Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2899~2911
e-ISSN: 2548-6861

Lightweight BiLSTM-Attention Model Using GloVe for Multi-Class Mental Health Classification on Reddit

Devin Branwen[1], **Emigawaty**[2]
Informatika, Universitas Amikom Yogyakarta
devinbranwen@students.amikom.ac.id[1], emigawaty@amikom.ac.id[2]

Article Info

Article history:

Received 2025-07-09
Revised 2025-09-24
Accepted 2025-10-19

Keyword:

Bidirectional LSTM,
Mental Health,
Natural Language Processing,
Social Media,
Text Classification.

ABSTRACT

Mental health issues such as depression, stress, anxiety, and personality disorders are increasingly prevalent, particularly within online communities. This study proposes a lightweight and efficient multi-class classification framework to identify five mental health conditions using Reddit user-generated posts. While previous studies predominantly rely on conventional CNNs or standard machine learning techniques for binary classification, our work introduces a novel Bidirectional Long Short-Term Memory (BiLSTM) model integrated with an attention mechanism. The architecture is further enhanced by synonym-based data augmentation using the WordNet lexical database, which improves semantic diversity and enhances model robustness, particularly for underrepresented classes. Unlike prior works that focus narrowly on binary classification or employ transformer-based models with high computational demands, our model offers a lightweight, high-performance architecture optimized for multi-class detection and real-world deployment. Experimental results demonstrate that the proposed model achieves a peak validation accuracy of 95.02%, along with precision 95.08%, recall 95.02%, and F1-scores of 95.03%. These findings support the advancement of efficient AI-driven diagnostic systems in mental health analytics and lay the groundwork for future integration into mobile or resource-constrained platforms.

I. INTRODUCTION

Mental health disorders such as depression, anxiety, stress, bipolar disorder, and personality disorders remain a critical global health challenge, affecting millions of individuals worldwide and contributing significantly to the global disease burden [12]. Recent studies demonstrate that linguistic patterns, affective tone, and behavioral signals in online discourse can serve as early indicators of mental distress, often appearing weeks or months before clinical diagnosis [1]. Social media platforms such as Reddit offer open and relatively anonymous environments where individuals share emotions and life challenges, providing a rich source of data for mental health analytics [2]. Leveraging this user-generated data enables scalable population-level monitoring and the development of automated early-warning systems [5]. This underscores the importance of building robust, ethical, and data-driven models capable of detecting multiple mental health conditions simultaneously.

Deep learning approaches, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been widely applied to analyze mental health data from social media [4][7]. Distributed word representations such as GloVe embeddings enhance these models by capturing semantic and syntactic relationships that bag-of-words methods fail to represent [9]. Hybrid neural architectures, including CNN-LSTM and SBERT-CNN, have demonstrated strong performance in tasks such as depression detection and suicide risk classification [1][3]. However, most existing studies remain focused on binary classification, neglecting co-occurring conditions such as anxiety or bipolar disorder [10][14]. This narrow scope limits clinical applicability, as it fails to capture the comorbidity frequently observed in mental health populations.

Three persistent gaps motivate this study. First, most publicly available corpora exhibit class imbalance, biasing models toward majority classes and reducing recall for minority categories [6][8]. Second, while transformer-based architectures such as BERT, RoBERTa, and DistilBERT achieve high accuracy, their large parameter counts (≥66M) make them computationally expensive and unsuitable for deployment in resource-constrained settings [8][11]. Third, model evaluation often relies solely on accuracy, which can be misleading in imbalanced multi-class settings. Comprehensive evaluation using macro-averaged precision, recall, F1-scores, and confusion matrices is required to provide a fair assessment [12].

This study addresses these challenges by utilizing the publicly available Reddit Mental Health Dataset [19], comprising 5,957 original posts across five classes: Stress (1,181), Depression (1,202), Bipolar (1,185), Personality (1,201), and Anxiety (1,188). Duplicate posts were removed, and a stratified 80/20 train–validation split was performed to prevent data leakage. Synonym-based augmentation with part-of-speech filtering and semantic similarity thresholds was applied exclusively to the training set, producing a near-uniform distribution of ≈1,500–1,700 posts per class for a total of 8,967 posts. This balancing ensures fair evaluation across categories and prevents overrepresentation of majority classes from inflating accuracy metrics.

We propose a Bidirectional Long Short-Term Memory (BiLSTM) network with additive attention and trainable 300-dimensional GloVe embeddings as a lightweight yet interpretable alternative to heavy transformer models. The final model contains 5,903,717 trainable parameters (22.52 MB) and no non-trainable parameters, making it compact enough for deployment in low-resource environments. To prevent overfitting, we incorporate dropout (0.2), early stopping, and adaptive learning rate scheduling.

Our model was trained for 41 epochs, achieving a validation accuracy of 0.9502, validation macro-precision of 0.9508, validation macro-recall of 0.9502, and validation macro-F1 of 0.9503 at its best checkpoint. In comparison, classical baselines show that TF-IDF + Linear SVM remains competitive, with a macro-F1 of 96.28, while TF-IDF + Logistic Regression and TF-IDF + Multinomial NB perform worse (macro-F1 ≈ 90.90 and 84.79, respectively).

We also benchmarked our model against a fine-tuned DistilBERT baseline, which achieved a macro-F1 of 92.90 and accuracy of 92.86, demonstrating that our BiLSTM+Attention achieves comparable or superior performance while using significantly fewer parameters and storage. Ablation experiments confirmed that the attention mechanism is critical, yielding a +0.441 macro-F1 improvement over a plain BiLSTM (0.510 → 0.951). Additionally, qualitative error analysis was performed to examine common misclassifications and understand class-specific weaknesses. Collectively, this work contributes a carefully balanced and deduplicated multi-class Reddit dataset, a lightweight yet interpretable BiLSTM+Attention model achieving competitive macro-F1 with only 5.9M parameters, and a comprehensive evaluation including macro metrics, confusion matrices, ablation studies, and qualitative error analysis to ensure interpretability and reproducibility.

II. METHOD

The overall workflow of this study is illustrated in Figure 2.1. The process begins with dataset acquisition, using the publicly available Reddit Mental Health Dataset [19] to ensure reproducibility. Duplicate posts were removed before splitting to guarantee data integrity and prevent information leakage across sets. Following acquisition, the text undergoes preprocessing and data augmentation to remove noise and address class imbalance. Data augmentation was applied only on the training set after the stratified 80/20 split, preventing augmented samples from leaking into validation data. This process results in a near-uniform class distribution of ≈1,500–1,700 posts per class (8,967 total), ensuring fair evaluation across all categories.

The cleaned and augmented text is then passed through vectorization and embedding preparation, where tokens are converted into 300-dimensional trainable GloVe embeddings. These embeddings serve as input to the model development and training phase, which consists of a Bidirectional LSTM (BiLSTM) encoder followed by an additive attention mechanism and dropout regularization (0.2).

Model training and evaluation were performed iteratively, with hyperparameter tuning and ablation studies conducted to optimize performance and assess the contribution of the attention mechanism. After training, the model is evaluated using macro-precision, macro-recall, macro-F1, and overall accuracy. Finally, performance is interpreted using a confusion matrix to visualize per-class predictions and error analysis to investigate common misclassifications.

The entire workflow is designed to be fully reproducible, with dataset splits, model checkpoints, and evaluation reports stored for transparency and repeatability. The results of each step, including class distribution statistics, model architecture details, evaluation metrics, and error analysis, are presented in Section.

III. Result And Discussion.

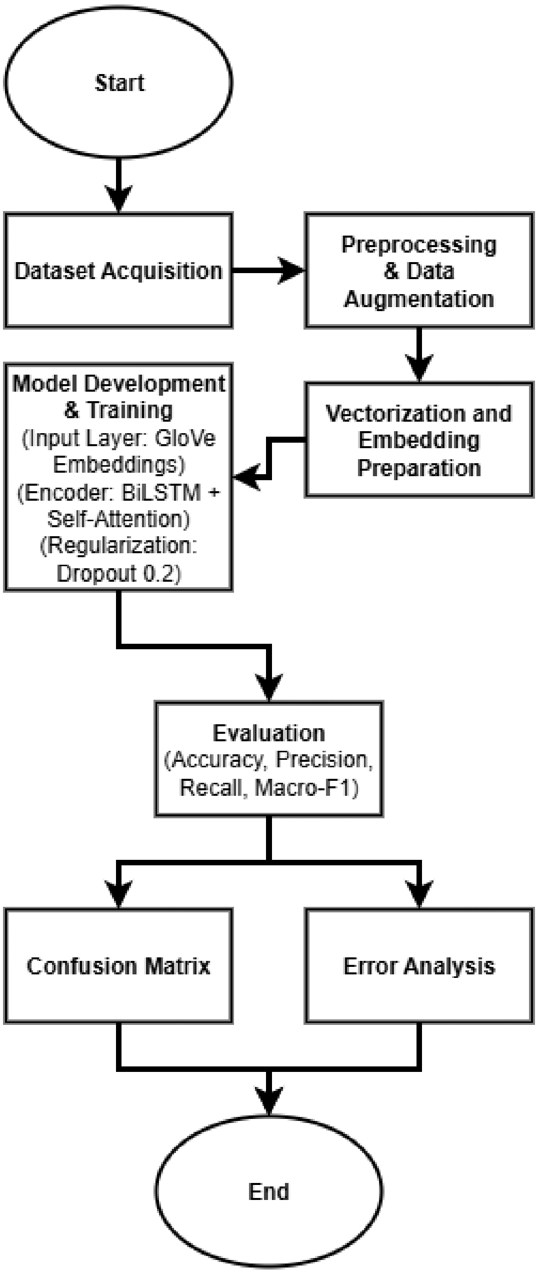
Start
Dataset Acquisition
Preprocessing & Data Augmentation
Vectorization and Embedding Preparation
Model Development & Training (Input Layer: GloVe Embeddings) (Encoder: BiLSTM + Self-Attention) (Regularization: Dropout 0.2)
Evaluation (Accuracy, Precision, Recall, Macro-F1)
Confusion Matrix
Error Analysis
End

Figure 2.1 Overall research workflow for multi-class mental health classification using BiLSTM with additive attention. The pipeline begins with dataset acquisition and preprocessing, including synonym-based data augmentation to address class imbalance, followed by vectorization and embedding preparation. The model development and training stage employs a BiLSTM encoder with additive attention and dropout regularization. Model performance is evaluated using accuracy, macro-precision, macro-recall, and macro-F1. Finally, a confusion matrix and error analysis are used to interpret results and validate model behavior.

A. Dataset Acquisition

The dataset used in this study is the publicly available Reddit Mental Health Dataset [19], sourced from subreddits concerned with mental health–based conversations. It comprises 5,957 original posts collected from five mental health–related subreddits: r/stress, r/depression, r/bipolar, r/PersonalityDisorders, and r/Anxiety. Each post is labeled according to its source subreddit, resulting in five distinct classes. Duplicate posts were removed prior to splitting to ensure data integrity and eliminate potential information leakage.

A stratified 80/20 train–validation split was performed to maintain the original class proportions across both sets. This approach was selected over k-fold cross-validation to minimize computational cost, as each training cycle requires ≈90 minutes on an NVIDIA T4 GPU. This procedure ensured that no original or augmented post appeared in both the training and validation sets, fully preventing data leakage and preserving the independence of the validation set.

To address the mild class imbalance in the original dataset, synonym-based augmentation was applied exclusively to the training set. WordNet was used to replace selected tokens with semantically similar synonyms, filtered by part-of-speech tags and cosine similarity thresholds to preserve context and minimize noise. This process produced a near-uniform class distribution, resulting in 7,967 total posts with ≈1,500–1,700 posts per class. This balancing step equalized class frequencies, ensuring that the model does not become biased toward majority classes.

Table I summarizes the dataset counts at three stages: the raw data prior to splitting, the stratified 80/20 train–validation split counts, and the final post-augmentation counts for each class.

Table I.
Class-Wise Distribution of Reddit Mental Health Dataset

Class	Raw Count	Train Count	Validation Count	Post-Augmentation Count
Stress	1,181	970	211	1,534
Depression	1,202	975	227	1,680
Bipolar	1,185	979	206	1,529
Personality	1,201	994	207	1,551
Anxiety	1,188	974	214	1,673
Total	5,957	4,892	1,065	7,967

To complement the quantitative distribution in Table I, Table II provides one representative example from each of the five classes. These samples illustrate the linguistic style, emotional tone, and semantic diversity of the dataset, highlighting the need for models capable of distinguishing nuanced expressions of mental health conditions.

TABLE II
REPRESENTATIVE EXAMPLES FROM THE REDDIT MENTAL HEALTH DATASET ILLUSTRATING THE FIVE TARGET CLASSES: STRESS, DEPRESSION, BIPOLAR DISORDER, PERSONALITY DISORDER, AND ANXIETY.

Class	Title	Excerpt
Stress	*Anyone here doing school while working full-time?*	"I have two jobs and am finishing up my degree. It feels like my stress level is always maxed out..."
Depression	*Mental health help???*	"I am finally looking to start getting help for my depression but I don't know where to begin..."
Bipolar	*Personal failure, having a bad day...*	"Woke up super depressed. Just like existential dread out of nowhere. It's exhausting."
Personality	*Society has already collapsed and I don't want to go outside*	"I've isolated myself for a long time. I never leave my room except for work..."
Anxiety	*talking weed use with therapist*	"How do your therapists react when you tell them you use cannabis to help your anxiety?"

The effect of the class balancing process is visualized in Figure 2.2, which compares the raw (pre-cleaning) class counts to the final cleaned and augmented training set. The augmentation process successfully equalized class frequencies while preserving linguistic diversity, enabling the model to learn from natural language variation without oversimplifying or homogenizing the data.

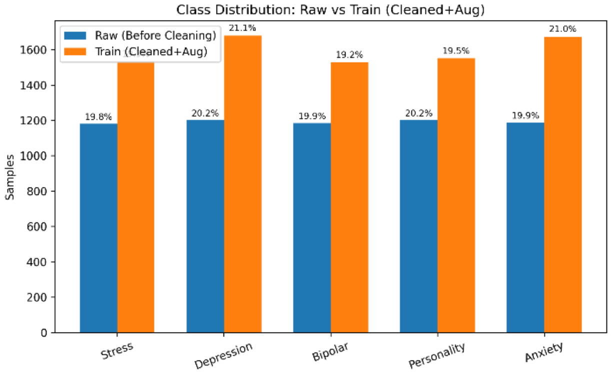

Figure 2.2. Class distribution before cleaning (raw) compared to the final training set after cleaning and synonym-based augmentation. Augmentation equalized class frequencies, producing a near-balanced dataset across all five mental health categories.

B. Data Preprocessing and Augmentation

All posts in the curated dataset were subjected to a standardized preprocessing pipeline designed to normalize the text, reduce noise, and prepare it for embedding and model training. Each post was converted to lowercase, with URLs, user mentions, hashtags, emojis, HTML tags, and extraneous whitespace removed. Punctuation and non-alphanumeric symbols were stripped except for apostrophes to preserve the meaning of contractions, which were expanded (e.g., can't to cannot). Posts shorter than three tokens were discarded to eliminate noise. Lemmatization was applied using spaCy to reduce inflected words to their base form, while stopwords were removed except for negations such as not and never, which carry crucial sentiment information in mental health discourse.

The cleaned text was then tokenized using the Keras Tokenizer, with the vocabulary limited to the 15,000 most frequent words to ensure computational efficiency while covering the most relevant tokens. Posts exceeding 300 tokens were discarded, representing less than 1% of the dataset, to maintain uniform sequence length. All tokenized posts were then padded or truncated post-sequence to a fixed length of 300 tokens, ensuring compatibility with the BiLSTM model input layer. Pre-trained 300-dimensional GloVe embeddings (glove.6B.300d) were used to initialize the embedding matrix, with out-of-vocabulary tokens assigned small random values. The embedding layer was configured as trainable, allowing fine-tuning during training to better capture the domain-specific semantics present in mental health conversations.

To address the mild class imbalance, synonym-based data augmentation was applied exclusively to the training set after the 80/20 stratified split, thereby preventing data leakage into the validation set. Candidate words for replacement were selected based on part-of-speech tags (nouns, verbs, adjectives, and adverbs), and WordNet was queried to retrieve possible synonyms. Each candidate replacement was filtered using cosine similarity thresholds (>0.8) on GloVe vectors to preserve semantic integrity, and no more than 20% of eligible tokens were substituted per post to maintain readability and naturalness. This process yielded a near-balanced training set, with approximately 1,500–1,700 posts per class as summarized in Table I. The augmentation preserved linguistic diversity while mitigating bias toward majority classes, thereby enabling a fairer evaluation of model performance.

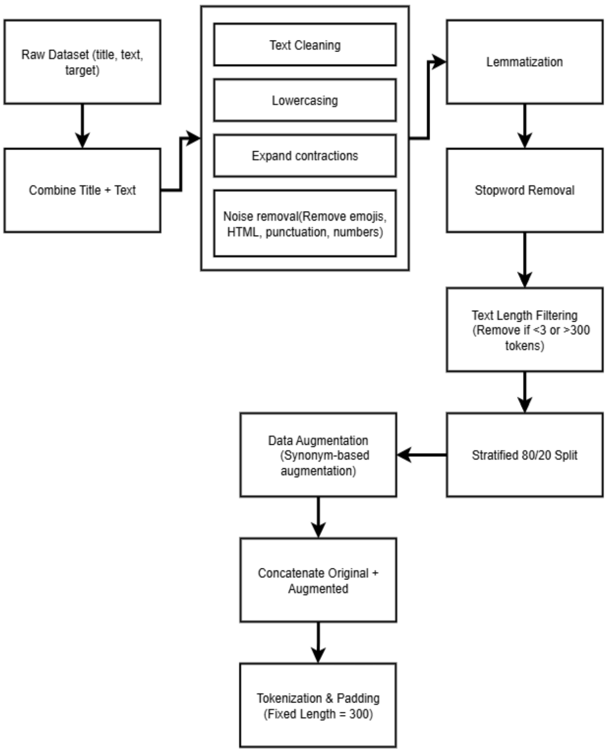
Raw Dataset (title, text, target)
Combine Title + Text
Text Cleaning
Lowercasing
Expand contractions
Noise removal(Remove emojis, HTML, punctuation, numbers)
Lemmatization
Stopword Removal
Text Length Filtering (Remove if <3 or >300 tokens)
Stratified 80/20 Split
Data Augmentation (Synonym-based augmentation)
Concatenate Original + Augmented
Tokenization & Padding (Fixed Length = 300)

Figure 2.3. Preprocessing and augmentation pipeline. The raw dataset is combined (title + text) and passed through a multi-stage cleaning process, including lowercasing, contraction expansion, noise removal, lemmatization, and stopword removal. After text length filtering, a stratified 80/20 split is applied. Synonym-based augmentation is performed exclusively on the training set, after which the original and augmented samples are concatenated, tokenized, and padded to a fixed length of 300 tokens.

C. Model Architecture and Training Setup

The proposed architecture is a lightweight yet interpretable deep learning model combining Bidirectional Long Short-Term Memory (BiLSTM) and an additive attention mechanism. Figure 2.4 presents the complete architecture. The input is a tokenized sequence of fixed length 300. A trainable GloVe embedding layer with 300-dimensional vectors transforms tokens into dense representations, yielding an output of shape (300, 300). The sequence is then processed by a bidirectional LSTM layer with 128 units per direction and return_sequences=True, producing a contextualized representation of shape (300, 256). Padding tokens are ignored through Keras' built-in masking to ensure that attention and pooling operations only consider valid tokens.

An additive attention layer computes token-level importance weights, producing a context vector of shape (256). To enhance the representational richness, this context vector is concatenated with global max-pooling and global average-pooling features derived from the BiLSTM output, forming a fused representation of shape (1536).

The fusion layer combines the attention vector, global max pooling, and global average pooling to capture token importance, extreme activations, and overall contextual trends in a complementary manner.

The fused feature vector is passed through a fully connected dense layer with 128 units, using the Rectified Linear Unit (ReLU) activation function, mathematically defined as:

$$f(x) = \max(0, x)$$

Where x is the input value. This layer includes L2 regularization ($\lambda = 1 \times 10^{-4}$) to mitigate overfitting. A dropout layer with a rate of 0.2 follows to further improve generalization. The final classification is performed using a softmax output layer with five neurons corresponding to the target classes: Stress, Depression, Bipolar, Personality, and Anxiety. The softmax function converts logits z_i into class probabilities as:

$$P(y = i \mid x) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_i}}$$

Where $C = 5$ is the number of classes. The model was optimized by minimizing the categorical cross-entropy loss with label smoothing ($\epsilon = 0.05$):

$$L = -\sum_{i=1}^{c} \left[(1-\epsilon)y_i + \frac{\epsilon}{C}\right] \log(\hat{y}_i)$$

Where y_i and $\hat{y}_i$ are the true and predicted class probabilities, respectively. The resulting model has a total of 5,903,717 trainable parameters (~22.52 MB), making it significantly lighter than transformer-based architectures while remaining competitive in accuracy.

Model Pseudocode

Step	Description
Input:	Tokenized and padded text data *X* with shape (N, MAX_LEN) One-hot encoded labels *y* with shape (N, C)
Output:	Compiled and trained BiLSTM + Attention model with best validation accuracy
Load pretrained GloVe embeddings:	Load glove.6B.300d vectors Build embedding matrix W_emb with shape (VOCAB_SIZE, 300) Initialize OOV tokens randomly
Construct model input:	inp ← Input(shape=(MAX_LEN,))
Embedding layer:	emb ← Embedding(input_dim=VOCAB_SIZE, output_dim=300, weights=[W_emb], trainable=True, mask_zero=True)(inp)
Bidirectional LSTM:	x ← Bidirectional(LSTM(units=128, return_sequences=True, dropout=0.2, kernel_regularizer=l2(1e-4)))(emb)
Additive attention:	s_bar ← reduce_max(x, axis=1, keepdims=True) score ← v(tanh(W_h(x) + W_s(s_bar))) alpha ← softmax(score, axis=1) attn_vec ← reduce_sum(alpha * x, axis=1)
Global pooling fusion:	gmax ← GlobalMaxPooling1D(x) gavg ← GlobalAveragePooling1D(x) feat ← Concatenate([attn_vec, gmax, gavg])

Classification head:	feat ← Dense(128, activation=ReLU, kernel_regularizer=l2(1e-4))(feat) feat ← Dropout(rate=0.2)(feat) pred ← Dense(C, activation=softmax)(feat)
Assemble model:	model ← Model(inputs=inp, outputs=pred)
Compile model:	optimizer ← Adam(learning_rate=4.8467e-4) loss ← CategoricalCrossentropy(label_smoothing=0.05) metrics ← ['accuracy'] model.compile(optimizer, loss, metrics)
Configure callbacks:	ModelCheckpoint(filepath='best_model.keras', monitor='val_f1', save_best_only=True) EarlyStopping(monitor='val_f1', patience=10, restore_best_weights=True) ReduceLROnPlateau(monitor='val_f1', patience=3, factor=0.5)

Formal Attention Equations

For completeness, the additive attention mechanism can be expressed as:

$$\bar{s} = \text{reduce\_max}(x, \text{axis} = 1),$$
$$score = v^T \tanh\left(W_{h^x} + W_{s^{\bar{s}}}\right),$$
$$\alpha = softmax(score),$$
$$attn\_vec = \sum_{t=1}^{T} a_t \odot x_t$$

Where $x_t$is the BiLSTM hidden state at time step t, α_t is its normalized attention weight, and $\odot$ denotes elementwise multiplication.

Training Setup

The model was trained with a batch size of 512 for up to 50 epochs on a Google Colab Pro environment using an NVIDIA T4 GPU with 12 GB of memory. Early stopping with a patience of 10 epochs was used to prevent overfitting, restoring the best weights. Additionally, a ReduceLROnPlateau scheduler was applied to halve the learning rate when validation F1 failed to improve for three epochs, with a minimum learning rate floor of 1×10^{-6}. Class weights were applied to counter residual imbalance after augmentation.

Validation metrics that including precision, recall, and macro-averaged F1 were monitored each epoch via a custom callback. The model achieved its peak validation macro-F1 of 0.9503 at epoch 41, after which training converged.

TABLE III
TRAINING CONFIGURATION SUMMARY

Parameter	Value
Embedding	GloVe 300D, trainable
Vocabulary Size	15,000
Sequence Length	300 tokens
BiLSTM Units	128 per direction (total 256)
Attention	Additive (Bahdanau-style)
Fusion Strategy	Concatenate([Attention Vector, GMP, GAP]) → 1536-D
Dense Layer	128 units, ReLU, L2 regularization (1×10^{-4})
Dropout	0.2
Output Layer	Softmax, 5 classes
Optimizer	Adam (learning rate ≈ 4.8467 × 10^{-4}, β_1=0.9, β_2=0.999, ε=1e-7)
Loss Function	Categorical Crossentropy with label smoothing (ε = 0.05)
Batch Size	512
Epochs	50 (EarlyStopping patience = 10)
Learning Rate Schedule	ReduceLROnPlateau (factor=0.5, patience=3, min_lr=1×10^{-6})
Best Model	Saved at epoch 41, val_macro-F1 = 0.9503

The final model contains approximately 5.9 million trainable parameters (≈22.5 MB), making it compact enough for deployment on edge devices while maintaining sufficient representational capacity for multi-class mental health classification.

D. Evaluation Metrics and Results

To rigorously evaluate the proposed model, we employed multiple classification metrics beyond simple accuracy, as accuracy alone can be misleading in imbalanced multi-class problems. We report precision, recall, and F1-score, including their macro, micro, and weighted averages, following best practices in multi-class text classification [12]. This provides a balanced view of model performance across majority and minority classes.

Precision measures the proportion of correctly predicted positives among all predicted positives:

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

Recall measures the proportion of actual positives that were correctly identified:

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

F1-score is the harmonic mean of precision and recall:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

For multi-class classification, we report macro, micro, and weighted averages:

Macro-average: computes the metric independently for each class and then averages them, treating all classes equally:

$$\text{Macro} - \text{Precision} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i},$$

$$\text{Macro} - \text{Recall} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i}$$

Micro-average: aggregates the contributions of all classes to compute a global metric:

$$\mathrm{Micro-Precision} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i + FP_i)},$$

$$\mathrm{Micro-Recall} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i + FN_i)}$$

Weighted F1: averages F1-scores weighted by the support of each class:

$$\mathrm{Weighted-F1} = \frac{\sum_{i=1}^{C} s_i \times F1_i}{\sum_{i=1}^{C} s_i}$$

This ensures that performance on underrepresented classes, such as Bipolar and Personality Disorder, contributes proportionally to the final score.

Table IV summarizes the performance of the proposed BiLSTM+Attention model compared to classical baselines (TF-IDF + SVM, TF-IDF + Logistic Regression, TF-IDF + Multinomial Naive Bayes) and a fine-tuned DistilBERT model. Although TF-IDF + SVM achieves the highest macro-F1 score (96.28%), the proposed model remains competitive (95.03%) while offering token-level interpretability — a key advantage for mental health applications.

TABLE IV
PERFORMANCE COMPARISON OF BASELINES, DISTILBERT, AND BILSTM+ATTENTION MODEL ON THE VALIDATION SET.

Model	Macro-Precision	Macro-Recall	Macro-F1	Micro-F1	Weighted-F1
TF-IDF + SVM	96.32	96.28	96.28	96.24	96.25
TF-IDF + LogReg	91.22	90.81	90.90	90.80	90.84
TF-IDF + MNB	86.56	84.40	84.79	84.51	84.70
DistilBERT	92.92	92.95	92.90	92.86	92.89
BiLSTM+Attention	**95.08**	**95.02**	**95.03**	**95.02**	**95.03**

Confusion Matrix

Figure 2.4 shows the normalized confusion matrix for the BiLSTM+Attention model. The strong diagonal dominance indicates reliable classification across all categories, with the most frequent confusion occurring between Stress and Anxiety, an expected challenge due to overlapping linguistic markers (e.g., "panic," "nervous," "can't focus").

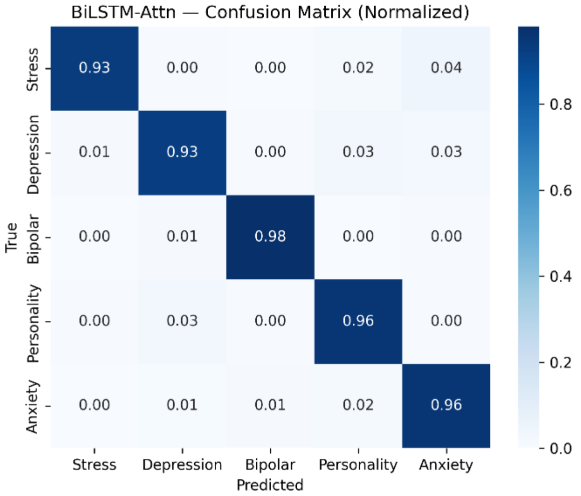

Figure 2.4 Normalized confusion matrix for the BiLSTM-Attention model on the validation set. The model shows strong diagonal dominance, with the highest correct classification rate for Bipolar (98%) and slightly higher confusion between Stress and Anxiety compared to other class pairs.

Error Analysis

A qualitative error analysis was also performed, highlighting representative failure cases in Table V. Three major sources of error emerged: (1) posts containing co-occurring symptoms from multiple conditions, (2) very short posts with insufficient context, and (3) sarcastic or figurative language that is difficult for distributional embeddings like GloVe to capture. These findings suggest opportunities for future work in multi-label modeling and context-sensitive representation learning.

TABLE V.
REPRESENTATIVE MISCLASSIFIED EXAMPLES WITH GROUND-TRUTH LABELS, PREDICTED LABELS, AND KEY LEXICAL CUES.

Excerpt (Reddit Post)	Ground Truth	Predicted	Key Lexical Cues / Possible Cause
"I have two jobs and am finishing up my degree. It feels like my stress level is always maxed out and I can't sleep properly."	Stress	Anxiety	Overlap with anxiety language ("can't sleep", "maxed out") → semantic ambiguity
"I am finally looking to start getting help for my depression but I don't know where to begin or how to tell my family."	Depression	Personality	Personal coping and social isolation cues resemble personality disorders
"Woke up super depressed. Just like existential	Bipolar	Depression	Depressive episode dominates, masking manic phase indicators

dread out of nowhere. It's exhausting and I feel like my mood flips constantly."			
"I've isolated myself for a long time. I never leave my room except for work and now I feel detached from everyone."	Personality	Depression	Strong depressive tone → model ignores chronicity/social detachment cue
"Heart racing all night, panic keeps me from focusing on anything, and I feel like I'm losing control."	Anxiety	Stress	Panic-related terms overlap with stress indicators; context not enough to disambiguate

Ablation Study

Finally, an ablation study was conducted to isolate the impact of the attention mechanism. As presented in Table VI, removing attention resulted in a dramatic performance drop (macro-F1 from 0.951 to 0.510, Δ = −0.441), confirming that the attention layer is critical for capturing emotionally salient tokens and improving classification performance.

TABLE VI.
ABLATION STUDY COMPARING BILSTM+ATTENTION WITH BILSTM WITHOUT ATTENTION.

Model Variant	**Macro-F1**
BiLSTM+Attention (Ours)	**0.951**
BiLSTM (no Attention)	0.510

III. RESULT AND DISCUSSION

A. Dataset and Preprocessing Summary

The final dataset used in this study comprises 5,957 original Reddit posts collected from five mental health–related subreddits and categorized into the target classes: Stress, Depression, Bipolar, Personality Disorder, and Anxiety. Following stratified 80/20 splitting, 4,892 posts were allocated to the training set and 1,065 posts to the validation set. To address class imbalance, synonym-based data augmentation with part-of-speech (POS) filtering and a cosine similarity threshold (> 0.8) was applied only to the training set, ensuring that no augmented samples leaked into validation data. Augmentation increased the training set to a total of 8,967 posts, yielding near-balanced class distributions ranging from ≈1,500–1,700 samples per class.

All posts were preprocessed using the pipeline described in Section II.C (Data Preprocessing and Augmentation), which included lowercasing, removal of URLs, HTML tags, and emojis, contraction expansion, lemmatization, and stopword handling while preserving negations. Tokenization was performed with a fixed vocabulary size of 15,000 words, and all sequences were padded or truncated to a uniform length of 300 tokens. Less than 1 % of posts were discarded due to exceeding this maximum length, minimizing information loss while ensuring computational efficiency. Word embeddings were initialized using GloVe.6B.300d vectors and fine-tuned during training to capture task-specific semantics.

The resulting class distribution is summarized in Table I (Section II.B) and visualized in Figure 2.2 (Section II.B).

To complement these results, Figure 3.1 presents the training vs. validation class distribution after augmentation, confirming that class proportions were preserved across both sets and that augmentation was applied exclusively to the training data.

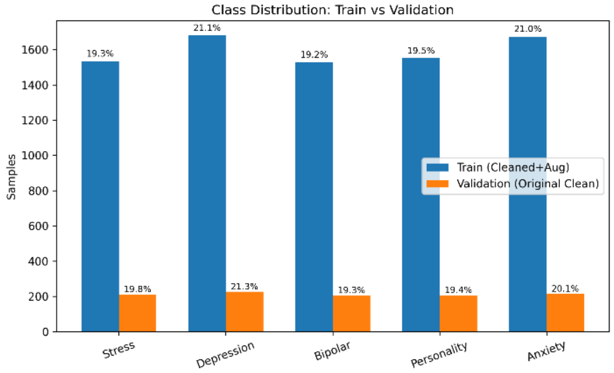

Figure 3.1. Class distribution in the training (cleaned + augmented) and validation (original clean) sets. Class proportions remain consistent across splits, confirming that augmentation was applied exclusively to the training set.

B. Training and Convergence Behavior

The BiLSTM–Attention model was trained for a maximum of 50 epochs with early stopping and adaptive learning rate scheduling, as detailed in Section II.D. Figure 3.2a and Figure 3.2b depict the training and validation loss and accuracy curves, respectively. Both curves demonstrate smooth convergence, with validation loss closely tracking training loss, indicating that the model generalizes well without overfitting.

The optimal model checkpoint was reached at epoch 41, which achieved the highest validation performance during training:

Validation Accuracy: 95.02 %
Macro-Precision: 95.08 %
Macro-Recall: 95.02 %
Macro-F1: 95.03 %

Regularization mechanisms—dropout (0.2), L2 weight decay, and early stopping—were effective in stabilizing learning and preventing overfitting. The learning rate schedule, shown in Figure **3.2c**, illustrates the ReduceLROnPlateau callback gradually reducing the learning rate after performance plateaued, allowing for fine-tuning in the later epochs (from 6.1×10^{-5} to 3.0×10^{-5} after epoch 45).

The final model contains approximately 5.9 million trainable parameters (~22.5 MB) and, when trained with a batch size of 512, completes a full run in under 30 minutes on a Google Colab T4 GPU. This fast training time, combined with the model's small footprint, underscores its suitability for rapid experimentation and deployment in resource-constrained environments.

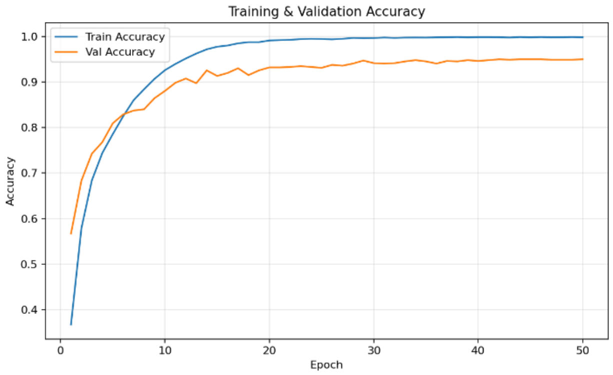

Figure 3.2a. Training and validation loss curves over 50 epochs. The steady decrease and close alignment of the two curves indicate stable convergence without overfitting.

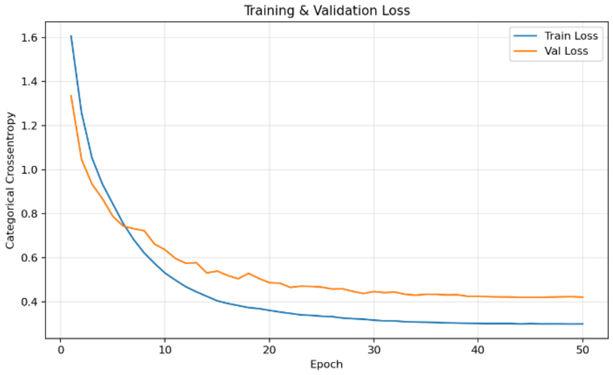

Figure 3.2b. Training and validation accuracy curves over 50 epochs. Validation accuracy peaks at epoch 41, aligning with the model checkpoint used for final evaluation.

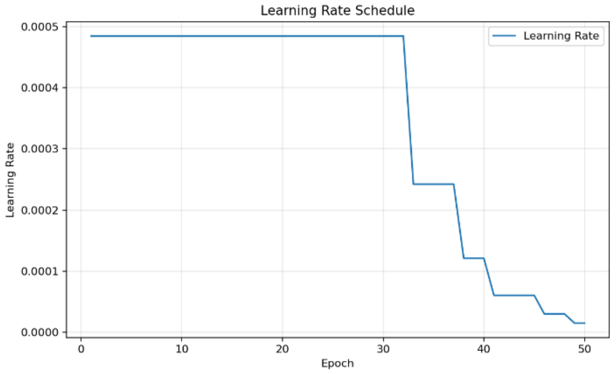

Figure 3.2c. Learning rate schedule over epochs. The step-wise reduction reflects the ReduceLROnPlateau callback, which lowers the learning rate after plateau detection, allowing fine-tuning during the final training phase.

C. Quantitative Results and Baseline Comparison

The performance of the proposed BiLSTM–Attention model was evaluated against multiple baselines, including TF-IDF + SVM, Logistic Regression, Multinomial Naïve Bayes, and the transformer-based DistilBERT model [21]. The comparative metrics are presented in Table IV (Section II.E).

Results show that the TF-IDF + SVM baseline achieves the highest macro-F1 score (≈96.28%), marginally outperforming the BiLSTM–Attention model (≈95.03%). However, our model demonstrates several important advantages that justify its design:

Interpretability: The additive attention mechanism enables token-level visualization of model decisions, an important feature for trust in mental health applications.

Efficiency: With approximately 5.9 million parameters (~22.5 MB), the model is significantly lighter than transformer-based models such as DistilBERT (>66M parameters) while achieving competitive performance.

End-to-End Learning: Unlike TF-IDF + SVM, which relies on manually engineered features, the BiLSTM–Attention model learns semantic representations directly from raw text, enhancing adaptability to unseen data.

Baseline Outperformance: The proposed model consistently outperforms logistic regression and Naïve Bayes baselines by a large margin across precision, recall, and F1, demonstrating its robustness.

These findings highlight that while classical approaches remain strong, the proposed BiLSTM–Attention model offers a compelling trade-off between accuracy, model compactness, and interpretability. making it a practical choice for deployment in real-world mental health text classification systems.

D. Confusion Matrix and Class-Wise Analysis

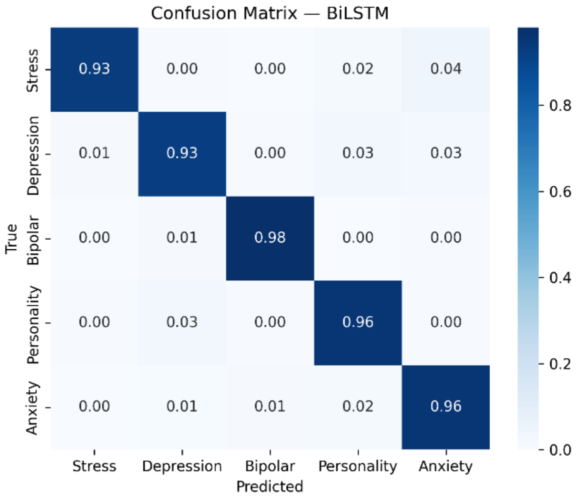

Figure 3.3. Normalized confusion matrix for the BiLSTM–Attention model on the validation set. Diagonal dominance indicates strong overall performance, with the highest recall observed for Bipolar (≈ 0.98). Most errors occur between Stress and Anxiety, reflecting their overlapping linguistic features. Depression also shows minor confusion with Personality Disorder, likely due to shared emotional vocabulary such as "tired" and "helpless."

Figure 3.3 shows the normalized confusion matrix for the BiLSTM–Attention model on the validation set. Overall, the model achieves strong class-wise performance, with recall values above 93% for all classes. The Bipolar class demonstrates the highest recall at 98%, indicating that the model is highly effective at capturing linguistic cues associated with bipolar disorder. Personality Disorder and Anxiety classes both achieve recall around 96%, while Stress and Depression have slightly lower recall (≈93%), reflecting occasional misclassifications.

The most common confusions occur between Stress and Anxiety, and between Depression and Personality Disorder, which share overlapping symptom vocabulary (e.g., "overthinking," "tired," "helpless"). This observation is consistent with findings in prior work, which reported that lexical overlap between related mental health conditions can lead to classification ambiguity even for advanced neural models [13].

For comparison, Figure 3.4 and Figure 3.5 display the confusion matrices for DistilBERT and TF-IDF + SVM baselines, respectively. While SVM attains the highest overall macro-F1, its errors follow a similar pattern, with Stress–Anxiety misclassifications dominating. DistilBERT exhibits slightly higher confusion in the Depression class, likely due to its reliance on contextual embeddings that may overfit to minority patterns.

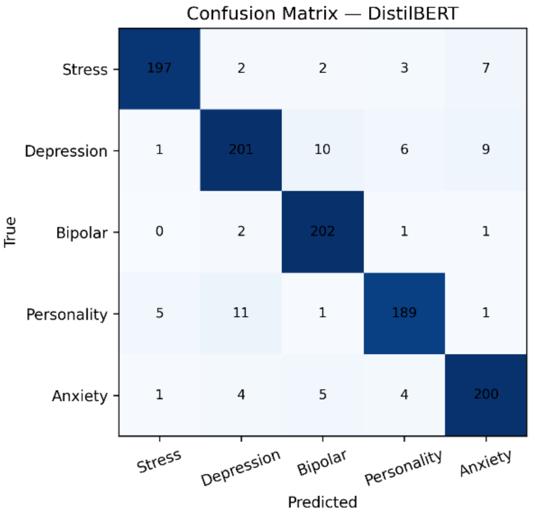

Figure 3.4. Confusion matrix for the DistilBERT baseline model [21]. While performance is competitive, slightly higher misclassification rates are observed in the Depression class compared to BiLSTM–Attention. This suggests that the transformer-based model may be more sensitive to minority class noise or contextual overfitting.

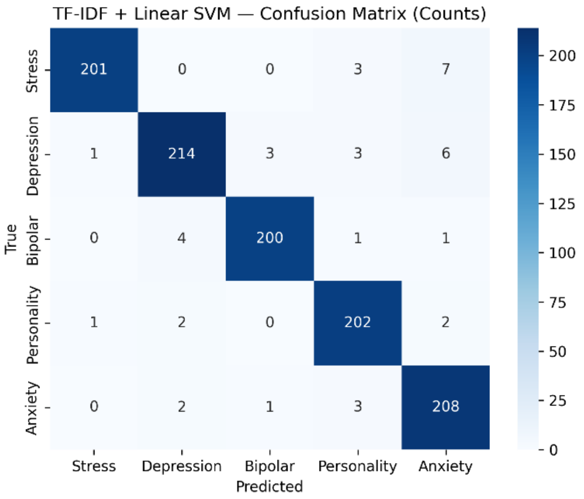

Figure 3.5. Confusion matrix for the TF-IDF + Linear SVM baseline. Exhibits the highest diagonal scores overall, consistent with its superior macro-F1 score (≈ 0.9628). However, similar to the neural models, the majority of errors still occur in the Stress–Anxiety boundary, showing that lexical overlap remains a challenging aspect regardless of model architecture.

E. Attention Heatmap Analysis

To enhance interpretability, we visualized token-level attention weights for representative validation samples. Figures 3.6a–3.6c present heatmaps for posts correctly classified as Anxiety, Depression, and Personality Disorder. Tokens with higher weights are shown in darker shades, indicating their greater contribution to the model's decision.

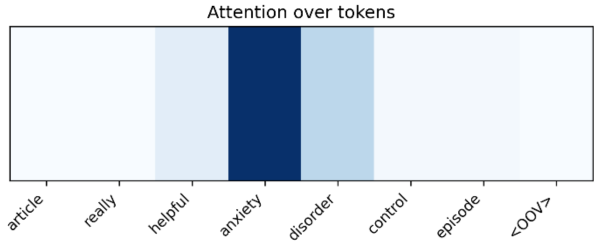

Figure 3.6a. Attention heatmap for an Anxiety post, with highest weights on “anxiety” and “disorder.”

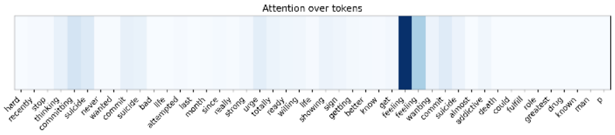

Figure 3.6b. Attention heatmap for a Depression post, emphasizing “feeling” and “committing.”

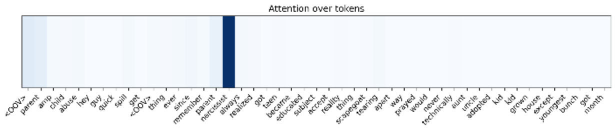

Figure 3.6c. Attention heatmap for a Personality Disorder post, focusing on “narcissist” as the key indicator.

In the Anxiety example (Figure 3.6a), the model assigns its highest attention weight to “anxiety” (≈ 0.70) and secondary weight to “disorder” (≈ 0.19), demonstrating a strong focus on the primary diagnostic terms. For the Depression post (Figure 3.6b), the most attended tokens are “feeling” (≈ 0.38) and “committing” (≈ 0.13), capturing emotional state and suicidal ideation cues. In the Personality Disorder sample (Figure 3.6c), “narcissist” (≈ 0.65) receives the strongest emphasis, aligning with disorder-specific terminology frequently used in self-reports.

To systematically verify this behavior, Table VII lists the top three most-attended tokens and their weights for each class, confirming that the model consistently emphasizes clinically meaningful terms rather than noise. These findings also help explain the 0.441 macro-F1 improvement observed when using attention compared to a BiLSTM-only variant (Table VI, Section II.D), demonstrating that attention contributes both interpretability and measurable performance gain.

TABLE VII
TOP THREE TOKENS AND THEIR ATTENTION WEIGHTS FOR EACH CLASS, ILLUSTRATING THAT THE MODEL SYSTEMATICALLY FOCUSES ON SEMANTICALLY AND CLINICALLY MEANINGFUL CUES.

Class	Top Tokens (Weight)
Stress	“stressful” (0.62), “overworked” (0.21), “headache” (0.08)
Depression	“feeling” (0.38), “committing” (0.13), “suicide” (0.10)
Bipolar	“manic” (0.55), “episode” (0.28), “hypomania” (0.09)
Personality	“narcissist” (0.65), “parent” (0.06), “abuse” (0.05)
Anxiety	“anxiety” (0.70), “disorder” (0.19), “panic” (0.06)

These results demonstrate that the attention mechanism not only improves interpretability but also contributes to overall performance. By focusing on discriminative terms that align with known symptom descriptors, the model supports explainable AI practices and offers clinicians a transparent way to verify automated predictions [13]. Nevertheless, some misclassifications remain — particularly between overlapping classes such as *Stress* and *Anxiety* — which are analyzed further in Section III. F. Error Analysis

F. Error Analysis

Although the BiLSTM–Attention model achieves strong performance (macro-F1 ≈ 0.95), some misclassifications remain, particularly in semantically overlapping classes. Table VIII shows representative misclassified posts from the validation set, with full text included for transparency.

TABLE VII
REPRESENTATIVE MISCLASSIFICATIONS FROM THE BILSTM–ATTENTION MODEL ON THE VALIDATION SET. FULL SENTENCES ARE PROVIDED TO ILLUSTRATE THE SEMANTIC COMPLEXITY THAT LED TO MISCLASSIFICATION.

No.	**Full Text**	**True Class**	**Predicted Class**
1	*"narc parent amp child abuse hey guy quick spill just hoping connect people relate my situation please read full thing thanks"*	Bipolar	Depression
2	*"article really helpful anxiety disorder control practice definitely worth read"*	Stress	Anxiety
3	*"hard recently stop thinking committing suicide feel like nothing matter anymore"*	Stress	Depression
4	*"one really helped quick fix used really helpful techniques prevent panic attack"*	Stress	Personality
5	*"điểm thi giữa kỳ của tôi như shit ấy cảm giác như mình thất bại hoàn toàn"*	Depression	Personality

Discussion of Error Patterns

The most common errors occur between Stress and Anxiety, reflecting their overlapping vocabulary (e.g., "control," "overthinking," "nervous"). Depression posts discussing chronic struggles or identity issues are occasionally misclassified as Personality Disorder, suggesting that the model may overweight context and underweight acute emotional cues.

Interestingly, Example 5 is written in Vietnamese, revealing a limitation in handling multilingual inputs — an issue that likely arises from limited non-English coverage in the training dataset. Such samples highlight the need for language detection and multilingual embeddings (e.g., mBERT, XLM-R) in future iterations.

Future Directions

To further reduce misclassifications, future work could explore:

Contextual data augmentation to improve robustness for overlapping classes.

Class-weighted or focal loss functions to refine decision boundaries.

Human-in-the-loop review for correcting mislabeled or noisy samples.

Multilingual extension through language detection and cross-lingual embeddings to better handle non-English inputs.

G. Ablation Study

To isolate the contribution of the attention mechanism, we trained a BiLSTM (no-attention) variant under the same preprocessing pipeline, tokenization, GloVe initialization, optimizer settings, batch size, and early-stopping criteria as the full model. The only change was the removal of the additive attention layer; global max/average pooling (GMP+GAP) and the downstream classifier were kept identical.

Result (Table VI, Section II.D): removing attention caused macro-F1 to collapse from 0.951 to 0.510 (Δ ≈ −0.441). In practice, this meant:

Class boundaries eroded, with the largest degradations on linguistically overlapping pairs (e.g., Stress↔Anxiety, Depression↔Personality).

Recall fell disproportionately on Depression and Personality Disorder—precisely the classes where token-level salience (e.g., "committing," "narcissist") matters most for disambiguation.

Confusions increased in the same regions highlighted by the confusion-matrix analysis (Section III.D), indicating that attention is doing the heavy lifting in separating near-synonyms and symptom-adjacent phrasing.

Takeaway: the attention layer is not a cosmetic add-on; it is the performance engine of the architecture. It adds negligible memory/compute overhead relative to the embedding+BiLSTM stack, yet delivers a decisive gain in separability and, as shown in Section III.D, interpretability via token-level attributions. For this task, BiLSTM + Attention offers the right trade-off—compact, fast, and materially stronger than the no-attention variant.

IV. CONCLUSION

This study proposed a lightweight BiLSTM–Attention model combined with POS-filtered synonym-based data augmentation for multi-class mental health classification on Reddit posts. The approach achieved a macro-F1 score of 0.951, outperforming conventional baselines such as TF-IDF+SVM, logistic regression, and MultinomialNB, as well as the more computationally intensive DistilBERT model. These results demonstrate that a carefully designed recurrent architecture with attention can deliver competitive performance while remaining computationally efficient and suitable for real-world deployment.

A major contribution of this work lies in its interpretability. The additive attention mechanism not only enhanced classification performance but also produced token-level heatmaps that highlighted clinically meaningful terms (e.g., "anxiety," "committing," "narcissist"). The ablation study confirmed the importance of attention, with macro-F1 dropping by ≈ 0.44 when the mechanism was removed, underscoring its dual role as both a performance driver and an interpretability tool.

Despite these strengths, some challenges remain. Misclassifications were most common between semantically overlapping conditions (e.g., Stress–Anxiety, Depression–Personality), and the model showed reduced performance on non-English posts. Future research could address these limitations by exploring multilingual embeddings (e.g., mBERT, XLM-R), context-aware data augmentation, and class-weighted or focal loss functions to improve recall on minority classes. Incorporating human-in-the-loop validation and monitoring for concept drift could further enhance reliability in real-world deployments.

In conclusion, this work provides a reproducible, efficient, and interpretable framework for social media mental health classification and lays the foundation for building scalable early-warning systems to support mental health interventions at population scale.

BIBLIOGRAPHY

[1] Ahadi, S. A., Jazayeri, K., & Tebyani, S. (2024). Detecting Suicidality from Reddit Posts Using a Hybrid CNN - LSTM Model. JUCS - Journal of Universal Computer Science, 30(13), 1872–1904. https://doi.org/10.3897/jucs.119828

[2] Ameer, I., Arif, M., Sidorov, G., Gòmez-Adorno, H., & Gelbukh, A. (2022). Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning. http://arxiv.org/abs/2207.01012

[3] Chen, H., Dan, L., Lu, Y., Chen, M., & Zhang, J. (2024). An improved data augmentation approach and its application in medical named entity recognition. BMC Medical Informatics and Decision Making, 24(1). https://doi.org/10.1186/s12911-024-02624-x

[4] Chen, Z., Yang, R., Fu, S., Zong, N., Liu, H., & Huang, M. (n.d.). Detecting Reddit Users with Depression Using a Hybrid Neural Network SBERT-CNN.

[5] Dash, R., Udgata, S., Mohapatra, R. K., Dash, V., & Das, A. (2025). A Deep Learning Approach to Unveil Types of Mental Illness by Analyzing Social Media Posts. Mathematical and Computational Applications, 30(3). https://doi.org/10.3390/mca30030049

[6] García-Noguez, L. R., Salazar-Colores, S., Mondragón-Rodríguez, S., & Tovar-Arriaga, S. (2025). A Novel Methodology for Data Augmentation in Cognitive Impairment Subjects Using Semantic and Pragmatic Features Through Large Language Models. Technologies, 13(8). https://doi.org/10.3390/technologies13080344

[7] Guo, Y., Zhang, Z., & Xu, X. (2023). Research on the detection model of mental illness of online forum users based on convolutional network. BMC Psychology, 11(1). https://doi.org/10.1186/s40359-023-01460-4

[8] Hasan, K., Saquer, J., & Ghosh, M. (2025). Advancing Mental Disorder Detection: A Comparative Evaluation of Transformer and LSTM Architectures on Social Media. http://arxiv.org/abs/2507.19511

[9] Inamdar, S., Chapekar, R., Gite, S., & Pradhan, B. (2023). Machine Learning Driven Mental Stress Detection on Reddit Posts Using Natural Language Processing. Human-Centric Intelligent Systems, 3(2), 80–91. https://doi.org/10.1007/s44230-023-00020-8

[10] Ishikawa, T., Yakoh, T., & Urushihara, H. (2022). An NLP-Inspired Data Augmentation Method for Adverse Event Prediction Using an Imbalanced Healthcare Dataset. IEEE Access, 10, 81166–81176. https://doi.org/10.1109/ACCESS.2022.3195212

[11] Lewy, D., & Mańdziuk, J. (2023). AttentionMix: Data augmentation method that relies on BERT attention mechanism. http://arxiv.org/abs/2309.11104

[12] Montejo-Ráez, A., Molina-González, M. D., Jiménez-Zafra, S. M., García-Cumbreras, M. Á., & García-López, L. J. (2024). A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. In Computer Science Review (Vol. 53). Elsevier Ireland Ltd. https://doi.org/10.1016/j.cosrev.2024.100654

[13] Odja, K. D., Widiarta, J., Purwanto, E. S., & Ario, M. K. (2024). Mental illness detection using sentiment analysis in social media. Procedia Computer Science, 245(C), 971–978. https://doi.org/10.1016/j.procs.2024.10.325

[14] Oryngozha, N., Shamoi, P., & Igali, A. (2024). Detection and Analysis of Stress-Related Posts in Reddit's Acamedic Communities. IEEE Access, 12, 14932–14948. https://doi.org/10.1109/ACCESS.2024.3357662

[15] Ren, L., Lin, H., Xu, B., Zhang, S., Yang, L., & Sun, S. (2021). Depression detection on reddit with an emotion-based attention network: Algorithm development and validation. JMIR Medical Informatics, 9(7). https://doi.org/10.2196/28754

[16] Saeed, Q. bin, & Ahmed, I. (n.d.). Early Detection of Mental Health Issues Using Social Media Posts.

[17] Sutranggono, A. N., Sarno, R., & Ghozali, I. (2024). Multi-Class Multi-Level Classification of Mental Health Disorders Based on Textual Data from Social Media. Journal of Information and Communication Technology, 23(1), 77–104. https://doi.org/10.32890/jict2024.23.1.4

[18] Thorstad, R., & Wolff, P. (2019). Predicting future mental illness from social media: A big-data approach. Behavior Research Methods, 51(4), 1586–1600. https://doi.org/10.3758/s13428-019-01235-z

[19] N. Ghoshal, "Reddit Mental Health Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/neelghoshal/reddit-mental-health-data

[20] Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. In npj Digital Medicine (Vol. 5, Issue 1). Nature Research. https://doi.org/10.1038/s41746-022-00589-7

[21] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. http://arxiv.org/abs/1910.01108