., & Hussain, S. H. (2025). An efficient LSTM based depression detection model from social media. In H. L. Gururaj, F. Flammini, & J. Shreyas (Eds.), Data Science & Exploration in Artificial Intelligence: Proceedings of the First International Conference on Data Science & Exploration in Artificial Intelligence (CODE-AI 2024), Bengaluru, India, 3–4 July 2024 (Vol. 1, pp. 175–180). A A Balkema/CRC Press. doi: 10.1201/9781003587392-23.
[15] Qasim, A., et al. (2025). Depression severity classification using Sentence Transformers on Reddit posts. IEEE Access.
[16] Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). Ablation studies in artificial neural networks. arXiv preprint arXiv:1901.08644.
[17] Tadesse, M.M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in Reddit using NLP and ML techniques. IEEE Access, 7, 44883–44893.
[18] Helmy, M., et al. (2024). Machine learning for depression detection in social media: Arabic and English corpora. Expert Systems with Applications.

[19] Bokolo, B.G., & Liu, Q. (2023). Deep learning-based depression detection from social media: comparative evaluation of ML and transformer techniques. Electronics, 12(21), 4396.
[20] IJACSA. (2025). Depression detection in social media using NLP and RoBERTa-BiLSTM approach. International Journal of Advanced Computer Science and Applications, 16(2).
[21] Al Asad, N., et al. (2024). BERT+BiLSTM pipeline for depression detection in English and Arabic with XAI. Frontiers in Artificial Intelligence.
[22] Thekkekara, J.P., Yongchareon, S., & Liesaputra, V. (2024). An attention-based CNN-BiLSTM model for depression detection on social media text. Expert Systems with Applications, 249, 123834.
[23] Prasad, P., & Jamadagni, H.S. (2023). Depression detection from social media using BiLSTM-CNN hybrid model. Journal of King Saud University — Computer and Information Sciences.
[24] Xin, C., & Zakaria, L. Q. (2024). Integrating BERT with CNN and BiLSTM for explainable detection of depression in social media contents. IEEE Access, 12, 161203–161212. https://doi.org/10.1109/ACCESS.2024.3488081.
[25] Vandana, Marriwala, N., & Chaudhary, D. (2023). A hybrid model for depression detection using deep learning. Measurement: Sensors, 25, 100587.
[26] Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30.
[27] Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of KDD 2016, 1135–1144.
[28] Malhotra, A., & Jindal, R. (2024). SHAP and LIME with BERT-based model for depressive and suicidal behavior analysis. Journal of Biomedical Informatics.
[29] Kokane, S., et al. (2024). Evaluating BERT, XLNet, RoBERTa and DistilBERT for depression detection on Twitter and Reddit. IEEE Access.
[30] Raj, A., et al. (2024). BERT model for depression detection in Reddit posts achieving F1-score exceeding 91%. Human-Centric Intelligent Systems.
[31] Qasim, A., et al. (2025). Transformer-based models for depression severity assessment from social media text. Frontiers in Artificial Intelligence.
[32] Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.
[33] Wang, Z., et al. (2025). A comprehensive evaluation of oversampling techniques for enhancing text classification performance. Scientific Reports, 15.
[34] Garg, M. (2023). Mental health analysis in social media posts: a survey. Archives of Computational Methods in Engineering, 30, 1819–1842.
[35] Taskiran, S. F., Turkoglu, B., Kaya, E., & Asuroglu, T. (2025). A comprehensive evaluation of oversampling techniques for enhancing text classification performance. Scientific Reports, 15, Article 21631. https://doi.org/10.1038/s41598-025-05791-7.
[36] Beniwal, R., & Saraswat, P.A. (2024). Hybrid BERT-CNN approach for depression detection on social media. The Computer Journal, 67(7), 2453–2472.
[37] DABLNet Authors. (2025). Multi-modal deep-attention-BiLSTM based early detection of mental health issues using social media. Scientific Reports.
[38] Bhuiyan, M.R., et al. (2025). CNN-BiLSTM with attention mechanisms for depression symptom classification. Journal of Healthcare Informatics Research.
[39] Li, Y., Zhou, S., & Xu, Y. (2022). RoBERTa-BiLSTM: A hybrid deep learning model for depression severity detection. Journal of Affective Computing, 11, 145–153.
[40] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Stanford CS224N Project Report. [Sentiment140 Dataset].
[41] Baydili, İ., Tasci, B., & Tasci, G. (2025). Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection. Behavioral Sciences, 15(3), 352. https://doi.org/10.3390/bs15030352
[42] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
[43] Alqarni, F., et al. (2025). Emotion-aware RoBERTa enhanced with emotion-specific attention and TF-IDF gating for fine-grained emotion recognition. Scientific Reports, 15.
[44] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. ICLR 2015.
[45] Losada, D.E., Crestani, F., & Parapar, J. (2020). eRisk 2020: Early risk prediction on the internet. In CLEF 2020 Conference and Labs of the Evaluation Forum.
† Reported on single dataset with sentiment-derived labels; no cross-platform validation.