International Journal of Innovative Research in Engineering
Volume 7, Issue 2 (March - April 2026), PP: 152-164.
https://www.doi.org/10.59256/ijire.20260702020
www.theijire.com
ISSN No: 2582-8746

Multi-Class Mental Health Detection With LSTM and BilLSTM Models

Dondapati Sasi Prasanna[1], Suneel Kumar Duvvuri[2]
[1]Student, M.Sc Computer Science, Government College (Autonomous), Rajahmundry, Andhra Pradesh, India.
[2]Assistant Professor, Department of Computer Science, Government College (Autonomous), Rajahmundry, Andhra Pradesh, India.

How to cite this paper:
Dondapati Sasi Prasanna[1], Suneel Kumar Duvvuri[2] "Multi-Class Mental Health Detection With LSTM and BilLSTM Models", IJIRE-V7I2-152-164.

Abstract: *Mental health disorders have become a major global concern, affecting millions of individuals worldwide. Traditional diagnostic methods are often time-consuming, reactive, and limited by accessibility and social stigma. With the rise of social media, users increasingly express their emotions through textual content, creating opportunities for early detection of mental health issues using computational techniques. This research focuses on developing a deep learning-based framework for multi-class mental health detection using Natural Language Processing (NLP). The study utilizes the Mental Distress (2026) dataset, which consists of social media text categorized into five classes: Depressed, Anxious, Frustrated, Suicidal, and Others. Unlike traditional binary classification approaches, this research addresses a more complex multi-class classification problem, where emotional states often overlap linguistically and semantically. Advanced preprocessing techniques, including text normalization, tokenization, stop word removal, and sequence padding, are applied. The text is converted into numerical form using embedding layers to capture semantic relationships. Two models, Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), are implemented and compared. The BiLSTM model, with its bidirectional learning capability, provides better contextual understanding. The models are evaluated using Stratified 5-Fold Cross-Validation to ensure reliability and robustness. Performance metrics such as accuracy, precision, recall, and F1-score are used for evaluation. Experimental results demonstrate that the BiLSTM model outperforms the LSTM model, achieving an accuracy of approximately 89%, compared to 86% for LSTM. The improvement is attributed to the model's ability to capture bidirectional context and resolve lexical ambiguities in emotional expressions. This research demonstrates the effectiveness of deep learning techniques in accurately detecting mental health conditions from textual data and highlights their strong potential for real-world applications. These include early warning systems for identifying psychological distress, AI-driven mental health monitoring tools, and digital support platforms that can assist individuals and healthcare professionals in timely intervention and decision-making.*

Keywords: *Mental Health Detection, NLP, Deep Learning, LSTM, BiLSTM, Social Media Analysis, Multi-Class Classification, Text Mining, Emotion Detection.*

I.INTRODUCTION

1.1 Global Mental Health Crisis: Mental health disorders have become one of the most pressing challenges in modern healthcare systems worldwide [1],[2],[3]. According to global health organizations, millions of individuals suffer from conditions such as depression, anxiety, and stress-related disorders, affecting their quality of life and overall productivity [4]. These disorders not only impact individuals but also place a significant burden on families, communities, and healthcare infrastructures. Despite increasing awareness, mental health issues continue to rise due to factors such as urbanization, lifestyle changes, and social isolation.

A major concern in addressing mental health is the widespread treatment gap, where a large proportion of affected individuals remain undiagnosed or untreated. This gap is primarily caused by limited access to healthcare services, social stigma associated with mental illness [5], and a shortage of trained mental health professionals. Traditional diagnostic approaches are often reactive in nature, focusing on treatment after symptoms become severe. As a result, there is a critical need for proactive and scalable solutions that can enable early detection and intervention.

1.2 Social Media as an Early Warning System: Social media platforms have transformed into digital spaces where individuals openly express emotions and psychological states [6],[7]. These platforms provide real-time, unfiltered data that can be analysed to detect early signs of mental distress [8]

The application of Machine Learning (ML) and Natural Language Processing (NLP) techniques enables automated

analysis of this large-scale unstructured data [9],[10]. By processing textual content, these techniques can detect linguistic cues, emotional tone, and sentiment patterns associated with mental health conditions. This makes social media a powerful early warning system, allowing researchers and practitioners to identify potential risks such as depression, anxiety, and suicidal ideation before they escalate into critical conditions.

1.3 Problem Statement: Most existing research in mental health detection focuses primarily on binary classification problems [11],[12], such as distinguishing between depressed and non-depressed individuals. While such approaches provide a basic understanding, they fail to capture the complexity and diversity of human emotions. Mental health conditions often exist on a spectrum, where individuals may simultaneously experience multiple emotional states, making binary classification insufficient for real-world applications.

This research focuses on multi-class classification of textual data into emotional states such as depressed, anxious, frustrated, suicidal, and others. It addresses challenges like lexical ambiguity and contextual dependency, where meanings vary based on word usage and surrounding context. To overcome these issues, advanced deep learning models are required to effectively capture complex semantic relationships in text.

1.4 Research Motivation: The motivation behind this study is the urgent need for accessible and efficient solutions for early mental health detection, as traditional healthcare systems often face delays in diagnosis and treatment. By leveraging advancements in artificial intelligence and data analytics, automated systems can help identify mental health issues at an early stage. Additionally, the growing availability of digital data enables scalable, real-time solutions that reduce the burden on healthcare professionals and improve diagnostic accuracy. Ultimately, this approach aims to provide timely support, enable early intervention, and enhance overall well-being by reducing the severity of mental health conditions.

1.5 Research Objectives

- The study aims to develop a robust deep learning model for multi-class mental health detection using textual data.
- It focuses on designing neural networks to accurately classify different emotional states.
- A key objective is to compare Long Short-Term Memory and Bidirectional LSTM models for better contextual understanding.
- The research seeks to improve accuracy through preprocessing, feature extraction, and handling class imbalance and lexical overlap.
- It aims to provide a scalable framework for real-world applications like early warning systems and digital mental health monitoring.

This research proposes a novel deep learning framework for multi-class mental health detection using social media text, classifying emotions into five categories: Depressed, Anxious, Frustrated, Suicidal, and Others. Unlike traditional binary approaches, it uses a stacked Bidirectional Long Short-Term Memory model to capture deeper contextual and sequential information. The study addresses lexical ambiguity and class imbalance through advanced preprocessing, negation-aware stop word handling, and class weighting to improve performance. Overall, it integrates multi-class classification with stacked BiLSTM and context-aware techniques to enhance accuracy and reliability in mental health detection systems.

II.LITERATURE REVIEW

The field of mental health detection has undergone significant transformation over the past decade, evolving from traditional machine learning techniques to more sophisticated deep learning approaches. Initially, research in this domain relied heavily on statistical and rule-based models that required manual feature engineering. While these methods provided a foundation for text classification tasks, they were limited in their ability to understand the complexity of human language, particularly in the context of emotional and psychological expression.

With improved computational power and large-scale datasets, deep learning models have become widely used in mental health analysis. They automatically extract features from raw text and capture complex patterns in language. These neural network-based approaches enhance the accuracy and reliability of mental health detection, especially in social media data where emotions are subtle and context-dependent.

2.1 Traditional Approaches: Traditional machine learning approaches played a crucial role in the early development of text classification systems. Algorithms such as Naïve Bayes [13], Support Vector Machines (SVM) [14], and Random Forest [15] were widely used due to their simplicity and effectiveness. Naïve Bayes, for instance, is a probabilistic classifier that assumes independence between features, making it computationally efficient but limited in capturing contextual relationships between words. Similarly, SVM is known for its high accuracy in classification tasks but fails to consider the sequential nature of language.

Algorithm	Advantages	Limitations
Naïve Bayes	Simple, fast, efficient	Ignores context and word relationships
SVM	High accuracy, effective	Cannot capture sequence information
Random Forest	Robust, reduces overfitting	No sequential learning capability

Table 1: Traditional Machine Learning Techniques

Random Forest, an ensemble learning method, improves classification by combining multiple decision trees. However, it cannot capture word order or sentence structure. It relies on feature extraction methods like Bag-of-Words and TF-IDF, which ignore semantic meaning and word relationships. Consequently, they struggle to accurately interpret complex linguistic patterns such as sarcasm, negation, and contextual dependencies [16].

2.2 Deep Learning Approaches: Deep learning approaches have significantly improved the performance of text classification tasks by addressing the limitations of traditional methods. These models improve classification accuracy. Recent studies have explored advanced deep learning models such as transformer-based and hybrid architectures for mental health detection [17], [18], [19] [20], [21] by addressing the limitations of traditional methods. Convolutional Neural Networks (CNNs) are effective in capturing local features and identifying important patterns within text, such as key phrases and n-grams [22],[23] However, CNNs are limited in capturing long-range dependencies, which are essential for understanding the overall meaning of a sentence.

Recurrent Neural Networks (RNNs) were introduced to handle sequential data, allowing models to process text in order [24]. Despite their advantages, RNNs suffer from the vanishing gradient problem, which limits their ability to retain long-term dependencies. Long Short-Term Memory (LSTM) networks address this issue by incorporating memory cells and gating mechanisms, enabling them to retain relevant information over long sequences [25],[26]. Bidirectional LSTM (BiLSTM) further enhances this capability by processing data in both forward and backward directions, Recent approaches also incorporate multimodal data such as text, speech, and images for improved detection [27], [28], [29]. Providing a more comprehensive understanding of context and improving classification accuracy [30].

Author/references	Year	Model / Approach	Dataset	Key Findings	Research Gap
De Choudhury et al. [4]	2013	Statistical + SVM	Twitter	Predicted depression before clinical diagnosis	Ignores word order (Bow limitation)
Coppersmith et al. [21]	2015	Linguistic Analysis	Social media	Identified psychological signals in text	No deep learning context modeling
Yates et al. [31]	2017	RNN/LSTM	Reddit	Improved sequential understanding	Limited long-term context
Orabi et al [24]	2018	CNN + LSTM	Social media	Combined local + sequential features	High complexity
Trotzek et al. [25]	2020	Neural + NLP	Reddit	Early depression detection	Limited scalability
Chancellor et al. [31]	2020	DL Survey	Multiple datasets	Reviewed predictive mental health systems	Lack of multi-class focus
Zeberga et al. [28]	2022	BiLSTM + BERT	Social media	Improved semantic understanding	Resource intensive
Singh et al. [29]	2022	LSTM	Text data	Achieved ~90% accuracy	No bidirectional context
Huda et al. [30]	2024	LSTM vs BiLSTM	Social media	BiLSTM outperformed LSTM	No multi-class analysis
Saeed & Cha [32]	2025	BiLSTM + Multimodal	Reddit	Higher F1-score (73.76%)	High complexity
Hasan et al. [27]	2025	LSTM vs Transformer	Reddit	Transformer performed best	High computational cost
Recent Study [26]	2026	BiLSTM + Attention	Multimodal	Improved feature fusion	Complex training

Table 2: Authors Summary Table for Mental Health Detection using Deep Learning

From Table 2, it is evident that significant progress has been made in mental health detection using machine learning and deep learning techniques. Early approaches such as Naïve Bayes and SVM were effective for basic classification but failed to capture contextual and sequential dependencies in text.

2.3 Research Gap: Despite the advancements in machine learning and deep learning techniques, several challenges remain in the field of mental health detection. One of the major limitations is the focus on binary classification in most existing studies [33], [32], where data is categorized into only two classes. in most existing studies, where data is categorized into only two classes. This approach oversimplifies the complexity of mental health conditions, which often involve multiple overlapping emotional states. As a result, there is a need for multi-class classification models that can better represent real-world scenarios.

A key challenge is handling overlapping emotions and contextual ambiguity in text, where meanings vary by context, making sentiment interpretation difficult. Additionally, advanced deep learning approaches like hybrid and attention-based models are still limited in mental health detection. Addressing this requires models that better capture semantic relationships and contextual dependencies.

Challenge	Description
Multi-class Classification	Most studies focus only on binary classification
Overlapping Emotions	Difficulty in distinguishing similar emotional states
Contextual Ambiguity	Words change meaning based on context
Limited Advanced Models	Lack of hybrid and attention-based approaches

Table 3: Identified Research Gaps

III.METHODOLOGY

3.1 Dataset Description: The present study utilizes the Mental Distress (2026) dataset [34], consisting of approximately 9,850 social media posts collected from various platforms and annotated into five categories: Depressed, Anxious, Frustrated, Suicidal, and others. The dataset captures diverse real-world emotional expressions, making it suitable for multi-class mental health detection tasks.Its unstructured nature includes informal language, slang, emojis, and inconsistent grammar, providing both rich context and preprocessing challenges. Additionally, class imbalance exists, particularly in critical categories like "Suicidal," requiring appropriate balancing techniques to ensure accurate and fair classification.

Parameter	Description
Dataset Name	Mental Distress (2026)
Total Samples	~9,850 posts
Data Type	Social media text
Classes	5 (Depressed, Anxious, Frustrated, Suicidal, Others)
Nature	Unstructured textual data

Table 4: Dataset Overview

3.2 Data Preprocessing: Data preprocessing is essential for preparing raw social media text, which often contains noise such as special characters, URLs, hashtags, and inconsistent formatting. To standardize the data, text is converted to lowercase and unnecessary elements like punctuation and special characters are removed. Stopword removal is applied to eliminate common words, while retaining important negation terms such as "not" and "no" to preserve sentiment. Tokenization splits the text into individual words or tokens [35], and sequence padding ensures uniform input length for efficient neural network processing.

Mathematical Representation of Preprocessing

Let the original text be represented as:

$$T = \{w_1, w_2, w_3, \ldots, w_n\} \quad (1)$$

After preprocessing:

$$T' = \text{Clean}(T) = T - S \quad (2)$$

Where:

- T= Original text
- S= Stopwords
- T′= Cleaned text

Sequence Padding Formula

$$X_{\text{padded}} = [x_1, x_2, \ldots, x_n, 0, 0, \ldots, 0] \quad (3)$$

Where all sequences are padded to a fixed length.

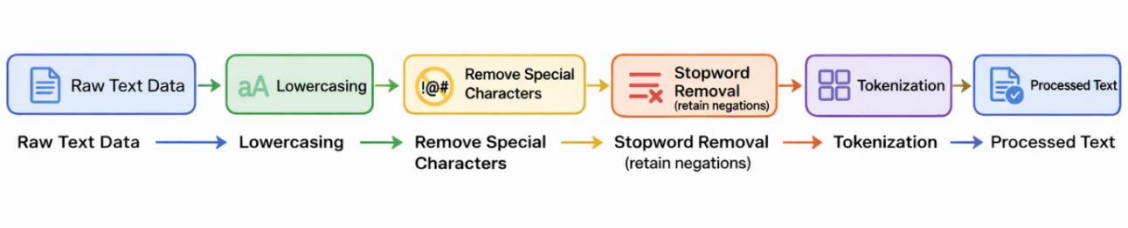

Fig 1: Data Preprocessing Pipeline

3.3 Feature Engineering

Feature engineering plays a vital role in transforming textual data into a numerical format that can be processed by deep learning models. In this study, word embeddings convert tokens into dense vectors that capture semantic relationships between words. Unlike traditional methods such as Bag-of-Words, embeddings preserve context by mapping similar words closer in vector space. An embedding layer is used within the neural network to learn these representations during training [31], [36] to learn these representations during training. This improves model performance by efficiently handling large vocabularies while preserving both syntactic and semantic information.

Embedding Representation

$$E \in \mathbb{R}^{V \times D} \quad (4)$$

Where:

- V= Vocabulary size
- D= Embedding dimension

Each word is represented as:

$$x_i \rightarrow \vec{e_i} \quad (5)$$

Technique	Description
Tokenization	Splitting text into words
Embedding	Converting words into vectors
Sequence Padding	Uniform input length
Vocabulary Indexing	Assigning integer IDs to words

Table 5: Feature Engineering Techniques

3.4 Handling Class Imbalance

Class imbalance is common in mental health datasets, where critical classes like “Suicidal” have fewer samples. This can bias models toward majority classes and reduce detection of minority classes. Applying class weighting during training gives more importance to underrepresented classes, improving fairness and overall model performance. As a result, class weighting improves overall model fairness and enhances performance across all classes.

Class Weight Formula

$$w_i = \frac{N}{C \times n_i} \quad (6)$$

Where:

- w_i= Weight of class i
- N= Total number of samples
- C= Number of classes
- n_i= Number of samples in class i

Technique	Purpose
Class Weighting	Assign higher importance to minority classes
Oversampling	Increase samples of minority classes
Undersampling	Reduce majority class samples

Table 6. Class Imbalance Handling Techniques

3.5 Model Architecture: The proposed system employs two deep learning architectures for multi-class mental health detection: the Long Short-Term Memory (LSTM) model [25] processes textual sequences… and the Stacked Bidirectional LSTM (BiLSTM) [30] further enhances this capability. The LSTM model processes text in a single forward direction, capturing temporal dependencies but limiting full contextual understanding. To address this, a stacked BiLSTM processes data in both forward and backward directions, capturing richer context. This architecture learns hierarchical features and improves detection of subtle emotional patterns in mental health analysis [37].

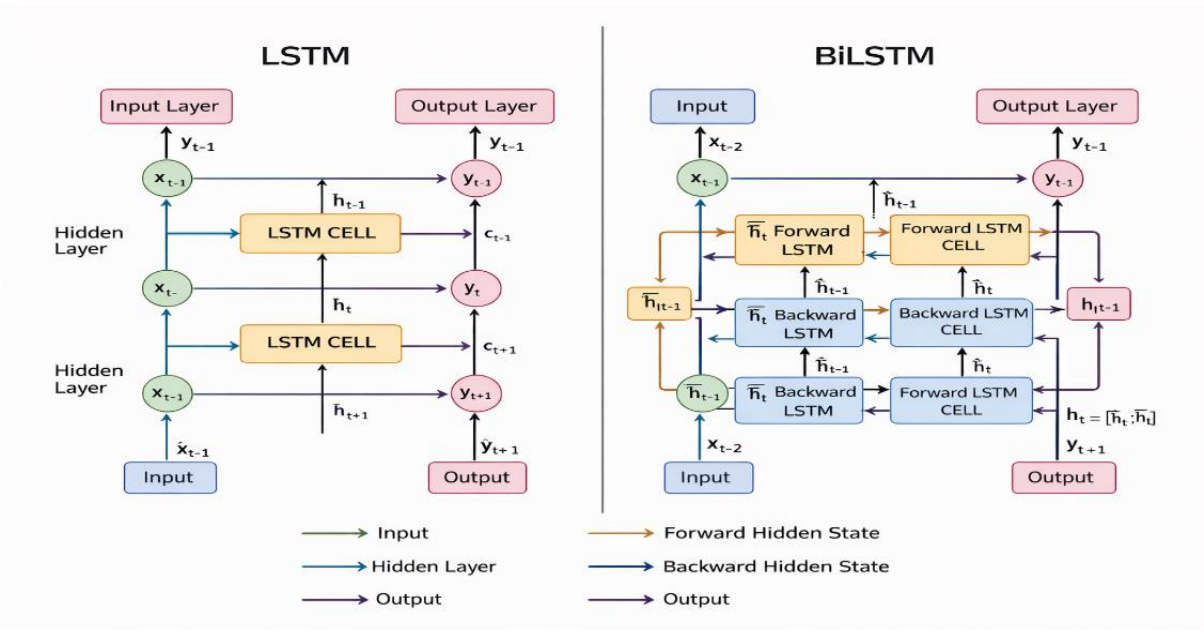

Fig 2: Lstm and Bilstm Architecture

3.6 Key Components: The embedding layer [31] is the first component of the model, that converts text into dense numerical vectors, mapping each word to a 200-dimensional representation to capture semantic relationships while reducing dimensionality compared to one-hot encoding. The model uses stacked BiLSTM layers with 128 and 64 units to capture short- and long-term dependencies. Dropout helps prevent overfitting, and a dense layer with Softmax activation performs final classification across five mental health categories.

Mathematical Representation of Embedding

$$E \in \mathbb{R}^{V \times D} \quad (7)$$

Where:

- V= Vocabulary size
- $D = 200$ Embedding dimension

LSTM Equations

The LSTM unit consists of three gates and a cell state:

Forget Gate:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (8)$$

Input Gate:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (9)$$

Candidate Cell State:

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (10)$$

Cell State Update:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (11)$$

Output Gate:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (12)$$

Hidden State:

$$h_t = o_t \cdot \tanh(C_t) \quad (13)$$

BiL STM Representation

BiLSTM processes sequences in two directions:

$$\overrightarrow{h_t} = \text{LSTM}(x_t) \quad (14)$$

$$\overleftarrow{h_t} = \text{LSTM}(x_t) \quad (15)$$

Final output:

$$y_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}] \quad (16)$$

Where $\oplus$denotes concatenation.

Layer Type	Configuration
Embedding Layer	200 dimensions
BiLSTM Layer 1	128 units
BiLSTM Layer 2	64 units
Dropout Layer	0.5
Dense Layer	SoftMax (5 classes)

Table 7: Model Architecture Summary

3.7 Performance Metrics Interpretation

Accuracy:

Accuracy measures the overall correctness of the model:

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total predictios}} \quad (17)$$

The high accuracy of BiLSTM indicates its effectiveness in classifying mental health categories.

Precision:

Precision indicates how many predicted positives are actually correct:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (18)$$

Higher precision in BiLSTM shows fewer false positives.

Recall:

Recall measures how many actual positives are correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (19)$$

BiLSTM achieves better recall, meaning it detects more true cases.

F1-Score:

F1-score balances precision and recall:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (20)$$

A higher F1-score indicates a better balance between precision and recall.

3.8 Training Strategy

The model uses the Adam optimizer for efficient convergence with adaptive learning rates. Categorical cross-entropy loss is applied for multi-class classification. To improve robustness and generalization, 5-fold cross-validation is used, where the dataset is split into five parts and trained and validated on different combinations.

This technique reduces overfitting and provides a more reliable evaluation of model performance Fig 2.

Categorical Cross-Entropy Loss

$$\text{L} = -\sum_{\text{i}=1}^{\text{C}} \text{y}_\text{i} \log(\widehat{y_\text{i}}) \quad (21)$$

Where:

- C= Number of classes
- y_i= True label
- $\widehat{y_i}$= Predicted probability

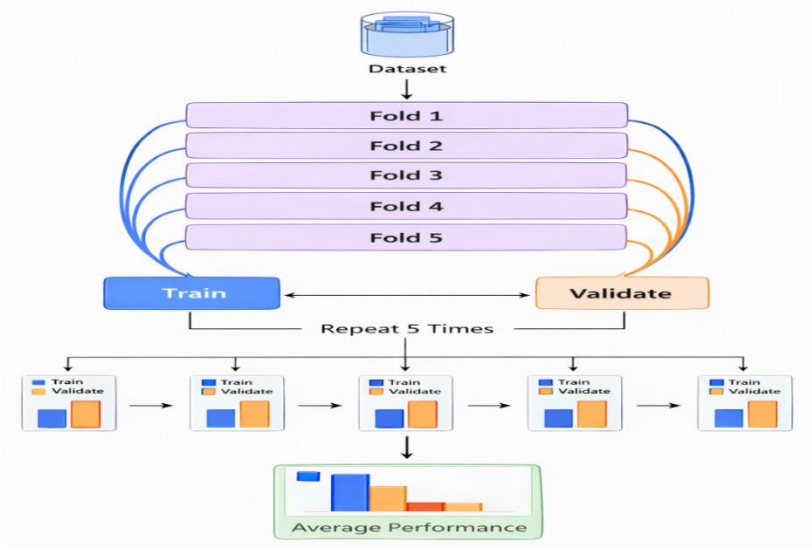

Fig 2: 5-Fold Cross-Validation

Parameter	Value
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Validation Method	5-Fold Cross-Validation
Epochs	10–20
Batch Size	32

Table 8. Training Parameters

3.9 Implementation Steps

The proposed mental health detection system follows a structured pipeline from data preprocessing to model evaluation. The steps are described as follows:

Step 1: Data Collection

The Mental Distress dataset containing labeled social media text across five emotional classes is loaded and checked for missing values and inconsistencies.

Step 2: Data Cleaning and Preprocessing

Text is converted to lowercase and cleaned by removing special characters, URLs, and punctuation. Stop words are removed while retaining negations like “not” and “no.” Tokenization is applied, followed by sequence padding for uniform input length.

Step 3: Text Representation

The processed text is converted into numerical form using tokenization and vocabulary indexing. An embedding layer transforms words into dense vectors capturing semantic relationships.

Step 4: Model Construction

Two models are developed: LSTM and Stacked BiLSTM. The architecture includes an embedding layer, followed by recurrent layers (LSTM/BiLSTM), dropout for regularization, and a dense output layer with SoftMax activation for multi-class classification.

Step 5: Model Training

Models are trained using the Adam optimizer with categorical cross-entropy loss. Stratified 5-fold cross-validation is used for balanced training and validation.

Step 6: Performance Evaluation

Model performance is evaluated using accuracy, precision, recall, and F1-score, along with a confusion matrix for class-wise analysis.

Step 7: Result Comparison

The performance of LSTM and BiLSTM models is compared, with BiLSTM showing better accuracy and contextual understanding due to bidirectional learning.

The system is implemented using deep learning and NLP libraries. TensorFlow is used as the backend for training and

optimization [38]. Keras provides a high-level API for building and implementing neural network architectures [39]. Scikit-learn is utilized for data preprocessing, model evaluation, and performance metrics computation [40] Natural Language Toolkit (NLTK) is used for text preprocessing tasks such as tokenization and stop word removal [35]

3.10 Algorithm for Multi-Class Mental Health Detection

Algorithm 1: Multi-Class Mental Health Detection using LSTM/BiLSTM

Input:

Mental Distress Dataset (text data with labels)

Output:

Predicted mental health class (Depressed, Anxious, Frustrated, Suicidal, Others)

Step 1: Load the dataset D
Step 2: Remove null or duplicate entries from D
Step 3: For each text sample Tin dataset:
a. Convert text to lowercase
b. Remove special characters and punctuation
c. Remove stopwords (retain negation words)
d. Tokenize text into words
Step 4: Convert tokens into sequences using vocabulary indexing
Step 5: Apply sequence padding to ensure uniform input length
Step 6: Split dataset into training and validation sets using 5-fold cross-validation
Step 7: Initialize embedding layer with dimension $D = 200$
Step 8: Build Model:
a. Add Embedding Layer
b. Add LSTM / BiLSTM layers
c. Apply Dropout for regularization
d. Add Dense layer with SoftMax activation
Step 9: Compile model using:
Optimizer = Adam
Loss Function = Categorical Cross-Entropy
Step 10: Train model on training data
Step 11: Evaluate model on validation data
Step 12: Compute performance metrics:
Accuracy, Precision, Recall, F1-Score
Step 13: Generate confusion matrix
Step 14: Compare LSTM and BiLSTM results
Step 15: Select best-performing model
Step 16: Output predicted class labels

The above algorithm summarizes the complete workflow of the proposed system, from data preprocessing to model evaluation and comparison.

IV.RESULTS AND DISCUSSION

4.1 Experimental Results: The performance of the proposed LSTM and stacked BiLSTM models is evaluated using accuracy, precision, recall, and F1-score. Trained on the Mental Distress dataset with 5-fold cross-validation, both models effectively detect mental health categories. However, the BiLSTM model performs better, achieving about **89.7% accuracy** compared to **86.0% for LSTM**, due to its ability to capture bidirectional contextual dependencies in text.

Metric	LSTM	BiLSTM
Accuracy	86.0%	89.7%
Precision	0.86	0.89
Recall	0.85	0.88

Table 9: Model Performance Comparison

4.2 Graphical Analysis: Graphical representation helps in better understanding and comparison of model performance. The following bar chart illustrates the comparison between LSTM and BiLSTM models across different evaluation metrics.

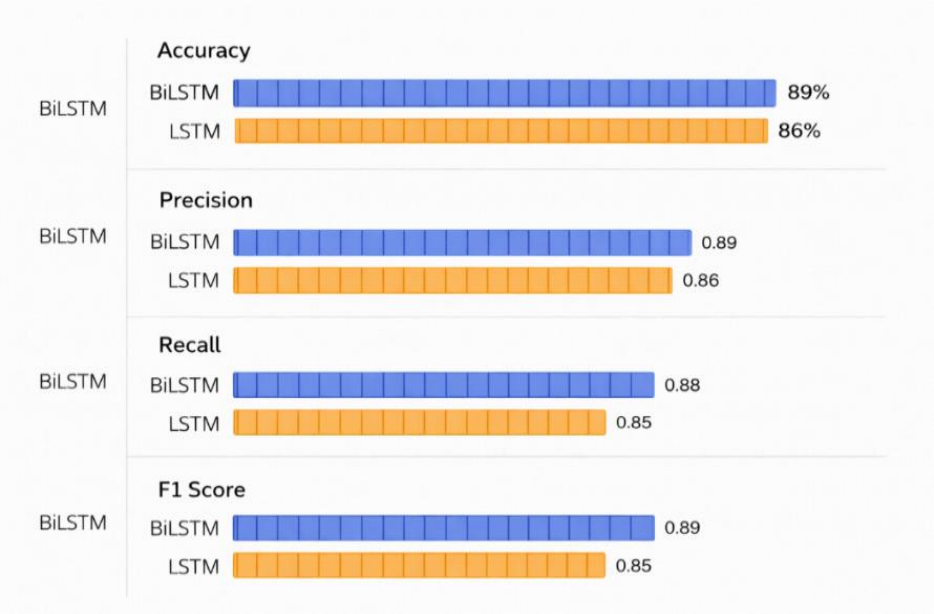

Fig 4: Model Performance Graph

The graph shows that the BiLSTM model outperforms the LSTM model across all evaluation metrics, especially in precision and F1-score, indicating better classification and fewer misclassifications. This highlights the benefit of bidirectional context in improving model performance.

4.3 Confusion Matrix Analysis: The confusion matrix evaluates classification performance by showing correct and incorrect predictions across all classes. In multi-class problems, it helps assess how well the model distinguishes between different emotional states.

The LSTM confusion matrix shows most predictions correctly classified along the diagonal, indicating good accuracy. The model performs well for "Anxious," "Frustrated," and "Others," but shows confusion between "Depressed" and "Suicidal," highlighting difficulty in distinguishing closely related emotions.

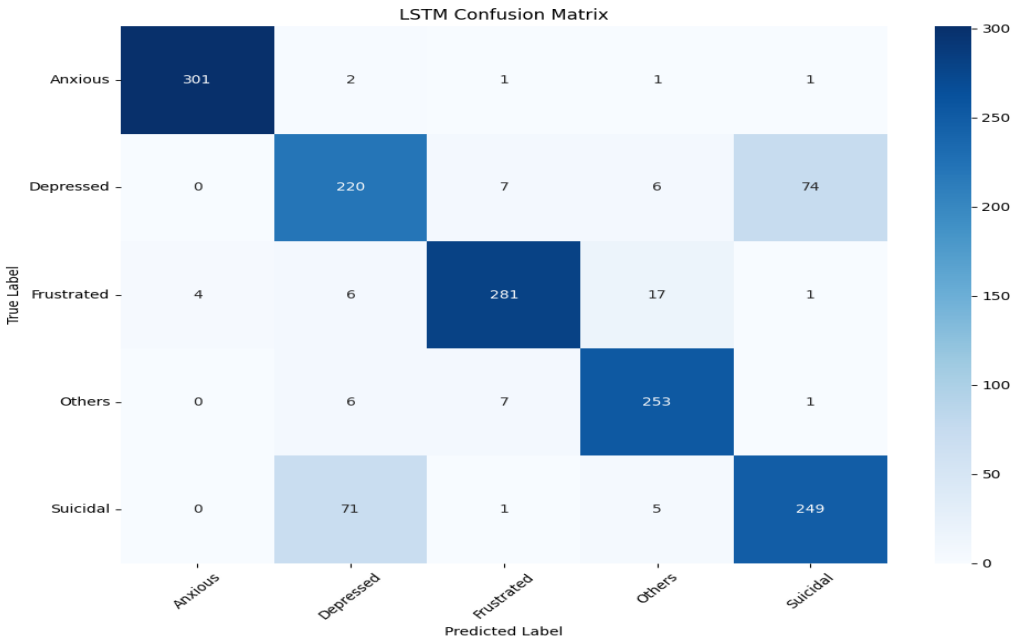

Fig 5: Confusion Matrix (LSTM Model)

The BiLSTM confusion matrix shows strong performance, with most predictions along the diagonal indicating high accuracy. The model effectively identifies "Anxious," "Frustrated," and "Suicidal" classes, with only minor confusion between "Depressed" and related categories.

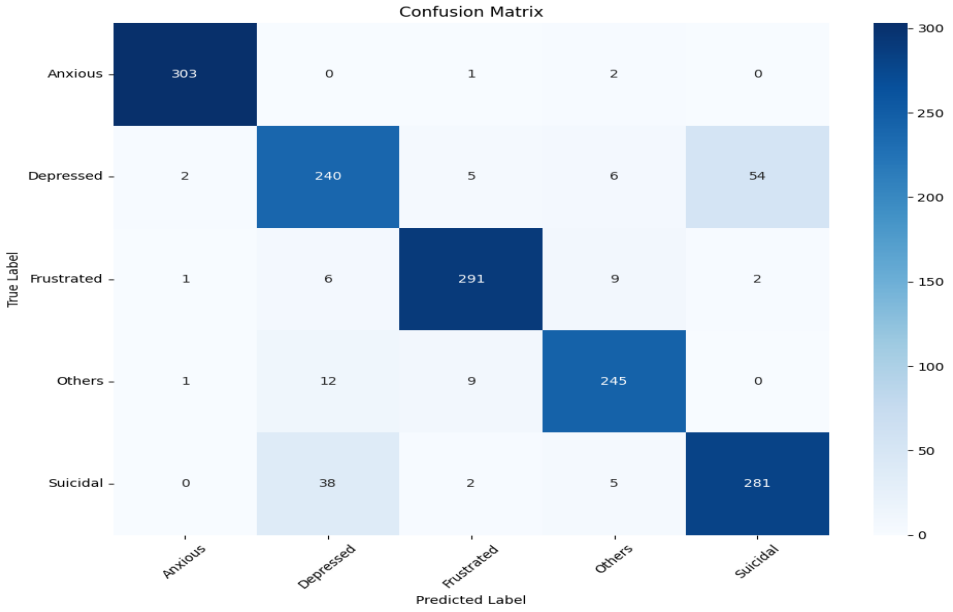

Fig 6: Confusion Matrix (BiLSTM Model)

4.4 Performance Metrics Interpretation Accuracy:
The LSTM model achieved an accuracy of **86.0%**, while the BiLSTM model achieved a higher accuracy of **89.7%**, indicating better overall performance.

Precision:
LSTM obtained a precision of **0.86**, whereas BiLSTM achieved **0.89**, showing that BiLSTM produces fewer false positive predictions.

Recall:
The recall value for LSTM is **0.85**, while BiLSTM achieved **0.88**, demonstrating that BiLSTM is more effective in identifying true cases.

F1-Score:
LSTM recorded an F1-score of **0.85**, while BiLSTM achieved **0.89**, indicating a better balance between precision and recall in BiLSTM.

4.5 Discussion: The results show that the BiLSTM model is effective for multi-class mental health detection due to its ability to understand context from both directions, making it more reliable than LSTM. However, it faces challenges in distinguishing similar emotions like anxiety and frustration, and class imbalance affects minority classes such as "Suicidal." Future improvements can include attention mechanisms and transformer-based models to enhance performance.

V.CONCLUSION AND FUTURE WORK

This research presented a deep learning-based approach for multi-class mental health detection using textual data derived from social media platforms. The study focused on addressing the limitations of traditional binary classification methods by introducing a more comprehensive multi-class framework capable of identifying diverse emotional states such as Depressed, Anxious, Frustrated, Suicidal, and Others. Through systematic data preprocessing, feature engineering, and model development, the study successfully demonstrated the effectiveness of advanced Natural Language Processing techniques in understanding complex human emotions.

The comparative analysis between LSTM and Stacked BiLSTM models revealed that the BiLSTM architecture outperforms the standard LSTM in terms of accuracy, precision, recall, and F1-score. This improvement is primarily attributed to the bidirectional processing capability of BiLSTM, which enables better contextual understanding of textual data. The use of embedding layers and sequence modeling further enhanced the model's ability to capture semantic relationships. Overall, the results validate that deep learning models, particularly BiLSTM, are highly suitable for mental health detection tasks and can serve as reliable tools for early identification of psychological distress.

5.1 Key Contributions: This research makes several important contributions to the field of mental health analytics and text classification. Firstly, it introduces a multi-class classification framework that better reflects real-world mental health conditions compared to traditional binary approaches. Secondly, the study implements and evaluates both LSTM and BiLSTM models, providing a clear comparison of their performance in handling contextual dependencies in textual data.
Additionally, the research incorporates effective preprocessing techniques, embedding strategies, and class imbalance handling methods to improve model performance. The use of stratified cross-validation ensures robustness and reliability of the results. The proposed framework can be extended and adapted for various real-world applications, making it a valuable contribution to both academic research and practical implementations in mental health monitoring systems.

5.2 Limitations of the Study: Despite achieving promising results, the study has certain limitations that need to be acknowledged. One of the primary limitations is the reliance on a single dataset, which may not fully capture the diversity of language and cultural variations present in real-world scenarios. The dataset, although comprehensive, may contain biases that could affect model generalization when applied to different populations or platforms.

Another limitation is related to the complexity of human emotions, which are often subtle and overlapping. The model may struggle to accurately distinguish between closely related emotional states such as anxiety and frustration. Additionally, the performance of minority classes, such as "Suicidal," may be affected due to class imbalance, despite the use of weighting techniques. These limitations highlight the need for further research and model improvements.

5.3 Future Scope: The future scope of this research is extensive and offers several opportunities for improvement and expansion. One potential direction is the integration of transformer-based models such as BERT, which have shown superior performance in natural language understanding tasks. These models can further enhance contextual understanding and improve classification accuracy. Additionally, incorporating attention mechanisms into the existing BiLSTM architecture can help the model focus on important words and phrases, leading to better interpretation of emotional content.

Another promising area is the extension of this system to multilingual datasets, enabling mental health detection across different languages and cultural contexts. Real-time deployment of the model in applications such as mobile apps, chatbots, and social media monitoring tools can significantly improve accessibility and early intervention. Furthermore, integrating this system with healthcare platforms can assist professionals in identifying at-risk individuals and providing timely support. Future research can also explore the use of multimodal data, such as images and voice inputs, to develop more comprehensive mental health detection systems.

Recent advancements in transformer-based architectures such as BERT [41] and attention mechanisms [42] can further enhance model performance and contextual understanding.

References

1. World Health Organization, "Comprehensive Mental Health Action Plan 2013–2030," 2021, *World Health Organization, Geneva, Switzerland*. [Online]. Available: https://apps.who.int/iris/handle/10665/345301
2. "Mental health summit Voices of people with lived experience in the WHO South-East Asia Region," 2024. [Online]. Available: http://apps.who.int/bookorders.
3. M. Freeman, "The World Mental Health Report: transforming mental health for all," *World Psychiatry*, vol. 21, pp. 391–392, Mar. 2022, doi: 10.1002/wps.21018.
4. M. Farahzadi, "Depression; Let's talk," *J. Community Health Res.*, Mar. 2017.
5. C. Lomas, "Global Mental Health Workforce Composition and National Depression Prevalence: Cross-National Evidence From the WHO Mental Health Atlas 2020 and the Global Burden of Disease Study 2019," *Int. J. Soc. Psychiatry*, p. 207640261424406, Mar. 2026, doi: 10.1177/00207640261424406.
6. J. C. Eichstaedt *et al.*, "Facebook language predicts depression in medical records," *Proceedings of the National Academy of Sciences*, vol. 115, no. 44, pp. 11203–11208, 2018, doi: 10.1073/pnas.1802331115.
7. B. Rudd and R. Beidas, "Digital Mental Health: The Answer to the Global Mental Health Crisis?," *JMIR Ment. Health*, vol. 7, p. e18472, Mar. 2020, doi: 10.2196/18472.
8. Z. Yingbo *et al.*, "The comprehensive clinical benefits of digital phenotyping: from broad adoption to full impact," *NPJ Digit. Med.*, vol. 8, p. 196, Mar. 2025, doi: 10.1038/s41746-025-01602-5.
9. M. Morales, S. Scherer, and R. Levitan, "A Linguistically-Informed Fusion Approach for Multimodal Depression Detection," Mar. 2018, pp. 13–24. doi: 10.18653/v1/W18-0602.
10. D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, vol. 2. 2008.
11. L. he *et al.*, "Deep Learning for Depression Recognition with Audiovisual Cues: A Review," Mar. 2021. doi: 10.48550/arXiv.2106.00610.
12. Y. Hussain, M. A. Zaheer, A. M. Khan, and A. S. Malik, "Depression detection using deep learning and large language models from multimodalities," *Front. Digit. Health*, vol. Volume 8-2026, 2026, doi: 10.3389/fdgth.2026.1759857.
13. A. Mccallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," *Work Learn Text Categ*, vol. 752, Mar. 2001.
14. C. Cortes and V. Vapnik, "Support-vector networks," *Chem. Biol. Drug Des.*, vol. 297, pp. 273–297, Mar. 2009, doi: 10.1007/%2FBF00994018.
15. L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, Mar. 2001, doi: 10.1023/A:1010950718922.
16. T. J. Devi and A. Gopi, "The Evaluation of Deep Learning Models for Detecting Mental Disorders Based on Text Summarization in Societal Analysis," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 3, pp. 1620–1628, 2024.
17. Q. Saeed and Y. Cha, "Multi-modal deep-attention-BiLSTM based early detection of mental health issues using social media posts," *Sci. Rep.*, vol. 15, Mar. 2025, doi: 10.1038/s41598-025-19141-0.
18. Y. Hussain, M. A. Zaheer, A. M. Khan, and A. S. Malik, "Depression detection using deep learning and large language models from multimodalities," *Front. Digit. Health*, vol. Volume 8-2026, 2026, doi: 10.3389/fdgth.2026.1759857.
19. J. Thekkekara and S. Yongchareon, "An Attention-Based BERT–CNN–BiLSTM Model for Depression Detection from Emojis in Social Media Text," *Big Data and Cognitive Computing*, vol. 9, p. 310, Mar. 2025, doi: 10.3390/bdcc9120310.
20. M. Nadeem, M. A. Abbasi, F. Ali, K. Mughal, S. Azhar, and A. Saeed, "A Deep Learning Approach to Early Detection of Depression Through Social Media Text Analysis," *International Journal of Social Sciences Bulletin*, vol. 1, no. 1, pp. 653–675, 2025, doi: 10.5281/zenodo.17435175.
21. S. C H, "A Deep Learning Framework for Depression and Major Depressive Disorder Detection from Social Media Text Using

MentalBERT and Multilayer Perceptron," Mar. 2025, doi: 10.22266/ijies2025.0930.16.
22. A. Amanat *et al.*, "Deep Learning for Depression Detection from Textual Data," *Electronics (Basel).*, vol. 11, Mar. 2022, doi: 10.3390/electronics11050676.
23. Md. Z. Uddin, K. Dysthe, A. Følstad, and P. Brandtzaeg, "Deep learning for prediction of depressive symptoms in a large textual dataset," *Neural Comput. Appl.*, vol. 34, pp. 1–24, Jan. 2022, doi: 10.1007/s00521-021-06426-4.
24. Y. Shen, H. Yang, and L. Lin, "Automatic Depression Detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based Model," Mar. 2022. doi: 10.48550/arXiv.2202.08210.
25. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Mar. 1997, doi: 10.1162/neco.1997.9.8.1735.
26. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Mar. 1997, doi: 10.1162/neco.1997.9.8.1735.
27. S. Jain, D. Rastogi, P. Nagrath, G. Stoian, D. Dănciulescu, and J. D, "Depression Detection Model Based on Facial, Verbal Features, and Motion Activity Using Deep Learning," *Traitement du Signal*, vol. 42, pp. 2499–2512, Mar. 2025, doi: 10.18280/ts.420505.
28. G. Lam, H. Dongyan, and W. Lin, "CONTEXT-AWARE DEEP LEARNING FOR MULTI-MODAL DEPRESSION DETECTION."
29. B. Diep, M. Stanojevic, and J. Novikova, "Multi-modal deep learning system for depression and anxiety detection," Mar. 2022. doi: 10.48550/arXiv.2212.14490.
30. P. Zhou *et al.*, "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification," Mar. 2016, pp. 207–212. doi: 10.18653/v1/P16-2034.
31. T. Mikolov, K. Chen, G. s Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Proceedings of Workshop at ICLR*, vol. 2013, Mar. 2013.
32. Y. Hussain, M. A. Zaheer, A. M. Khan, and A. S. Malik, "Depression detection using deep learning and large language models from multimodalities," *Front. Digit. Health*, vol. Volume 8-2026, 2026, doi: 10.3389/fdgth.2026.1759857.
33. L. he *et al.*, "Deep Learning for Depression Recognition with Audiovisual Cues: A Review," Mar. 2021. doi: 10.48550/arXiv.2106.00610.
34. F. Y. Prity, T. C. Munira, S. Ahmed Shayeed, and M. J. U. Chowdhury, "MentalDistress: A multi-class social media text dataset for mental health–related emotion detection," 2026, *Mendeley Data*. doi: 10.17632/b42wr437hg.1.
35. S. Bird, "NLTK: The natural language toolkit," Mar. 2006. doi: 10.3115/1225403.1225421.
36. T. Mikolov, K. Chen, G. s Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Proceedings of Workshop at ICLR*, vol. 2013, Mar. 2013.
37. D. P. Patinavalasa and D. Suneel Kumar, "Scalable Email Spam Detection Using BiLSTM with Large-Scale Hybrid Datasets," *International Journal of Recent Trends in Multidisciplinary Research*, p. 96, Mar. 2026, doi: 10.59256/ijrtmr.20260602016.
38. M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," Mar. 2016, doi: 10.48550/arXiv.1603.04467.
39. A. Zhang, Z. Lipton, M. Li, and A. Smola, "Dive into Deep Learning," Mar. 2021. doi: 10.48550/arXiv.2106.11342.
40. F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, Mar. 2012.
41. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Mar. 2018. doi: 10.48550/arXiv.1810.04805.
42. A. Vaswani *et al.*, "Attention Is All You Need," Mar. 2017, doi: 10.48550/arXiv.1706.03762.