SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 9377-9387
Open Access. Online & Print

DOI: 10.5281/zenodo.12426919

DYNAMIC MULTIMODAL FUSION FOR MENTAL HEALTH DISORDER DETECTION FROM SOCIAL MEDIA DATA

Karnam Akhil[1], Manchikatla Srikanth[2*], Reshma Mulpuri[3], Koushik Shanker Nagula[4], Talhah Mohammed[5], Rajamahendravarapu Venkata giridhar[6]

[1,2]Assistant Professor, Department of CSE*
[3,4,5,6]Student, Department of CSE
[1,2,3,4,5,6]VNR Vignana Jyothi Institute of Engineering and Technology
[1,2,3,4,5,6]Hyderabad, India
Email: [1]akhilresearch18@gmail.com; [2]srikanth_m@vnrvjiet.in; [3]reshmachowdaryy27@gmail,com; [4]koushiknagula1@gmail.com; [5]talhah.mohammed.10@gmail.com; [6]giridharclash52@gmail.com

Received: 03/12/2025
Accepted: 13/03/2026

Corresponding Author: Manchikatla Srikanth (srikanth_m@vnrvjiet.in)

ABSTRACT

Rapidly growth in social media technologies has increased the possibilities of identifying mental health disorder types via the use of user-generated material. This study outlines the components of a multimodal deep learning platform to categorize individuals into pathological conditions of mental disorder, using data obtained from Reddit. The proposed methodology uses a variety of feature modalities to create a set of semantic and behavioural characteristics for determination of mental health disorder: text features from the textual property of the post title and content are converted into representations that reflect the semantics of text, providing a means of modelling textual properties via BERTweet, while temporal and metadata features (e.g. time of post; length of message; and content traits) are modelled through an LSTM for data with an associated temporal component. Unlike prior works that use static fusion, the proposed model introduces a dynamic gating-based adaptive fusion mechanism. The approach captures temporal behavioral patterns alongside semantic signals, enabling effective integration of heterogeneous features. The performance characteristics of the proposed methodology were evaluated using an open-source dataset collected from Reddit that has been portioned into Pathological and Non-Pathological classifications with an output precision of 0.7857; recall of 0.7908; and F1-score of 0.7866. A comparative evaluation and cross-validation were performed against baseline classification techniques, demonstrating the robustness of the proposed methodology when compared with traditional (text-only) and multimodal classification methods. The results demonstrate the need for the integration of behavioral features within and in addition to the textual features for effective identification of individuals with mental health disorders.

KEYWORDS: Mental Health Detection, Social Media Analysis, Multimodal Learning, BERTweet, LSTM, Cross-Modal Fusion, Explainable Artificial Intelligence, Reddit Dataset, Deep Learning, Natural Language Processing.

I. INTRODUCTION

Background and Motivation

One mental health disorder that has become a global social health concern that worries numerous individuals is depression. Depression is regarded as a complex mental disorder that makes people feel incessant deep melancholy and despair and lose interest in activities they engage in day-to-day and have impaired thinking skills. Clinical depression causes irreversible alterations in the thinking, behavior, physical well-being and quality of life of a person. Global health researches have shown that millions of people at one time or another in their lives experience depression which is a major cause of disability and low productivity all over the world. [1].

Multiple biological and psychological and social factors work together to create depression which affects people. Biological factors refer to the genetic predisposition of individuals which interacts with hormonal imbalances and neurological changes that affect brain chemistry. Psychological causes often involve prolonged stress, trauma, low self-esteem, and negative thinking patterns. Social and environmental factors such as isolation, financial difficulties, academic or professional pressure, and relationship issues also play a significant role. Digital environments and social media platforms enable users to display their emotional challenges through online platforms which have emerged as a new way. Users on these platforms unintentionally create mental health signals through their shared thoughts about loneliness and anxiety and depression [2].

The health sector is still struggling with getting depression proper diagnosis as people do not have proper knowledge on mental illnesses. The conventional diagnostic methods depend on three main elements which include clinical interviews and self-reporting questionnaires and expert evaluation methods. The approaches prove to be effective but they take a lot of time to implement and are also subjective limiting their possibility of growth. The fear of being judged by the society and lack of knowledge of mental health and treatment opportunities make people avoid getting professional treatment. Health professionals have challenges in diagnosing cases since 80 percent of individuals with mental disorders and 90 percent of individuals with substance use disorders are not diagnosed.

The increase in social media platforms like Reddit and Twitter and the forums on the internet have increased exponentially and researchers can scale user-generated content to make inferences on mental health. Users tend to reveal their emotions, their experience and struggle in everyday life in a very natural and crude manner. These valuable linguistic signals can be used to determine the psychological conditions that exist in these textual expressions. It is not possible to analyze because the amount of data is large in comparison to the usual operational levels and informal language and other types of writing are used. Automated systems have also been useful as the organizations need proper methods of diagnosing mental illness through text analysis.

Deep learning and Natural Language Processing (NLP) have passed to another level of their development as the latest accomplishments have made it possible to understand and learn human language less difficult. The Bidirectional Encoder Representations from Transformers (BERT) model and its versions on the basis of transformer architecture can be defined as being highly accurate in the ability of the model to analyze the interrelations between words within texts. The models enable researchers to determine latent linguist traits which are not within existing machine learning strategies. The analysis of people in terms of mental states could not be implemented only on documents written by them as it required additional information. Analysis of the user behavior that contains the posting patterns and active times of the user and the content of shared materials can be carried out to make a psychological assessment. Researchers also explore multimodal solutions that include text along with other non-text elements to address existing problems in their studies. The system utilizes linguistic elements along with user activity patterns and time-based information to form complete user activity profiles. The system also enhances mental health condition detection with its ability to trace user verbal expressions and their behavioral patterns during different time periods. Research indicates that using multiple data modalities in mental health classification yields more accurate results. In this context, our work created a multimodal deep learning framework which uses Reddit data to identify mental health disorders.The system uses more than 700000 Reddit posts that cover different mental health conditions to create its classification system which goes beyond traditional methods that dependon text data [3]. The system uses BERTweet which is a transformer-based language model that researchers developed to understand social media language for generating detailed semantic and emotional understanding of user content. The system uses text information while also utilizing additional data which

includes posting times and text lengths and content indicators which show NSFW status to create user behavior models. The multimodal learning system employs a dynamic gating-based fusion method for efficiently combining temporal features and text features while simultaneously capturing semantic and behavioral characteristics of each pair of objects within a dataset.

Objectives

1. To develop a robust deep learning model capable of classifying mental health conditions from social media text data.
2. To improve prediction accuracy through a combination of text features and behavioral and temporal metadata.
3. To assess the proposed system's performance using standard metrics: accuracy, precision, recall, and F1-score.
4. To use Explainable Artificial Intelligence (XAI) techniques for explaining model predictions and uncovering insights regarding which features influence mental health classification.

Acronyms

Table 1. Acronyms used in the study

Acronym	Full Form
NLP	Natural Language Processing
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
XAI	Explainable Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BERTweet	BERT model pre-trained on Twitter data
GCM	Gate-Controlled Multimodal
NSFW	Not Safe For Work

Mathematical Symbols

Table 2. Mathematical Symbols

Symbol	Description
x_i	Input text for i-th sample
t_i	Temporal feature vector
y_i	True label
H_i	BERTweet embedding
v_i^{text}	Text feature vector
v_i^{temp}	Temporal feature vector
z_i	Fused feature representation
g_i	Gating value
o_i	Output logits
W_t, W_p, W_g	Learnable weight matrices
f(.)	Classification Function
$\mathcal{L}$	Loss function
θ	Model parameters
η	Learning rate
∇	Gradient operator
σ	Sigmoid function
tanh	Hyperbolic tangent activation
N	Number of samples
C	Number of classes

Paper Organization

The remainder of this paper is organized as follows. Section II presents the literature survey of existing approaches and work on mental health detection. Section III describes the proposed methodology, about dataset, our proposed architecture, and algorithms. Section IV presents the experimental results and discussion, along with performance evaluation and analysis. Section V concludes the paper, and Section VI outlines the future scope of the proposed work.

II. LITERATURE SUMMARY

As the recent events in the field of social-media-based mental health-detecting techniques have shown, a considerable shift in the traditional application of a text-based approach to an innovative, deep-learning-based and multimodal system has already occurred. The first research was mainly focused on exploring the language patterns on user-generated texts but this approach failed to detect behavioral and temporal patterns related to mental health disorders. In order to mitigate these drawbacks, Saeed et al. [3] proposed a multimodal framework, uniting both textual and time-related features acquired during the posts created in Reddit, and it was discovered that the incorporation of pattern-based approach in regards to user activity can significantly enhance the classification procedure. On this idea, Saeed and Cha [4] expanded on the multimodal learning by adding BiLSTM and LSTM-based architectures and attention-based fusion to such a framework wherein the model is able to learn both behavioral and semantic dependencies successfully.

At the same time, several studies have made an effort to establish deep learning algorithms that can be used to improve the quality of classification in mental health detection. Dash et al. [5] embraced implementation of the LSTM network architecture with preprocessing methods such as stemming and lemmatization to address the problem of multi-class classification of social media data. Similarly, Shao et al. [6] had solved the dilemma of feature sparsity in high-dimensional social media data by introducing semantic preprocessing model and adaptive feature selection model models that resulted in improved model generalization and predictive accuracy. These directions, even though one can achieve considerable advancements in the conventional machine learning methodologies, are disadvantaged by the fact that they are

contingent upon a maneuver of fixed or flat text representations, which limits their ability to identify complex contextual interactions.

Transformer based architecture development has been making great contributions towards improving the performance of the mental health classification systems in that they have been able to better give a contextual interpretation of language. RoBERTa proved to be an effective tool in detecting subtle patterns in language and contextual dependencies of data present on Reddit that are difficult with traditional models as illustrated by Murarka et al. [7]. In a like manner, the application to BERT and RoBerta models [8] revealed that contextual embeddings show better performance in contrast to traditional and deep learning algorithms. Besides, with the recent advancement of large language models (LLMs), the topic of which is relevant to Shah et al. [9], has shown a high performance in terms of classification as they have the ability to model long-range dependencies and complex semantic connections. But, these models need a lot of computational power, and are therefore less applicable to real-time or resource-constrained models.

Along with increasing predictive power, there is an increasing interest in the area of merely increasing the model interpretability and transparency as well. The likelihood of using explainable artificial intelligence algorithms in the detecting depressions models introduced by Bao et al. [10] provided a more meaningful interpretation of the lifespan predictions, as well as increasing the validity of the predictions in the natural environment. Similarly, Islam et al. [11] proposed an ensemble transformer-based system having post-hoc explaining algorithms to enhance issues such as the unbalanced classes and explainability. Their style performed better in terms of putting in good performance in the decisions of minority classes and also excellent explanations to model decisions.

Moreover, there is interest in hybrid and multimodal approaches because there is an opportunity of the combination of different features representations. A hybrid deep learning algorithm that is integrated using convolutional neural networks (CNN) and long short-term memory (LSTM) networks was suggested by Tejaswini et al. [12] to absorb local and sequential properties of social media text. This strategy showed better performance than isolated models by taking advantage of complementary advantages of various architectures. These however are yet to incorporate the blend of advanced transformer based contextual embeddings and the multimodal fusion methods that encompass both the semantic and behavior dynamic to be adequately captivated.

It is necessary to mention that the literature analyzed displays the shift towards the modern machine learning methods to the innovative deep learning, transformers and multimodal modalities in mental health detection. Despite significant progress made with regards to the accuracy of classifications, as well as understanding the context, and interpretability, several of the issues still persist and these include high calculation, some heterogeneous features integration, and scale. Their inadequacies have led to a need to have valid and decipherable multimodal theories that can be used to make use of textual and behavioral data effectively. The current work is driven by these research gaps where it suggests a multimodal architecture which brings together transformer-based text representations and temporal feature modeling, and an efficient fusion mechanism, to result in better performance in the mental health classification tasks.

III. METHODOLOGY

3.1 Problem Statement

Given a dataset of social media posts, shown in Eq.1

$$D = \{(x_i, t_i, y_i)\}_{i=1}^{N}, \quad \text{Eq.1}$$

where

x_i represents the textual content (title and selftext),

t_i denotes the temporal and metadata features, and

$y_i \in \{1, 2, ..., C\}$ is the corresponding mental health class label, the objective is to learn a function (f(.)) that accurately predicts the class label for each input instance.

The textual input x_i is transformed into a contextual embedding using a transformer-based model, while the temporal features t_i capture user behavioral patterns such as posting time and content characteristics. These features are represented as shown in Eq.2

$$x_i \rightarrow v_i^{text} \ , \ t_i \rightarrow v_i^{temp} \quad \text{Eq.2}$$

The goal is to learn a mapping function: As shown in Eq.3

$$f(x_i , t_i) = y_i \quad \text{Eq.3}$$

where the fused representation (z_i) is obtained by combining textual and temporal features as Shown in Eq.4

$$z_i = g_i . v_i^{text} + (1-g_i) . v_i^{temp} \quad \text{Eq.4}$$

Here, g_i is a gating function defined as shown in Eq.5

$$g_i = \sigma(W_g . \tanh(W_t . v_i^{text} + W_p . v_i^{temp})) \quad \text{Eq.5}$$

The model is trained to minimize the classification loss function as shown in Eq.6

$$\mathcal{L} = -\sum_{i=1}^{N} (y_i \log(\text{Softmax}(o_i))) \quad \text{Eq.6}$$

3.2 Dataset Description

The study uses a dataset called "Mental Disorders Identification from Reddit NLP," which was procured through Kaggle [13]. It includes a very extensive corpus of user-created data collected from various subreddits related to mental health. Posts reflect actual user communication as people describe their interactive feelings and life events and how they relate to their mental health state. The corpus consists of around 700,000 individual user posts that have been labeled by researchers related to many different mental illness categories, such as anxiety, depression, bipolar disorder, borderline personality disorder (BPD), and schizophrenia. Because of the breadth/depth of mental illness types represented and the high density of data points available, this dataset serves as an excellent building block for developing data-rich classification algorithms.

Table 3. Dataset Details

Attribute / Property	Description
Dataset Name	Mental Disorders Identification (Reddit NLP)
Source	Kaggle
Total Records	Approx. 700,000+ posts
Number of Classes	6–7 mental health categories
Classes	Anxiety, Depression, Bipolar Disorder, BPD, Schizophrenia, Mental Illness
Platform	Reddit
Data Type	Text + Metadata

The dataset includes both text data, as well as records of behavioral patterns associated with the posting behaviors. This additional metadata provides more insight into the date/time behaviors of the users who post, allowing for richer analyses than would have been available if only using text data. The dataset must be pre-processed prior to modeling, i.e., to reduce noise, resolve missing values, and create a consistent textual representation. Preprocessing of the dataset is assumed to follow two phases, namely the creation of label encodings to represent each unique item as an integer, and the splitting of data into training and test sets.

3.3 Proposed Methodology

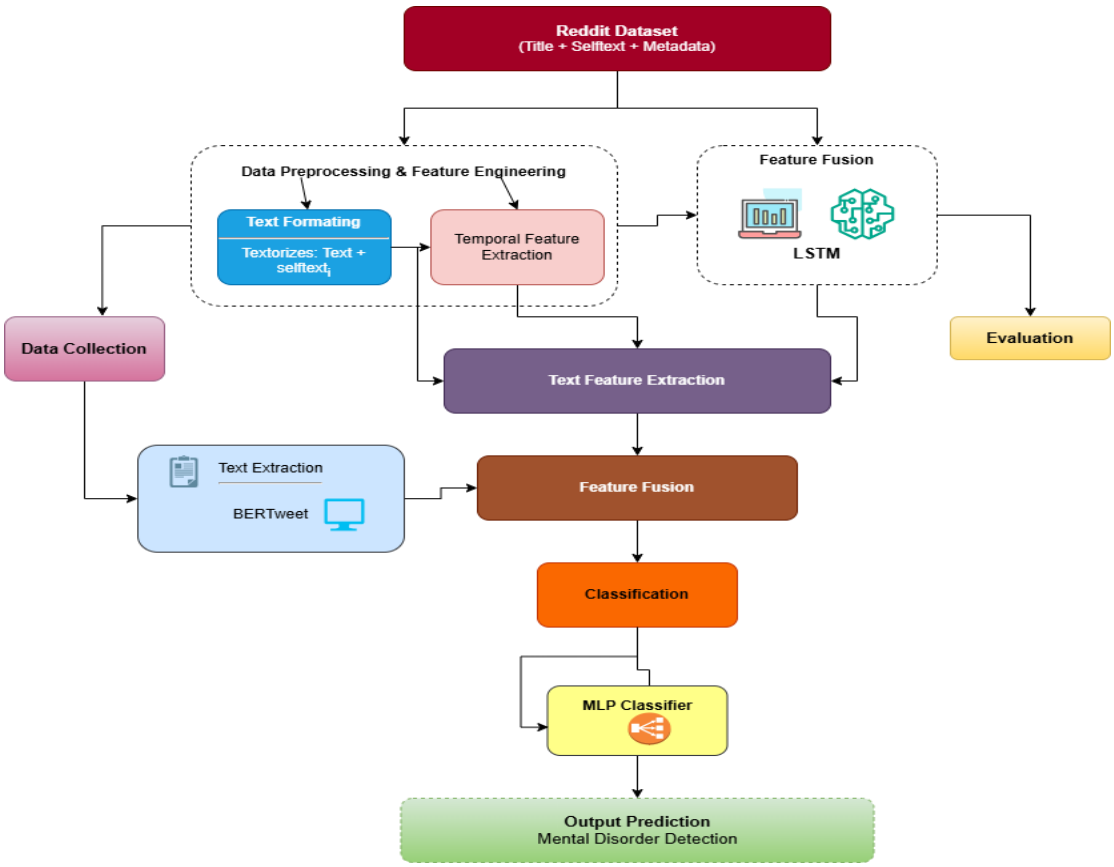

Figure 1. Proposed Multimodal Architecture for Mental Health Disorder Detection

The following architectural design is a multimodal architecture that incorporates both textual and behaviour-based data to classify mental health. An overview of an example of a Reddit dataset used in the proposed modelling is shown in Figure 1 and includes the textual content (title and selftext) with the associated metadata. Prior to using the data to classify mental health, the data will be pre-processed and featured. The textual data will be cleaned and formatted to remove any undesired characteristics, and the temporal-based characteristics that correspond to the time of posting and the behaviour of the users will be extracted.

The textual data will then be used to process through BERTweet via contextual embeddings to obtain semantic-based characteristics from social media data. Meanwhile, the temporal and metadata characteristics will also be processed via an LSTM network to model the sequential and behavioural characteristics. A dynamic gating-based fusion mechanism has been employed to combine the two types of feature representations by adaptively balancing the contribution from each type of feature to the final representation.

After combining the two resultant feature sets, the features will then be run through a classification module, a multi-layer perceptron (MLP). The MLP applies multiple non-linear transformations to learn complex relationships amongst the understanding of the data and will provide the output from which the final classification can be made (with probabilities associated with each of the underlying classifications).

3.4 Algorithms

Algorithm 1

Algorithm 1: Data Preprocessing and Feature Construction

Input:

Dataset $D = \{(title_i, selftext_i, ts_i, o_i, y_i)\}_{i=1}^{N}$

Output:

Processed dataset $D' = \{(input\_ids_i, mask_i, t_i, y_i)\}_{i=1}^{N}$

Procedure:

1. Foreach sample $i \in \{1, 2, ..., N\}$, construct textual input:

$x_i = concat(title_i, selftext_i)$

2. Convert timestamp ts_i into datetime format:

$dt_i = datetime(ts_i)$

3. Extract normalized temporal features:

$t_i = [\frac{hour(dt_i)}{23}, \frac{dow(dt_i)}{6}, \frac{len(xi)}{L}, o_i]$

4. Encode class labels:

$y_i \rightarrow \{0,1,...,C-1\}$

5. Tokenize textual input using BERTweet tokenizer:

$(input\_ids_i, mask_i) = Tokenizer(x_i)$

6. Construct processed dataset:

$D' = \{(input\_ids_i, mask_i, t_i, y_i)\}$

7. Split dataset into training and validation sets:

$D' = D_{train} \cup D_{val}$

The above algorithmdescribes the preprocessing of data & feature construction that is generally used to prepare the dataset prior to the model training as it ensures to improve predictive performance. As the first step , the textual components of each Reddit post that includes the title and selftext, are combined to form a unified input representation. Then the conversion of timestamp information is done into a standard datetime format, from which it extracts and normalizes relevant temporal features such as hour of posting and day of the week. In addition to this to capture user behavior patterns auxiliary features such as text length and content indicators are incorporated. BERTweet tokenizer is used to tokenize the text data to generate input IDs and attention masks suitable for transformer-based processing. Finally, a structured dataset is obtained and divided into training and validation sets, ensuring efficient learning and evaluation.

Algorithm 2

Algorithm 2: Multimodal Learning with Gated Fusion

Input:

Training and validation sets D_{train} , D_{val}

Output:

Optimal model parameters θ^*

Procedure:

1. Initialize model parameters θ^*, including BERTweet encoder, temporal LSTM, projection layers, and gating mechanism.
2. For each batch $(input\_ids_i, mask_i, t_i, y_i) \in D_{train}$:
3. Compute contextual embeddings using BERTweet:

$H_i = BERTweet(input\_ids_i, mask_i)$

4. Apply masked average pooling:

$v_i^{text} = \frac{\sum H_i \,.mask_i}{\sum mask_i}$

5. Encode temporal features using LSTM:

$v_i^{temp} = LSTM(t_i)$

6. Project both modalities into common space:

$z_i^{text} = W_t v_i^{text}$, $z_i^{temp} = W_p v_i^{temp}$

7. Compute gating value:

$g_i = \sigma(W_g \,.\, \tanh(z_i^{text} + z_i^{temp}))$

8. Fuse multimodal representations:

$z_i = g_i \,.\, z_i^{text} + (1 - g_i) \,.\, z_i^{temp}$

9. Compute logits and loss:

$o_i = f(z_i)$, $\mathcal{L} = -\Sigma\, y_i \log(Softmax(o_i))$

10. Update parameters using gradient descent:

$\theta = \theta - \eta\, \nabla\theta\, \mathcal{L}$

11. Repeat steps until convergence over all epochs.
12. Evaluate on validation set and select best model: $\theta^* = \text{argmax } F1(D_{val})$

The Algorithm 2 presents the proposed multimodal learning framework with gated fusion for mental health classification(depression). Here, the model uses BERTweet encoder to obtain contextual embeddings, while behavioral patterns are captured using LSTM network by modeling temporal and metadata features. A dynamic gating-based fusion mechanism has been employed to combine the two types of feature representations by adaptively balancing the contribution from each type of feature to the final representation. Then to predict the corresponding mental health category fused representation is passed through a classification network. Later the model is trained using a cross-entropy loss function and optimized through gradient-based updates. The selection of final model parameters are based on validation performance.

IV. RESULTS AND DISCUSSIONS

The effectiveness of the suggested multimodal mental health classification scheme is measured according to typical measures such as precision, recall, and F1-score. The model has a precision of 0.7857, recall of 0.7908 and an F1-score of 0.7866, which means balance and trustworthy classification performance between a variety of the classification classes. A set of experiments and visualizations are used in order to perform a further analysis of the model effectiveness.

Evaluation Metrics

Table 4. Evaluation results of Proposed model

Model	Precision	Recall	F1-Score
Our Proposed Model	0.7857	0.7908	0.7866

Comparison

Table 5. Results Comparison with baseline model

Model	Method	Features Used	F1 Score
CNN	Text-only	Text	0.7034
LSTM	Text-only	Text	0.7157
BiLSTM	Text-only	Text	0.7222
Base Model [3]	BiLSTM +LSTM + Attention	Text + Temporal	0.7376
Deep-AttentionModel [4]	Multimodal Attention	Text + Temporal	~0.75
Proposed Model	**BERTweet+LSTM + Gating**	**Text+Temporal + Metadata**	**0.7866**

We've evaluated the proposed model against both existing models (including the baseline) and those that have been previously evaluated on the same Reddit mental health dataset (Table 5). The traditional models (CNN, LSTM and BiLSTM) have a relatively low performance because they use only the text features of the data. A baseline multimodal model with temporal features from [3] had an F1-score of 0.7376. An extended version of this baseline multimodal model that used deep attention from [4] outperformed the baseline model. The proposed model produced an F1-score of 0.7866, thus outperforming both the baseline model and all other existing multimodal models. This improvement is attributed to transformer-based representations combined with adaptive fusion of semantic and behavioral features.

Training Dynamics

Figure 2 shows the change in the training and validation loss with the number of epochs. The loss in training is gradually decreasing which means that model parameters are well learnt. The validation loss first declines and afterwards levels off, indicating that the model reaches a convergence without much overfitting. Good generalization ability is indicated by the low difference between the training and validation loss.

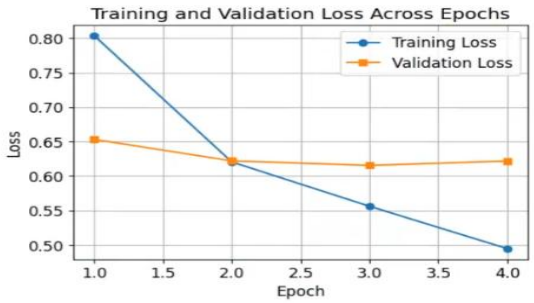

Figure 2: Training and Validation Loss across Epochs

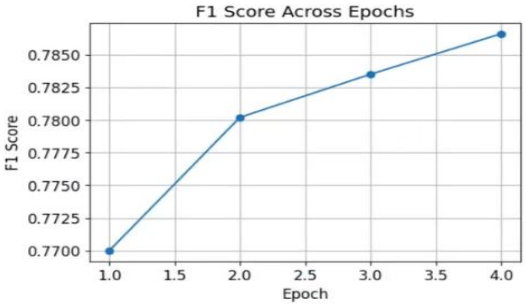

Figure 3: F1 Score across Epochs

As Figure 3 depicts, the F1-score becomes better and better throughout the epochs, which means that the model is able to learn to be discriminative with the course of time. The increasing trends depict stable training behaviour and correct optimization, and eventually results in better classification.

Model Behavior and Fusion Analysis

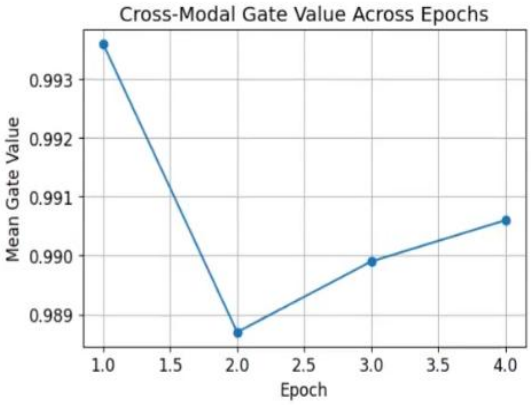

Figure 4: Cross-Modal Gate Value Across Epochs

Figure 4 shows how the values of cross-modal gates evolved as training progressed. The somewhat consistent values of the gates suggest that the model has a good balance on the input of the textual and temporal features. This is a confirmation that the gating mechanism is an important requirement to adapt multimodal information integration.

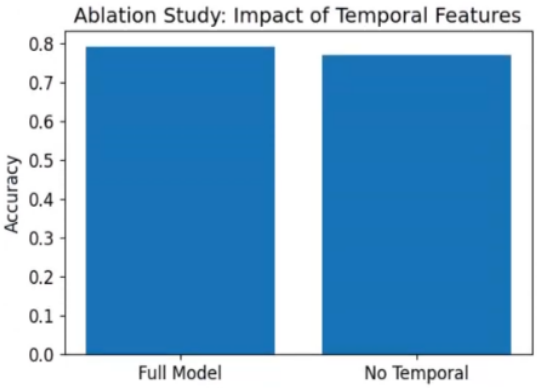

Figure 5: Ablation Study – Impact of Temporal Features

The ablation experiment in Figure 5 contrasts the results of the entire model with the results of the model without the use of temporal features. The duly noted increase in accuracy with incorporation of temporal features underscores the importance of the features in the process of acquisition of user behavior patterns. This can be interpreted to show that the prediction of the model is better when it uses temporal context compared to text-only models.

Feature Representation Analysis

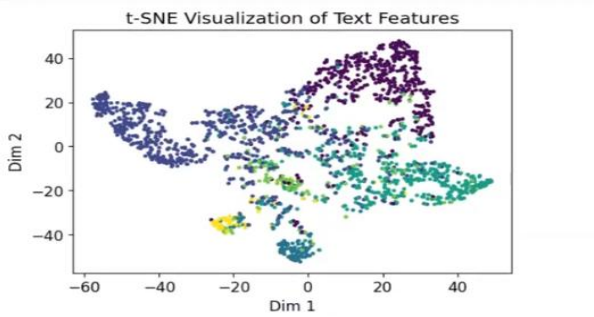

Figure 6: t-SNE Visualization of Learned Feature Representations

The t-SNE plot of the learned feature embedding is presented in figure 6. It is possible to observe distinctive clustering of the data points representing various classes, signaling the model to be learning meaningful and distinguishable representations. But some minor overlaps, between clusters, might imply natural similarities between some mental health conditions.

Performance Evaluation

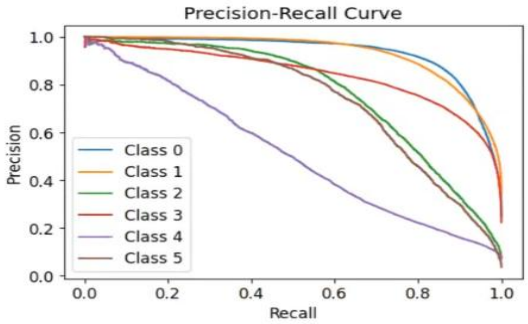

Figure 7: Precision-Recall Curve for Multi-Class Classification

Figure 7 above presents the precision-recall curves that illustrate the accuracy of model in the different classes. The majority of classes are very precise with a large variety of recall values, which suggests good performance in classification. The fact that the curves of some of the classes used to vary implies that there are some difficulties associated with the class imbalance & feature overlap.

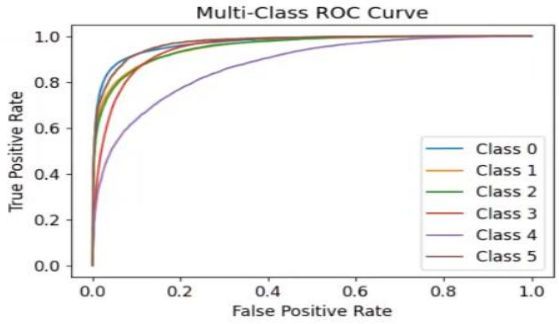

Figure 8: Multi-Class ROC Curve

Figure 8 shows all the classes Receiver Operating Characteristic (ROC) curves. The curves are clustered at the top-left and the false positive rates are low and high true positive rates imply a high concentration of the curves. This shows that the proposed model has a powerful discriminative capability in terms of several mental health categories.

In Figure 9, the confusion matrix gives a breakdown on the performance of classification. The scores of high values along the diagonal show the correct prediction accuracy of most classes. These misclassifications are mostly seen in between classes that overlap in any of their language or behavioral traits, which is most likely in the task of classifying in mental health because of the similarity.

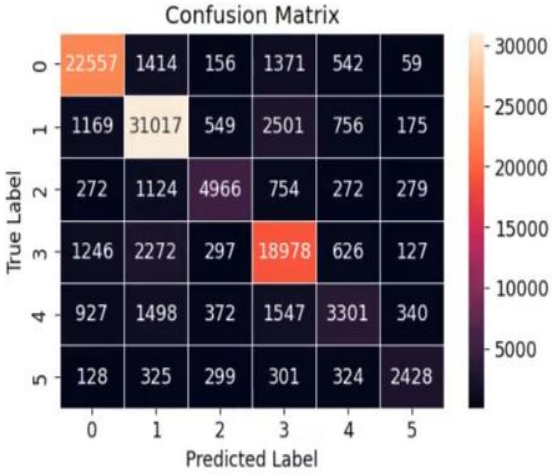

Figure 9: Confusion Matrix

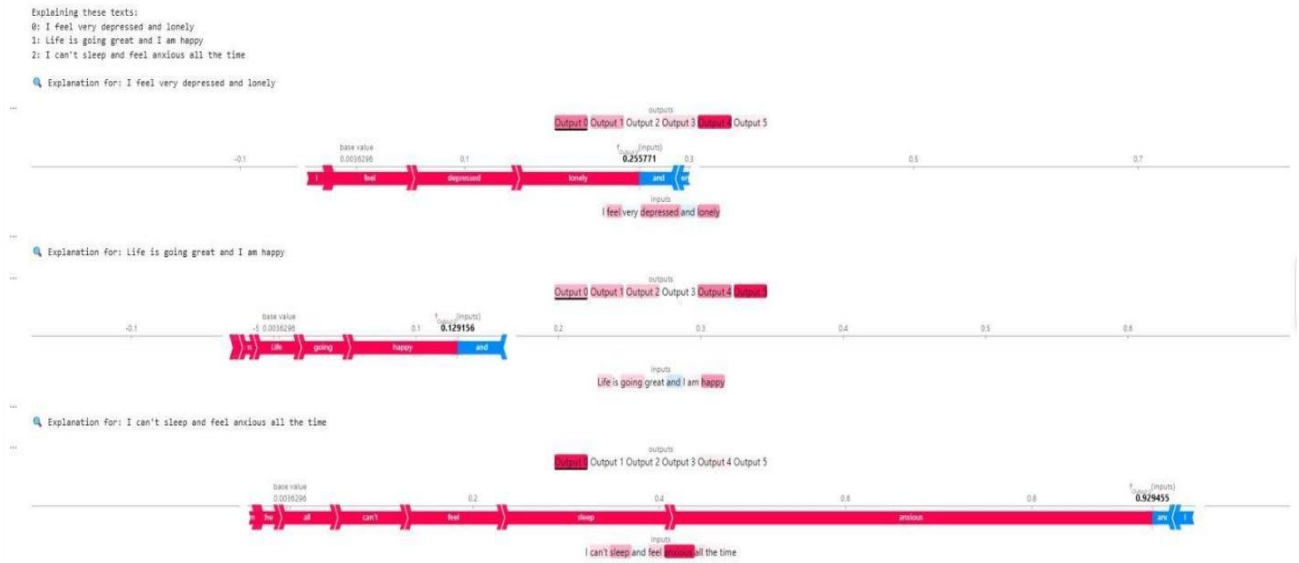

Figure 10. Explainable AI Visualization

Figure 10 above presents the explainability analysis which demonstrates the contribution of individual words towards the model predictions by an interpretable visualization method. Whether it was a positive contribution to a specific class or it was a negative influence, areas can be highlighted in blue or red, respectively. Indicatively, depressed and lonely, among other terms, highly predict mental health-related predictions, whereas positive terms like happy have a large impact on non-pathological-based predictions. This discussion has shown that the suggested model is able to obtain high predictive accuracy but also generate valuable and explainable information, which means that explainable artificial intelligence can be used in mental health detection.

The reason is that, the experimental findings clearly indicate the effectiveness of the proposed multimodal architecture in capturing the aspects of behavior and semantics of user data. The training dynamics substantiate constant convergence and good generalization, and the ablation study emphasizes the significance of temporal attributes towards performance enhancement. The visualization of the feature also confirms that the model is able to learn discriminative representations, and class separation becomes possible.

In addition, the precision-recall and ROC analyses reveal high classification performance in all the classes, whereas the confusion matrix will give information on the issues of classes. The fusion of cross-modal gating ensures a dynamic mix of feature fusion where different modalities can have a dynamic balance within the model. Despite the small numbers of misclassifications that still remain being caused by similar features of mental health disorders, the overall performance indicates that,

the proposed approach has an optimal trade-off of accuracy, robustness, and interpretability.

V. CONCLUSION

The paper provides a multimodal deep learning architecture of categorizing mental health, which is based on the data on social media. The given model includes admitting textual representations in the model of BERTweet and temporal and behavioral information through a dynamic gating-based fusion mechanism that integrates semantic and behavioral features, to provide the comprehensive perception of the content generated by the users. The experimental data shows that the model exhibits a balanced response with an F1-score of 0.7866 that allows on the one hand proving its usefulness in the characterization of both semantic and behavioral trends. The addition of time factors and cross-modal integration to performance significantly advances classification as compared to the conventional text only methods. All in all, the given system offers an effective and scalable way of detecting mental conditions automatically, which can be used in the early diagnosis and monitoring systems.

VI. FUTURE SCOPE

1. It is also possible to extend the model and introduce additional modalities such as the data on the interaction with the user, sentiments, and contextual metadata to increase performance in the classification.
2. Transformer architectures can be evolved to more advanced and bigger in size such that it may offer contextual significance and equally maximize on the availability of computational efficiency.
3. Continuous mental health monitoring systems can be developed with the help of integration of real-time data processing which may be used to perform the early intervention.
4. More explainable AI approaches may be incorporated, to increase transparency, and provide clinically significant explainable interpretations of model predictions.
5. The framework may be modified to pivot on cross-platform analysis in order to maximize generalization by merging the information on several social media such as Twitter, Reddit, and forums.

REFERENCES

[1] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. AAAI Conference on Weblogs and Social Media.
[2] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). Quantifying mental health signals in Twitter. CLPsych Workshop.
[3] Saeed, Q. B., & Cha, Y. (2025). Early detection of mental health issues using social media posts. Scientific Reports.
[4] Saeed, Q. B., & Cha, Y. (2025). Multi-modal deep-attention BiLSTM based early detection of mental health issues using social media posts. Scientific Reports.
[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.
[6] Shao, H., Zhu, M., & Zhai, S. (2025). Mental health diagnosis in the digital age: Harnessing sentiment analysis on social media platforms upon ultra-sparse feature content.
[7] Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2021). Classification of mental illnesses on social media using RoBERTa. Proceedings of ACL Workshop.
[8] BERT and RoBERTa models for enhanced detection of depression in social media text. (2024). Procedia Computer Science, Elsevier.
[9] Shah, S. M., Gillani, S. A., Baig, M. S. A., Saleem, M. A., & Siddiqui, M. H. (2024). Advancing depression detection on social media platforms through fine-tuned large language models. arXiv.
[10] Bao, E., Pérez, A., & Parapar, J. (2024). Explainable depression symptom detection in social media.
[11] Islam, S., Haque, R., Khan, M. A., et al. (2026). Ensemble transformer with post-hoc explanations for depression emotion and severity detection. iScience.
[12] Tejaswini, V., Sathya Babu, K., & Sahoo, B. (2024). Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions.
[13] Mental Disorders Identification (RedDIT). (2022, November 24). Kaggle. https://www.kaggle.com/datasets/kamaruladha/mentaldisorders-identification-reddit-nlp
[14] Franco-Martín, M. A., Muñoz-Sánchez, J. L., Sainz-de-Abajo, B., Castillo-Sánchez, G., Hamrioui, S., & de la Torre-Díez, I. (2018). A systematic literature review of technologies for suicidal behavior prevention. Journal of Medical Systems, 42(4).

[15] Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research. Journal of Personality and Social Psychology, 51(6), 1173.
[16] Bin Saeed, Qasim, and YoungJin Cha. "Multi-modal deep-attention-BiLSTM based early detection of mental health issues using social media posts." *Scientific reports* vol. 15,1 35152. 8 Oct. 2025, doi:10.1038/s41598-025-19141-0
[17] Losada, D. E., Crestani, F., & Parapar, J. (2017). CLEF 2017 eRisk overview: Early risk prediction on the internet. International Conference of the Cross-Language Evaluation Forum for European Languages, 3–15.
[18] Rissola, E. A., Losada, D. E., & Crestani, F. (2021). A survey of computational methods for online mental state assessment on social media. ACM Transactions on Computing for Healthcare, 2(2).
[19] Gkotsis, G., et al. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. Scientific Reports, 7(1), 1–10.
[20] Trotzek, M., Koitka, S., & Friedrich, C. M. (2020). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering, 32(3), 588–601.