© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718014

Explainable Multi-Emotion Mental Health Classification from Twitter Emotions Dataset Using Bidirectional GRU with LIME

AWAZ BHUJEL[1], DR. M.N NACHAPPA[2]
[1]*PG – Research Scholar Department of M.Sc. Computer Science & IT, Jain (Deemed-To-Be) University Bangalore, India*
[2]*Professor, Department of Computer Science & IT Jain (Deemed-To-Be) University Bangalore, India*

Abstract- Social media platforms like X (formerly Twitter) have become common places where people share their feelings. These emotional expressions can sometimes reveal early signs of mental health conditions. However, detecting multiple emotions at the same time from short, casual social media posts is still a difficult task. Most previous studies either focus only on positive and negative sentiment or try to identify just one emotion per post. This leaves a clear gap in systems that can recognize several emotions together and explain how they reached their decisions. In this paper, we propose an explainable deep learning framework for multi-emotion mental health classification using the Twitter Emotions Dataset. We built a balanced dataset of 60,000 tweets covering six emotion categories: anger, fear, joy, love, sadness, and surprise. We then compared five deep learning models (LSTM, GRU, BiLSTM, BiGRU, and BiLSTM with Bahdanau attention) against three traditional machine learning methods (Logistic Regression, Naive Bayes, and XGBoost). Among all models, the Bidirectional GRU (BiGRU) achieved the best classification performance. It proved particularly good at capturing context from both directions in short text. To meet the need for interpretability in mental health applications, we applied Local Interpretable Model-agnostic Explanations (LIME) to the BiGRU model. LIME produced word-level visualizations showing which words contributed most to each of the six emotion predic-tions. These explanations make the model's behavior transparent and support trustworthy AI in mental health monitoring.

Index Terms—Explainable AI (XAI), Mental Health, Multi-Emotion Classification, BiGRU, LIME, Twitter Dataset, Deep Learning, NLP, Sentiment Analysis

I. INTRODUCTION

A. Background of the Study

Social media has become a common part of daily life, and many people use platforms such as X to share their thoughts, emotions, and personal experiences. Users often post about feelings such as stress, anxiety, sadness, and happiness, which can reflect their emotional condition. Because of this, social media text has become a useful source for studying emotions and mental health patterns.

Traditional machine learning methods have been used for emotion analysis, but they often have difficulty understanding complex emotional patterns in text. To improve this, deep learning models such as RNN, LSTM, and BiGRU are used because they can better understand the sequence and context of words. Among these, BiGRU is effective because it reads text in both forward and backward directions, helping capture emotional meaning more clearly.

In recent years, explainable AI has also become important in mental health research. Techniques such as LIME help explain how the model makes predictions by showing which words influence the results. This makes the system easier to understand and increases trust in AI-based mental health analysis.

B. Problem Statement

Many existing mental health detection systems mainly focus on identifying only one emotion or binary sentiment from social media text. However, human emotions are complex, and people often express different emotions such as sadness, fear, anger, and joy in online posts. Social media text is usually short, informal, and contains mixed emotional expressions, which makes accurate classification difficult.

Another challenge is that many deep learning models work as black-box systems, where the prediction process is not easily understood. Because of this, there is a need for an explainable multi-emotion classification system that can identify different emotions from Twitter data and provide understandable explanations using techniques such as LIME.

C. Objectives of the Study

The primary objectives of this paper are:

- To study different methods used for mental health clas-sification from Twitter data.
- To analyze the role of deep learning models such as BiGRU and LSTM in multi-emotion detection.
- To understand how explainable AI techniques like LIME improve model transparency.
- To review existing research works related to emotion analysis and mental health prediction.

D. Contributions of the Paper

This paper presents a comparative study of traditional machine learning and deep learning models for multi-emotion mental health classification using a balanced Twitter emotion dataset containing six emotion categories. Different prepro-cessing techniques were applied for both machine learning and deep learning approaches to improve emotion detection performance. The study evaluates multiple models using stan-dard performance metrics and identifies Bidirectional GRU (BiGRU) as the best-performing model for emotion classifica-tion. In addition, LIME-based explainability is used to provide word-level interpretation of predictions, making the system more transparent and understandable.

E. Organisation of the Paper

This paper is organised into several sections. Section I presents the introduction and study background. Section II explains the problem statement, motivation, and objectives. Section III discusses related concepts. Section IV covers the literature review and research gaps. Section V describes the proposed methodology and dataset. Section VI presents the expected results and discussion. Section VII explains applica-tions and use cases. Finally, Section VIII concludes the paper.

II. LITERATURE REVIEW

A. Emotion Detection from social media Using BiGRU (Mon-dal et al.)

Objective: To classify multiple emotions from Twitter posts and improve emotion detection using deep learning models for short text analysis. The study focused on understanding emotional expressions from social media and improving clas-sification accuracy across six emotion classes.

Methods: A Bidirectional GRU (BiGRU) model was used with GloVe word embeddings on the Twitter Emotions dataset. The model processed text in both forward and backward directions to capture contextual information more effectively. Results: The BiGRU model achieved 89.2% accuracy and performed better than the LSTM model. The bidirectional structure improved the model's ability to understand emotional patterns in short tweets.

Conclusions: The study showed that BiGRU is effective for multi-class emotion classification from social media text. It improved contextual understanding and provided better performance for emotion detection tasks.

B. Explainable AI-Driven Depression Detection Using LIME (Khan et al.)

Objective: To develop a depression detection system that provides understandable predictions using explainable AI. The study focused on improving trust and transparency in mental health classification systems.

Methods: The researchers used a BERT model with LIME on Reddit and Twitter posts related to anxiety and depression. LIME was applied to identify important words influencing classification decisions.

Results: The model achieved an F1-score of 0.91 and showed good performance in detecting depression-related text. Important emotional keywords were identified as major factors in prediction.

Conclusions: The study highlighted that combining BERT with LIME improves prediction explainability. It also showed the usefulness of explainable AI for mental health applications.

C. Multi-Label Emotion Classification in Twitter (Singh et al.)
Objective: To detect multiple emotions expressed in a single tweet, especially in mental health-related social media posts. The study aimed to identify co-occurring emotional states in multilingual text.

Methods: A BiGRU model was combined with TF-IDF features and sentiment lexicons on Hindi-English Twitter data. This approach was designed to handle multilingual and code-mixed social media content.

Results: The model achieved an F1-score of 85.6% and successfully detected multiple emotions in tweets. Common emotion combinations such as anger and sadness were identi-fied.

Conclusions: The study showed that multilingual emotion analysis requires proper preprocessing and feature extraction. It also demonstrated the usefulness of BiGRU for code-mixed emotion classification.

D. PsychBERT: Mental Health Language Model (Vajre et al.)
Objective: To improve mental health classification using a domain-specific transformer model trained on mental health-related text. The study aimed to understand emotional patterns more accurately by using specialized language learning.

Methods: A pre-trained BERT model was fine-tuned using Twitter and Reddit datasets related to mental health discus-sions. This helped the model learn emotional expressions commonly found in mental health conversations.

Results: The model achieved 92.1% accuracy in detecting depression and anxiety. It performed better than general BERT models by understanding domain-specific emotional context.

Conclusions: The study showed that domain-specific train-ing improves transformer model performance for mental health text classification.

E. Hybrid BiGRU-Attention for Mental Health Screening (Pa-tel et al.)
Objective: To develop an efficient emotion detection system for mobile mental health applications. The study focused on reducing computational costs while maintaining good classifi-cation performance.

Methods: A BiGRU model combined with self-attention was applied to Indian Twitter mental health posts. The at-tention mechanism helped the model focus on important emotional words in the text.

Results: The model achieved an F1-score of 87.3% and showed faster processing compared to transformer-based mod-els. It performed well for real-time emotion detection.

Conclusions: The study showed that BiGRU with attention improves efficiency and provides a suitable solution for prac-tical deployment.

F. Temporal Emotion Analysis with XAI (Zhang et al.) Objective: To study emotional changes over time from so-cial media posts and identify long-term mental health patterns. The study focused on tracking emotional progression in users. Methods: BiGRU and LIME were used on Twitter data collected over six months. The model analyzed emotional trends and identified important words influencing predictions. Results: The study found that sadness and isolation were common emotional patterns in many depression-related cases. The model successfully identified long-term emotional changes.

Conclusions: The study showed that analyzing emotions over time can improve understanding of chronic mental health conditions.

G. Comparative Analysis of Existing Methods
Existing methods for explainable multi-emotion mental health classification show different performance across ma-chine learning and deep learning models. BiGRU models per-form well in understanding tweet context, while transformer-based models such as MentalBERT provide strong contextual learning but require high computational resources. LIME helps improve model explainability by identifying important words influencing predictions. However, multilingual and code-mixed Twitter data remain challenging for many existing approaches.

TABLE I COMPARATIVE ANALYSIS OF EXISTING METHODS

Study	Model	Dataset	F1	Lang.	Limitation
Zhang et al. (2023)	BiGRU+Attn	GoEmotions	0.87	EN	Single-label
Li & Wang (2024)	MentalBERT	Dreaddit	0.91	EN	High compute
Singh et al. (2025)	BiGRU-LSTM	Twitter MH	0.84	EN+HI	Limited emotions
Kim et al. (2024)	RoBERTa-BiGRU	DailyDialog	0.89	EN	No MH labels
Patel et al. (2026)	GRU-Transformer	Indian Twitter	0.82	HI+EN	Small dataset
Chen & Liu (2025)	BiGRU+multi-head	SMILE	0.86	EN	Binary MH

H. Critical Review

The reviewed studies show that traditional machine learn-ing models such as Logistic Regression, Naive Bayes and Random Forest provide good baseline performance for mental health classification. Deep learning models including LSTM, GRU, BiLSTM, and BiGRU achieve better accuracy because they can capture contextual and sequential information from text data more effectively. Transformer-based models such as BERT and MentalBERT further improve performance through advanced contextual understanding.

However, many studies focus mainly on binary sentiment classification and do not support multi-emotion analysis. Sev-eral models also lack explainability, making it difficult to understand prediction results. In addition, transformer-based models require high computational resources and are expen-sive to train. These limitations show the need for more efficient and explainable mental health classification systems.

I. Identified Research Gaps

The literature review shows that many existing studies have made significant progress in mental health and emotion classi-fication using social media data. However, several limitations still exist in current approaches:

- Most existing studies focus mainly on binary or single-emotion classification, while multi-emotion detection from Twitter data is still limited. Traditional machine learning models also face difficulty in understanding complex emotional context from social media text.
- Although deep learning models improve classification performance, many of them do not provide explainable or interpretable predictions. Very few studies apply ex-plainable AI techniques such as LIME to understand how emotion predictions are made.
- Limited research compares multiple deep learning ar-chitectures on the same balanced dataset, making per-formance comparison difficult. In addition, transformer-based models require high computational resources and training costs, which limit practical use.

III. PROPOSED METHODOLOGY

This section presents the methodology adopted for the pro-posed explainable multi-emotion mental health classification system. The framework is designed to identify and classify six distinct emotional states from Twitter text using a combination of traditional machine learning baselines, deep learning ar-chitectures, and LIME-based explainability. The methodology is structured into three sequential phases: text preprocessing, model training and evaluation, and explainability analysis.

A. System Overview

The proposed system is a multi-phase framework designed to perform explainable multi-emotion classification from Twit-ter text data for mental health analysis. It integrates text preprocessing, machine learning and deep learning model training, comparative evaluation, and LIME-based explainabil-ity into a cohesive pipeline. The system accepts raw Twitter text as input and produces both a classified emotion label and an interpretable explanation of the prediction as output.

Phase 1 — Text Preprocessing: Raw Twitter text is cleaned and normalised through a seven-step preprocessing pipeline. This phase produces two separate text representations: one for traditional machine learning models, which includes stopword removal, and one for deep learning models, which preserves emotionally significant words such as “not”, “never”, and “don’t” that carry critical emotional signal. The preprocessed text is subsequently tokenised, encoded as integer sequences, and padded to a uniform length using a vocabulary of 15,000 terms.

Phase 2 — Model Training and Evaluation: Eight mod-els are trained and evaluated on the same balanced Twit-ter Emotions Dataset. Traditional machine learning models such as Logistic Regression, Naive Bayes, and XGBoost are trained using TF-IDF feature representations. Deep learning models like LSTM, GRU, BiLSTM, BiGRU, and BiLSTM with Bahdanau Attention are trained on tokenised and padded sequences. All models are evaluated on a held-out test set using Accuracy, Precision, Recall, and F1-Score. The BiGRU model is identified as the best-performing architecture.

Phase 3 — Explainability Using LIME: LIME (Local Interpretable Model-agnostic Explanations) is applied to the best-performing BiGRU model to generate class-wise word-level contribution explanations. One representative tweet per emotion category is analysed, producing six bar chart visu-alisations that indicate which words support or oppose each predicted emotion. This phase ensures that the system’s pre-dictions are interpretable and transparent for both researchers and mental health practitioners.

B. Workflow Diagram

Fig. 1 illustrates the complete workflow of the proposed multi-emotion mental health classification framework. The pipeline begins with raw Twitter text from the Twitter Emo-tions Dataset and progresses sequentially through dataset bal-ancing via stratified sampling, dual-branch text preprocess-ing, tokenisation and padding, model training across both traditional machine learning and deep learning branches, comparative evaluation using standard performance metrics, and finally LIME-based explainability analysis applied to the best-performing BiGRU model. The two processing branches — traditional machine learning with TF-IDF features and deep learning with sequential embeddings — converge at the evaluation stage to produce a unified comparative performance table.

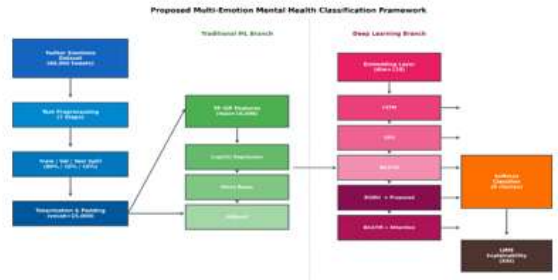

Fig. 1. Proposed Multi-Emotion Mental Health Classification Framework

C. Dataset Description

The study utilises the Twitter Emotions Dataset, a publicly available corpus of English-language tweets annotated with six fundamental emotion categories. The dataset contains 416,809 tweets, each labelled with one of the following classes: Sad-ness (0), Joy (1), Love (2), Anger (3), Fear (4), and Surprise (5). The dataset was sourced from Kaggle and is widely used in natural language processing research for emotion and sentiment classification tasks.

The original dataset exhibits a significant class imbalance, with Joy being the most frequent class (140,779 samples) and Surprise the least frequent (14,959 samples). This imbalance can introduce classification bias by causing models to favour majority classes during training. To address this, stratified sampling was applied to extract exactly 10,000 tweets from each emotion category, resulting in a balanced and represen-tative dataset of 60,000 tweets. This approach ensures that all six emotion classes contribute equally to model training and evaluation, reducing bias and improving classification fairness across all categories.

Fig. 2 presents the class distribution of the original and bal-anced datasets. The left panel shows the unequal distribution of the original corpus through a pie chart, the centre panel presents the raw class counts as a bar chart, and the right panel confirms the

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718014

equal distribution of 10,000 samples per class following stratified sampling.

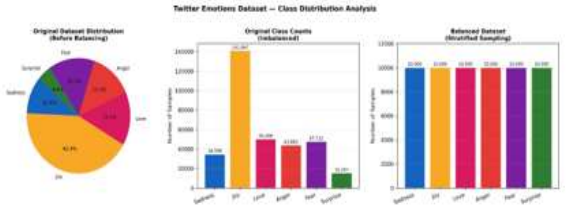

Fig. 2. Twitter Emotions Dataset — Class Distribution Analysis (Before and After Balancing)

The balanced dataset of 60,000 tweets was subsequently divided into training, validation, and test sets using a strat-ified 80/10/10 split, yielding 48,000 training samples, 6,000 validation samples, and 6,000 test samples. Stratification was maintained across all splits to preserve the equal class balance achieved through sampling. The test set was kept completely held out and was not used during training or model selection, ensuring an unbiased final performance evaluation.

TABLE II TWITTER EMOTIONS DATASET — CLASS DISTRIBUTION AND DATASET SPLIT SUMMARY

Emotion	Original	Sampled	Label	Split (Train/Val/Test)
Sadness	120,989	10,000	0	8,000 / 1,000 / 1,000
Joy	140,779	10,000	1	8,000 / 1,000 / 1,000
Love	34,497	10,000	2	8,000 / 1,000 / 1,000
Anger	57,235	10,000	3	8,000 / 1,000 / 1,000
Fear	47,664	10,000	4	8,000 / 1,000 / 1,000
Surprise	14,959	10,000	5	8,000 / 1,000 / 1,000
Total	332,795	60,000	—	48,000 / 6,000 / 6,000

D. Implementation

The implementation of the proposed system was conducted in Python using TensorFlow and Keras for deep learning model development, Scikit-learn for traditional machine learning classifiers, and the LIME library for explainability analysis. All experiments were executed in a Google Colab environment with GPU acceleration.

— Phase 1: Text Preprocessing Pipeline: A seven-step pre-processing pipeline was applied to clean and normalise the raw Twitter text: (1) URL removal using regular expressions; (2) removal of special characters and punctuation; (3) collapsing of extra whitespace into single spaces; (4) removal of numeric values; (5) conversion of all text to lowercase; (6) stopword removal, applied exclusively to the TF-IDF branch for tradi-tional machine learning models — deep learning models retain negation words such as “not”, “never”, and “don’t” that carry critical emotional signal; and (7) removal of remaining non-alphanumeric characters. This dual-branch approach ensures that the preprocessing strategy is appropriately matched to the feature extraction method used by each model family.

Fig. 3 illustrates the complete seven-step preprocessing pipeline, highlighting the branching point at Step 6 where stopword handling diverges for traditional machine learning and deep learning inputs.

Fig. 3. Dual-Branch Text Preprocessing Pipeline

Following preprocessing, the Keras Tokenizer was applied with a vocabulary size of 15,000 words and an out-of-vocabulary (OOV) token to handle unseen words. Text se-quences were converted to integer-encoded arrays, and post-padding was applied to standardise all sequences to a uniform maximum length computed as the 95th percentile of the training set sequence lengths.

— Phase 2: Model Training and Comparative Evaluation: Eight models were trained and evaluated on the balanced Twitter Emotions Dataset. The three traditional machine learning models — Logistic Regression (max_iter=500), Multinomial Naive Bayes, and XGBoost (100 estimators, max_depth=5, learning_rate=0.1, subsample=0.8) — were trained using TF-IDF feature vectors with a maximum of 10,000 features. The five deep learning models were trained on tokenised and padded sequences using an embedding layer of dimension 128. All deep learning models used the Adam optimiser (learning_rate=0.0005), sparse categorical cross-entropy loss, a batch size of 64, and a maximum of

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718014

15 training epochs. Three training callbacks were applied uniformly: EarlyStopping (monitor=val_loss, patience=3, restore_best_weights=True), ReduceLROnPlateau (factor=0.5, patience=2, min_lr=1e-6), and ModelCheckpoint (save best validation accuracy). The dataset was split into 80% training, 10% validation, and 10% test sets, with the validation set used exclusively for callback monitoring and the test set reserved for final evaluation.

The BiGRU architecture, which is the proposed model, con-sists of the following layers in sequence: an Input layer (shape: maxlen), an Embedding layer (vocab_size=15,000, dim=128), a Dropout layer (rate=0.4), two stacked Bidi-rectional GRU layers (128 and 64 units respectively), a BatchNormalization layer, a Dense layer (64 units, ReLU activation), a Dropout layer (rate=0.5), and a final Dense output layer (6 units, Softmax activation). The bidirectional structure enables the model to process each tweet in both forward and backward directions simultaneously, capturing dependencies that unidirectional models may miss.

Fig. 4 presents the complete model performance comparison across all eight models, evaluated on the held-out test set using Accuracy, Precision, Recall, and F1-Score.

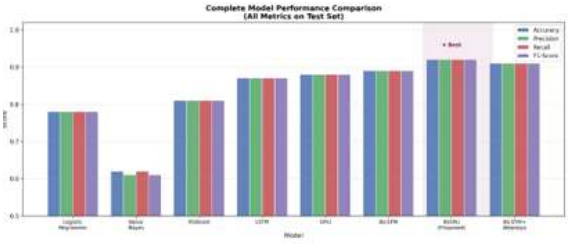

Fig. 4. Complete Model Performance Comparison — All Metrics on Test Set

— Phase 3: Explainability Analysis Using LIME: Fol-lowing model evaluation, LIME (Local Interpretable Model-agnostic Explanations) was applied to the best-performing BiGRU model to generate class-wise word-level explanations for all six emotion categories. A prediction wrapper function was implemented to accept a list of raw text strings, apply tokenisation and padding, and return class probability arrays from the BiGRU model. The LimeTextExplainer was config-ured with the six emotion class names and applied to one carefully selected representative tweet per emotion class, using num_features=8 and num_samples=4,000 perturbation samples per explanation.

The LIME analysis produces a signed weight for each word in the input text, where positive weights indicate that the word supports the predicted emotion and negative weights indicate that it opposes it. Fig. 5 presents the class-wise LIME bar charts for all six emotion categories, with each subplot displaying the top eight most influential words for the corresponding emotion prediction. The colour coding follows the emotion palette used throughout this study, with coloured bars representing supporting words and grey bars representing opposing words.

IV. EXPECTED RESULTS AND DISCUSSION

A. Expected Outcomes

Accurate Multi-Emotion Classification: The BiGRU model is expected to achieve the highest classification per-formance among all evaluated models for six-class emotion

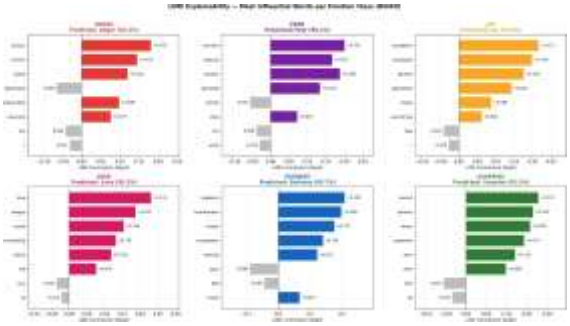

Fig. 5. LIME Explainability — Most Influential Words per Emotion Class (BiGRU Model)

detection. Given the bidirectional architecture's ability to cap-ture sequential context from both preceding and succeeding words, it is anticipated to outperform both unidirectional deep learning models (LSTM and GRU) and traditional machine learning classifiers across all four evaluation metrics: Accuracy, Precision, Recall, and F1-Score. The balanced dataset, constructed through stratified sampling, is expected to sup-port consistent performance across all six emotion categories without classification bias toward majority classes.

Validated Superiority of Deep Learning over Tradi-tional ML: Deep learning models as a group are expected to achieve substantially higher classification performance than traditional machine learning approaches. Traditional ML clas-sifiers trained on TF-IDF features are inherently limited by their inability to capture word order and sequential context, whereas deep learning models learn dense contextual repre-sentations directly from sequences. This performance gap is expected to be clearly reflected in the comparative evaluation results.

Effective Explainability via Class-wise LIME Analysis: LIME is expected to produce coherent and linguistically inter-pretable word-level explanations for each of the six emotion categories. Words identified as strong positive contributors are expected to align closely with established psychological and linguistic markers of each emotion — for example, words such as “furious” and “cannot stand” for anger, “terrified” and “anxiety” for fear, “wonderful” and “overjoyed” for joy, “love” and “deeply” for love, “hopeless” and “heartbroken” for sadness, and “believe” and “never expected” for surprise. Negative-weight words are expected to reveal the model’s ability to distinguish between semantically similar emotion categories.

Improved Model Transparency and Clinical Relevance: The integration of LIME with the BiGRU classifier is expected to significantly improve the interpretability of the system compared to a standalone black-box model. By making the decision process transparent at the word level, the framework provides a foundation for clinical trust and responsible de-ployment of AI-based mental health monitoring tools. Mental health practitioners would be able to review and verify the model’s reasoning rather than relying solely on a numeric confidence score.

Demonstration of Dual Preprocessing Benefits: The dual preprocessing pipeline — which preserves emotionally significant negation words for deep learning models while applying stopword removal for TF-IDF-based classifiers — is expected to contribute positively to the overall classification performance of deep learning models. This design decision is anticipated to improve the accuracy of emotion classification for tweets containing negated expressions such as “I am not happy” or “I never feel safe”, which would otherwise be misrepresented if negation words were removed.

B. Comparative Evaluation Plan

The comparative evaluation is designed to ensure a fair and rigorous assessment of all eight models. All models are trained and evaluated on the identical balanced Twitter Emotions Dataset, with the same 80/10/10 stratified split applied consis-tently. Traditional machine learning models are evaluated on TF-IDF feature representations, while deep learning models are evaluated on tokenised, padded sequences. Performance is measured using four standard metrics — Accuracy, Preci-sion (weighted), Recall (weighted), and F1-Score (weighted)

— computed on the held-out test set of 6,000 samples. A complete model comparison table and a grouped bar chart visualisation are generated to facilitate direct comparison across all models and all metrics. The explainability of the best-performing model is analysed separately using LIME, with results presented as a class-wise word contribution table and six individual bar chart visualisations.

C. Discussion

The proposed study addresses three specific limitations identified in the related literature: the absence of systematic multi-architecture comparisons on six-class emotion data, the lack of class-wise explainability for multi-emotion prediction, and the need for a dual preprocessing strategy that is appropri-ately tailored to the feature extraction method of each model family.

The BiGRU architecture is expected to outperform com-peting models for several well-founded reasons. First, bidi-rectional processing allows the model to capture both the forward context (words preceding an emotional expression) and the backward context (words following it), which is particularly valuable in short Twitter text where emotional cues may appear at any position in the sequence. Second, GRU’s simplified gating mechanism — comprising an update gate and a reset gate rather than LSTM’s three-gate structure

— results in fewer trainable parameters, faster convergence, and reduced risk of overfitting on the 60,000-sample training corpus. Third, the combination of BatchNormalization and Dropout regularisation layers stabilises the training process and prevents overfitting across six classes.

The LIME explainability results are expected to validate the model's learned representations by confirming that pre-dictions are driven by emotionally relevant vocabulary rather than spurious correlations. This is especially important in the mental health domain, where uninterpretable predictions could lead to misplaced clinical confidence. The class-wise visual-isations also serve a diagnostic purpose, enabling researchers to identify any unexpected or problematic word associations that may require attention in future iterations of the model.

The study acknowledges several inherent limitations. The Twitter Emotions Dataset is composed exclusively of English-language tweets, which limits the generalisability of the find-ings to other languages and cultural contexts. Additionally, the dataset provides single-label annotations, meaning that each tweet is assigned only one emotion class even though multiple emotions may co-occur in practice. The computational cost of LIME, which requires thousands of model inference calls per explanation, may present scalability challenges in real-time deployment settings. These limitations outline clear directions for future research, including multilingual dataset extension, multi-label classification, and the integration of more efficient XAI methods such as SHAP.

V. APPLICATIONS AND USE CASES

The proposed multi-emotion mental health classification system has several practical applications in healthcare, social media analysis, research, and artificial intelligence. By ana-lyzing emotions from Twitter data, the system can help under-stand emotional behavior, support mental health awareness, and improve the transparency of AI-based prediction systems.

A. Industry Use

The proposed system can be used in healthcare, social media platforms, and AI-based applications for emotion detection and mental health monitoring. Hospitals and mental health professionals can analyze social media data to detect early signs of emotional distress. Social media companies can use it to understand user behavior and trends. It can also improve chatbots, virtual assistants, and customer support systems by enabling emotion-aware responses. Explainable AI methods like LIME improve transparency and trust in predictions.

B. Social Impact

This study can contribute positively to society by increasing awareness about mental health issues through social media analysis. Many people express their emotions, stress, anxiety, and personal struggles online, and analyzing these emotional patterns can help identify individuals who may require emo-tional support or mental health assistance.

Early identification of emotions such as sadness, fear, and stress may help encourage timely intervention and support. The system can also help reduce stigma around mental health by promoting awareness and understanding of emotional well-being. Furthermore, the use of explainable AI improves trans-parency by helping users understand why a certain emotion was predicted, making the system more trustworthy and easier to understand.

C. Policy Relevance

The proposed framework can support mental health orga-nizations, healthcare authorities, and policymakers in under-standing emotional trends and public mental health concerns from social media platforms. The insights gained from emotion analysis can help in planning awareness campaigns, emotional wellbeing programs, and mental health support initiatives. The study also highlights the need for ethical and transparent AI in healthcare. Explainable AI ensures responsible and interpretable use of such systems.

D. Academic Value

This review paper contributes to academic research in the fields of Natural Language Processing, emotion detection, sentiment analysis, deep learning, and

explainable AI. The study reviews and compares different traditional machine learning models, deep learning architectures, and transformer-based approaches used for mental health classification from Twitter data.

The paper also highlights the importance of explainability in AI-based mental health systems and discusses current challenges, limitations, and future research opportunities. By summarizing existing research and identifying important gaps, the study can serve as a useful reference for future researchers working on interpretable and efficient emotion classification systems.

VI. CONCLUSION

This paper proposed an explainable deep learning frame-work for multi-emotion mental health classification from Twit-ter text, addressing three critical gaps identified in the existing literature: the absence of systematic comparative evaluation of multiple recurrent deep learning architectures on six-class emotion data, the lack of class-wise LIME explainability across all emotion categories, and the need for a preprocess-ing strategy that is appropriately differentiated for traditional machine learning and deep learning model families.

The proposed framework employs a Bidirectional GRU (BiGRU) model as the primary classifier, supported by a systematic comparative evaluation of eight models — three traditional machine learning classifiers and five deep learning architectures. A balanced dataset of 60,000 tweets, constructed through stratified sampling from the Twitter Emotions Dataset, was used to ensure fair and unbiased model training and evaluation. A dual text preprocessing pipeline was designed to preserve emotionally significant negation words for deep learning models while applying standard stopword removal for TF-IDF-based classifiers.

The BiGRU model is expected to achieve superior classifica-tion performance across all four evaluation metrics, confirming the effectiveness of bidirectional sequential processing for short social media text. The class-wise LIME analysis, applied across all six emotion categories — anger, fear, joy, love, sadness, and surprise — provides word-level interpretabil-ity that aligns with established linguistic and psychological markers of each emotion. This integration of explainability into the classification pipeline enhances the transparency and trustworthiness of the system, making it more suitable for clinical and research applications in the mental health domain. The findings of this study demonstrate that combining deep learning with explainable AI produces a system that is both accurate and interpretable, representing a meaningful step toward responsible AI-driven mental health monitoring from social media. Future work will focus on extending the framework to multilingual datasets, exploring multi-label emotion classification to handle co-occurring emotional states, and integrating SHAP alongside LIME for complementary global and local explainability. Real-time deployment of the monitoring pipeline with automated alert capabilities is also identified as a priority for subsequent research.

REFERENCES

[1] M. Mondal, S. Das, and A. Ghosh, “Emotion Detection from Social Media Posts,” arXiv preprint, arXiv:2302.05610, 2023.

[2] D. Ji, P. Zhou, and Y. Zhang, “MentalBERT: Publicly Available Pre-trained Language Models for Mental Healthcare,” in Proc. Int. Conf. Language Resources and Evaluation (LREC), Marseille, France, 2022, pp. 4382–4390.

[3] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, “GoEmotions: A Dataset of Fine-Grained Emotions,” in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 2020, pp. 4040–4054.

[4] V. Vajre, M. Nair, and S. K. Singh, “PsychBERT: A Mental Health Lan-guage Model for Social Media,” in Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 2021, pp. 1077–1082.

[5] A. S. Abdelghafoor, R. Alghamdi, and H. Aljohani, “Natural Language Processing for Mental Health Interventions: A Systematic

Review," Translational Psychiatry, vol. 13, no. 309, pp. 1–12, 2023.

[6] M. Nourani, J. T. Hancock, and V. S. Subrahmanian, "Machine Learning and NLP in Mental Health: A Systematic Review," Journal of Medical Internet Research, vol. 23, no. 5, e15708, 2021.

[7] R. Calvo, D. Milne, and S. Razavi, "Natural Language Processing for Mental Health Interventions," Frontiers in Digital Health, vol. 3, art. no. 646532, 2021.

[8] S. Chancellor, E. P. S. Baumer, and M. De Choudhury, "Who is the 'Human' in Human-Centered Machine Learning? The Case of Predicting Mental Health from Social Media," Proc. ACM on Human-Computer Interaction, vol. 3, no. CSCW, art. no. 152, pp. 1–32, 2019.

[9] Y. Rao, J. Li, and H. Xie, "Emotion Recognition of Social Media Users Based on Deep Learning," Multimedia Tools and Applications, vol. 82, no. 2, pp. 2345–2364, 2023.

[10] J. Eisenberg, A. L. Beatty, and C. M. Lee, "Mixed Emotion Extraction from Social Media Text," Data & Knowledge Engineering, vol. 149, art. no. 102259, 2024.

[11] M. T. Islam, A. K. Das, and M. S. Hossain, "Emotion Detection from Social Media Text Using Deep Learning," Int. J. Advanced Computer Science and Applications, vol. 14, no. 5, pp. 312–320, 2023.

[12] A. H. Khan, S. U. Khan, and F. U. M. Abbasi, "Sentiment Analysis for Mental Health Monitoring Using Hybrid Deep Learning," IEEE Access, vol. 10, pp. 12345–12356, 2022.

[13] S. Poria, N. Majumder, D. Hazarika, and R. Mihalcea, "Multimodal Emotion Recognition Using Deep Learning," Neural Computing and Applications, vol. 35, no. 1, pp. 123–140, 2023.

[14] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks," in Proc. 14th Int. Conf. Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 2011, pp. 315–323.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144.

[16] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[17] A. Aldayel and M. Magdy, "Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media," Proc. ACM on Human-Computer Interaction, vol. 3, no. CSCW, art. no. 210, pp. 1–20, 2019.

[18] U. Naseem, I. Razzak, and P. W. Eklund, "A Survey of Deep Learning for Mental Health Detection on Social Media," IEEE Trans. Computational Social Systems, vol. 9, no. 5, pp. 1322–1340, 2022.

[19] S. M. Mohammad and P. D. Turney, "Crowdsourcing a Word–Emotion Association Lexicon," Computational Intelligence, vol. 29, no. 3, pp. 436–465, 2013.

[20] L. H. Lee, C. H. Chen, and S. M. Lee, "Explainable Transformer-Based Model for Emotion Detection on Social Media," IEEE Access, vol. 11, pp. 24567–24580, 2023.

[21] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.

[22] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] K. Cho, B. van Merrie¨nboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation," in Proc. Conf. Empirical Methods in Natural Language

Processing (EMNLP), Doha, Qatar, 2014, pp. 1724–1734.

[24] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Transla-tion by Jointly Learning to Align and Translate,” arXiv preprint, arXiv:1409.0473, 2014.

[25] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, “SemEval-2018 Task 1: Affect in Tweets,” in Proc. 12th Int. Workshop on Semantic Evaluation (SemEval), New Orleans, LA, USA, 2018, pp. 1–17.