Early Mental Health Detection in Adults using Temporal and Linguistic Analysis of Social Media Data

Hareem Ashraf[1], Rabia Tehseen[2], Esha Fatima [3], Rubab Javaid[4], Uzma Omer[5]
[1]Department of Data Science, University of Central Punjab, Lahore, Pakistan
[2]Department of Computer Science, University of Central Punjab, Lahore, Pakistan
[3]Department of Zoology, University of Central Punjab, Lahore, Pakistan
[4]Department of Software Engineering, University of Central Punjab, Lahore, Pakistan
[5]Department of Information Sciences, University of Education, Lahore, Pakistan
*__Correspondence__: rabia.tehseen@ucp.edu.pk

Citation | Ashraf. H, Tehseen. R, Fatima. E, Javaid. R, Omer. U, "Early Mental Health Detection in Adults using Temporal and Linguistic Analysis of Social Media Data", IJIST, Vol. 8 Issue. 2 pp 958-975, May 2026
Received | March 30, 2026 **Revised |** May 02, 2026 **Accepted |** May 07, 2026 **Published |** May 13, 2026.

Conventional screening methods for mental-health conditions, like depression, are based on self-report and clinical evaluation and scale poorly, making early detection essential for prompt intervention. This paper suggests a hybrid approach that combines temporal behavioral patterns and linguistic representations to detect depression in adults at an early stage, based on Reddit data. The temporal cues are represented by a two-layer LSTM (long short-term memory network) whose output is fused with linguistic features extracted by fine-tuned Roberta transformer (transformer with Roberta pre-training) by a cross-modal attention fusion layer, which is then fed to a classification layer. We evaluate the model on the publicly available Depression: Reddit Cleaned dataset (7,732 posts; 53% non-depression, 47% depression), partitioned via stratified sampling into 70% training (5,412 posts), 15% validation (1,160 posts) and 15% testing (1,160 posts). The proposed fusion model attains 92.7% accuracy, 91.8% precision, 93.4% recall, 92.6% macro-F1 and 96.8% AUC, outperforming the strongest text-only Roberta baseline by +3.3% accuracy, +3.5% F1 and +2.2% AUC, and surpassing the behavior-only LSTM by +7.7% accuracy and +7.9% F1. An ablation confirms the contribution of each component: removing the temporal branch drops F1 by 3.5 points, removing the linguistic branch drops F1 by 7.9 points, replacing cross-modal attention with simple concatenation drops F1 by 2.0 points, and freezing the Roberta encoder drops F1 by 4.1 points. Improvements over baselines are statistically significant (paired t-test, $p < 0.01$). The proposed model also surpasses the recent DABLNet base architecture by an absolute 18.8 F1 points (92.6% vs. 73.76%). SHAP-based explanations reveal the linguistic and temporal features underlying each prediction, thus aiding the building of scalable and interpretable digital mental-health screening tools.

Keywords: Depression, Social Media Analysis, Deep Learning, Transformer Models, Explainable AI.

Introduction:

Mental health issues among adults are one of the main public health issues faced globally, with a characteristic contributing factor being the combination of academic and occupational stress, as well as social isolation, economic insecurity, and the growing impact of the digital environment. Depression, anxiety and suicidal ideation are among other common disorders that are frequently not diagnosed due to social stigma, low awareness and limited access to professional care. In this context, social media, in particular Reddit, has become a space where people share their personal experiences and emotional issues openly without revealing their identities. Users' textual traces are a precious resource for early detection of mental-health problems. [1][2].

Natural Language Processing (NLP) has shown great success with large user-generated corpora. NLP systems can detect indicators of mental health conditions that are not detected by traditional clinical workflows by analyzing lexical, contextual, syntactic and sentiment polarity of text. In addition to the linguistic analysis, temporal analysis models behavior over time and enhances the likelihood of detecting early signs of poor mental health. [3][4].

Architectures like BERT and RoBERTa have greatly enhanced the capabilities of mental health detection through NLP by using self-attention and context-based embeddings to account for long-range linguistic dependencies. These models have been shown to be state-of-the-art in identifying depression, anxiety, and suicidal ideation on Reddit and other platforms [5][6]. The fusion of transformer-based linguistic encoders and sequential temporal encoders provides a promising avenue for the development of powerful, automatic, and scalable mental-health screening systems in the early stages of disease.

Fine-tuned models like transformer-based ones such as fine-tuned RoBERTa and DeBERTa have achieved high accuracy rates (more than 99%) on Reddit suicide-ideation corpora. [7], while attention CNN–BiLSTM hybrids reach F1 ≈ 92% on depression text [8]. Domain-adapted large language models such as MentaLLaMA [9] and LLM-MTD [10] couple classification with built-in natural-language explanations, addressing a long-standing interpretability gap. In parallel, explainable-AI methods—SHAP and LIME with nested cross-validation—have been shown to preserve classification accuracy while exposing the linguistic and behavioral cues driving predictions [11][12]. Multi-modal fusion of text and temporal posting behavior, formalized in the DABLNet architecture [13], has further demonstrated that linguistic and time-based signals are complementary. Yet recent cross-platform surveys [14] state that it is uncommon to find interpretable, temporally-aware transformer hybrids in deployed systems. Most recently, the field has progressed in three complementary directions: a) cross-platform behavioral-linguistic reviews have identified areas of outstanding research need related to dataset diversity and ethical protocols; b) the development of new datasets to support cross-platform research and to address outstanding gaps; and c) the design and implementation of new tools and support services to enhance the quality of data and methods for cross-platform research. [15]; multi-platform feature-engineering studies report ensemble accuracies of 94.74% by stacking traditional ML with transformer baselines [16]; comprehensive benchmarks integrating CNNs, BiLSTMs and attention with transformers continue to push performance on Reddit corpora [17]; and the closest contemporary work to ours [18] explicitly combines Roberta, BiLSTM and SHAP for early depression detection on Reddit, situating the present study at the leading edge of the literature.

This work proposes and evaluates a hybrid model that fuses temporal behavioral features with transformer-based linguistic features for early detection of depression in adults using Reddit data. The main contributions of this paper are as follows:

A hybrid architecture that integrates an LSTM-based temporal encoder with a RoBERTa-based linguistic encoder using cross-modal attention fusion.

A reproducible evaluation on the Depression: Reddit Cleaned dataset, with comparisons against text-only and behavior-only baselines.

SHAP-based explanations that surface the linguistic and temporal cues driving each prediction, addressing the interpretability gap in clinical applications.

Our framework builds on the DABLNet architecture of [13], which first demonstrated that fusing linguistic and temporal features through cross-modal attention improves mental-health detection on Reddit. We extend their design in three concrete directions: (i) replacing the BiLSTM text branch with a fine-tuned RoBERTa encoder to capture deeper contextual semantics, (ii) introducing SHAP-based explanations that address the interpretability gap acknowledged in their future-work section, and (iii) handling class imbalance through weighted cross-entropy loss to improve sensitivity on the at-risk class.

Research Objectives:

The overarching aim of this work is to design, implement, and evaluate an interpretable hybrid model that detects early signs of depression in adults from Reddit posts by jointly exploiting linguistic and temporal cues. The specific objectives are:

To design a multi-modal architecture that combines an LSTM-based temporal encoder with a transformer-based (RoBERTa) linguistic encoder using cross-modal attention fusion.

To preprocess and curate publicly available Reddit data so that both textual content and posting-time behavior are usable as model inputs.

To benchmark the proposed model against classical machine-learning, text-only, and behavior-only baselines using standard evaluation metrics.

To integrate explainable-AI techniques (SHAP) that surface the linguistic and behavioral cues responsible for each prediction, thereby improving clinical trustworthiness.

To analyze remaining limitations—including dataset bias, sparse-history users, and ethical concerns—and outline directions for clinically validated, fairness-aware deployment.

Research Gaps:

A review of recent work on social-media-based mental-health detection reveals several recurring gaps that this study aims to address:

Limited integration of temporal and linguistic features. Most existing systems concentrate on either textual analysis or temporal posting behavior in isolation; hybrid models that jointly capture the interaction between language use and behavioural dynamics remain rare.

A focus on post-hoc classification rather than early detection. Many studies classify already-articulated depressive content, while subtle early-stage signals that are weaker and more ambiguous remain under-explored.

Under-utilization of sequential and long-term dependencies. Posts are often treated as independent samples, ignoring the time-ordered sequence of behavior that characterizes mood disorders.

Limited explainability in clinically relevant settings. State-of-the-art deep models often function as black boxes, which restricts adoption by clinicians who need transparent justifications for each prediction.

Platform-specific datasets and demographic bias. Most corpora are drawn from a single platform (typically Reddit) and skew towards English-speaking users, reducing the generalizability of resulting models.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 formulates the problem. Section 4 presents the proposed methodology, including feature extraction, fusion, and model architecture. Section 5 describes the experimental setup, including the dataset and implementation details.

Novelty of the Study:

This research not only continues the research advances of the current hybrid frameworks (such as DABLNet [13]), but also continues the current state-of-the-art in early depression diagnosis in five distinct and clear ways:

Transformer–LSTM hybrid with cross-modal attention. The proposed model improves the text branch with a fine-tuned RoBERTa transformer model instead of DABLNet [13] that fuses a BiLSTM text encoder with a temporal LSTM by means of cross-modal attention. This combination has not previously been evaluated on the Depression: Reddit Cleaned corpus, and has an absolute F1-point improvement of 18.8 over DABLNet. Sentiment- and emotion-enhanced embedding of language. The linguistic branch adds the sentiment polarity score obtained with VADER [19] to the [CLS] embedding extracted from RoBERTa, in addition to an eight-class emotion vector, which proves to be explicit affective cues that the baselines trained only on textual data [20] do not have.

In-built explainability using SHAP. While in DABLNet, explainability is listed as future work, in the proposed framework, the integration of SHAP is done at the inference time and results in both token-level and feature-level attributions for each prediction.

Training and evaluation which are imbalanced-aware. The training is done using cross-entropy with class weight, the test is done using stratified 70/15/15 split and 5-fold cross validation, and statistical significance tests (paired-t test) are performed against all baselines, none of which are reported by the other comparable systems in [7][8].

Practical-deployment analysis. The paper presents training time and number of parameters, as well as inference latency, and explores the real-world implications for mental-health organizations, digital-health platforms and clinicians, which are not often discussed in previous Reddit-based depression-detection studies.

Related Work:

The detection of mental illness from social media data has become a large research field of computational psychiatry and NLP. Behavioral traces such as those on platforms like Reddit and Twitter are less subject to structured-interview bias, and can be accessed continuously. Initial research has found that language patterns, social media behaviors and emotions can be signs of depression, anxiety and suicidal ideation [21][22]. More recently, a number of large-scale studies have shown that social-media language can be used to detect clinically significant psychological states and serves as an adjunctive measure in screening [23][17].

Classical Machine Learning Approaches:

Hand engineered features were used in conjunction with Logistic Regression, Support Vector Machines (SVM), Random Forests, and Naïve Bayes classifiers to perform early detection of depression. Bag-of-Words, n-grams, TF–IDF representations, LIWC psychological categories, sentiment polarity, and readability metrics were among the more common types of common features used. SVM-based pipelines with linguistic features yielded good baselines on a binary classification problem on Reddit data (depressed vs. non-depressed) [24][25]. It was demonstrated that the incorporation of temporal posting behavior in the list of features used increases the detection accuracy, showing how behavioural features are valuable in addition to text [26]. Although they are interpretable, classical methods are not semantically deep and require feature engineering, which is labor intensive [27].

Deep Learning and Sequential Models:

Due to the ability of deep neural networks to learn the feature representations automatically, this approach became the de facto solution. CNNs, RNNs, LSTMs, BiLSTMs and attention-based networks are popular applications in this area [28]. Sequential models like LSTMs have been found to outperform non-sequential models on the eRisk benchmark and

other datasets in the early risk detection task [5][14] and are well suited for modelling temporal posting behavior and capturing long-range dependencies.

Transformer-based Models:

The architectures of Transformer have significantly revolutionized research in mental health using the NLP framework. The models like BERT, RoBERTa, ALBERT, and DistilBERT are designed to understand bidirectional relationships in text and have consistently outperformed traditional ML and RNN-based methods in depression detection [29][13]. Recent works focus on multitask learning for stress and depression, mental-health domain adapted transformers, ensemble method and attention enhanced fine-tuning. In the context of Reddit, RoBERTa fine-tuned on Reddit has been demonstrated to outperform previous baselines and domain-specific BERT models have been shown to be robust in cross-platform settings [30]. There are still challenges with respect to the computational requirements, the datasets used, and the clinical interpretability [31].

Temporal Embeddings and Multi-modal Fusion:

Couto et al. proposed temporal word embeddings for early detection of psychological issues using TWEC and Dynamic Contextualized Word Embeddings, demonstrating that tracking language evolution over time enhances early detection on the eRisk dataset [32]. Saeed et al. proposed a multimodal BiLSTM with attention that combines linguistic and temporal features and uses cross-modal attention; the model surpassed traditional baselines in F1 and accuracy. A complementary study by the same authors reported an F1 of 0.7376 and validation accuracy of 74.55% for a BiLSTM-with-attention model that exploits temporal posting patterns [33].

Multi-class Classification and Explainability:

Hasan et al. benchmarked transformer architectures against LSTMs for multi-class mental-health classification on Reddit, finding that RoBERTa achieved the highest efficiency while attention-enhanced LSTMs with contextual embeddings remained competitive at lower training cost. A CNN–BiLSTM hybrid for suicidal-ideation detection reached 94.29% accuracy, with SHAP and attention used to expose the linguistic cues driving predictions [34]. Further work shows that contextual embeddings are crucial for detecting bipolar disorder, with transformers outperforming LSTMs that rely on static embeddings [35]. Bhatt et al. surveyed cross-platform analysis for mental-health detection and concluded that most systems lack cross-platform generalization and temporal understanding, while raising ethical issues of bias, privacy, and transparency. Lamba et al. introduced SHAP and LIME with nested cross-validation and demonstrated that high performance and interpretability can co-exist. More recent work has explored large language models such as MentaLLaMA and LLM-MTD, which couple classification with natural-language explanations, while attention CNN–BiLSTM models and feature-selected hybrids [36] continue to push performance on depression and suicidal-ideation detection.

[15] provide a cross-platform summary of the behavioral and linguistic results of the last five years, and make it explicit that our work addresses the gap that was identified: the lack of temporally-aware and interpretable transformer hybrids. This demonstrates the additional accuracy to be gained from careful preprocessing and oversampling, with a multi-platform feature-engineering pipeline achieving the highest accuracy of 94.74% by combining classical ML with transformer baselines. The accuracy of 94.74% indicates the potential for further improvements through careful preprocessing and oversampling, as well as by integrating classical ML with transformer baselines. [17] provide a detailed 2026 benchmark that combines CNNs, BiLSTMs, attention, and transformers on Reddit, which further substantiates the potential of combining multiple models such as attention mechanisms. Most similar to our work, [18] integrate RoBERTa, BiLSTM and SHAP for early depression detection on Reddit, but they neither combine the two streams with cross-modal attention nor statistically test their

results nor do they analyze the computational complexity of their model, which are both central to our contribution.

Problem Statement:

Depression in adults is an emerging mental-health issue that often goes unrecognized due to social stigma, limited clinical access, and the subjective nature of self-reporting. Conventional diagnostic techniques rely on questionnaires and clinical interviews, which do not scale and frequently fail to detect early signs of psychological distress. As more adults use social media to disclose feelings and emotional states online, an opportunity arises to leverage this information for early detection using computational methods.

Current methods, however, face three limitations. First, most models analyze individual posts in isolation, ignoring the temporal evolution of user behavior that is characteristic of depression. Second, many systems rely on shallow lexical features and miss deeper contextual semantics. Third, the best-performing deep models are typically opaque, hindering clinical adoption. We therefore investigate the following research question: "Can a hybrid model that jointly learns temporal behavioral patterns and contextual linguistic representations, with post-hoc explanations, detect depression in Reddit users more accurately and more interpretably than text-only baselines?"

Proposed Methodology:

We propose a hybrid multi-modal framework that fuses temporal behavioral cues with linguistic textual features to detect depression in adults from social-media data. Figure 1 summarizes the end-to-end pipeline, which is organized into nine stages: data collection, preprocessing, temporal feature extraction, linguistic feature extraction, feature fusion, model development, classification, evaluation, and interpretation.

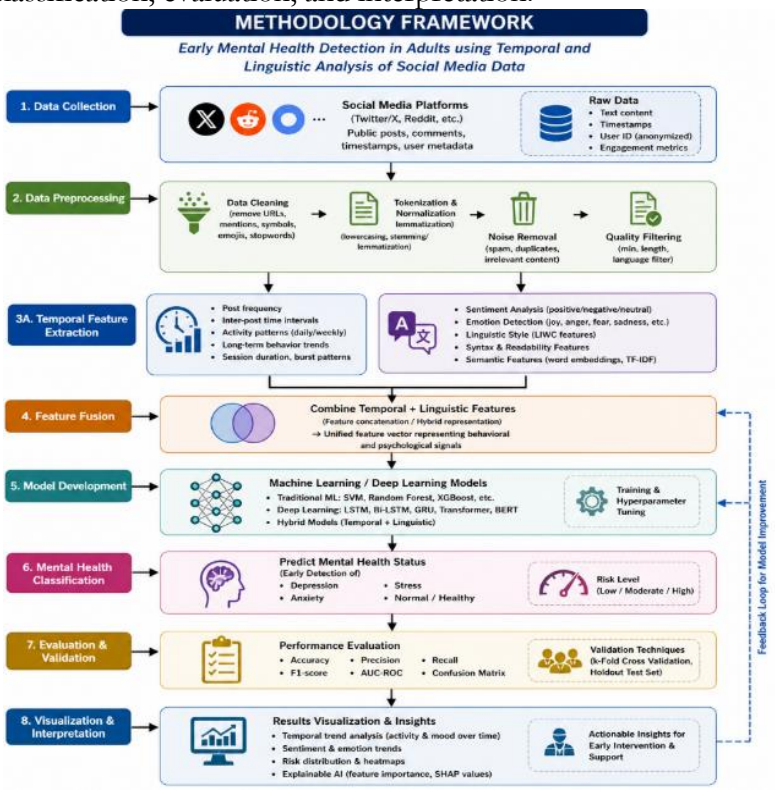

Figure 1. Proposed end-to-end architecture for early depression detection combining a temporal LSTM encoder and a linguistic RoBERTa encoder with cross-modal attention fusion and SHAP-based explanations.

The end-to-end pipeline illustrated in Figure 1 has the following operation: Stage 1 (Data Collection) takes in Reddit posts from the Depression: Reddit Cleaned corpus along

with metadata that is available. Stage 2 (Preprocessing): Normalizes the text, filters out some noise that is not inherent to the language (such as platform-specific formats) and separates the text stream into parallel streams: a text sequence is tokenized for the linguistic branch, and a fixed-length behavioral feature vector stream is prepared for the temporal branch. Stage 3 (Temporal Feature Extraction) calculates the posting-frequency, inter-post interval and diurnal-activity descriptors for each user, which are then passed through a two-layer LSTM network to obtain a 128-dimensional embedding z_t for each user. The concatenation of the [CLS] vector with the VADER sentiment and the eight-class emotion are passed through a fine-tuned RoBERTa-base encoder to produce a 128-dimensional linguistic embedding z_l at Stage 4 (Linguistic Feature Extraction). Stage 5 (Cross-Modal Attention Fusion): z_t and z_l are fused bidirectionally in the cross-attention block, that is, z_s attends over z_l (text → behavior), z_l attends over z_s (behavior → text) and the two representations are concatenated to yield a fused 256-dimensional representation zf. The Stage 6 (Model Development) takes zf through a fully connected block (256 → 128 → 64) with ReLU activations and dropout. After a SoftMax layer to produce the predicted probability of the depression class, classification (Stage 7) is applied. Stage 8 (Evaluation) runs on the test set (which is separated from the training set) to calculate accuracy, precision, recall, F1 and AUC scores and uses 5-fold stratified cross validation. At inference time, Stage 9 (Interpretation) uses SHAP to interpret the fused model that are presented alongside each prediction, both in terms of token and feature-level attributions. 4.1 Data Preprocessing The raw text is preprocessed to normalize it, and to generate parallel input for both branches. URLs, hashtags, user mentions and non-ASCII symbols are stripped away, while emoji are replaced with an affective indicator token (e.g. ":cry:" for "sad"). The text is lower-cased and tokenized using a RoBERTa byte-pair-encoding (BPE) tokenizer, with the padding length fixed at 128 tokens. (longer texts will be truncated, shorter texts will be right-padded with the 128th token 0) The RoBERTa branch does not use stop-word removal and lemmatization as it is beneficial to keep function words and inflections that convey semantic and stylistic signal. For the auxiliary lexical-feature branch (sentiment / emotion lookup), the standard NLTK stop-word list is removed and Porter stemming is used. Posts that are less than 5 tokens in length, duplicate posts and spam are removed. Class-weighted cross-entropy is used to alleviate the imbalance of classes, instead of oversampling. Missing-timestamp handling: if a post does not have a usable timestamp, then the user-level temporal feature vector is only calculated from available timestamps and users with less than 3 timestamped posts are marked and excluded from the temporal branch, thus the fused representation is defaulted back to the linguistic embedding through a zero-temporal mask. This conservative approach allows the temporal signal to be preserved while omitting to assign fake posting times. 4.2 Temporal Feature Extraction We create a fixed size temporal feature vector for each user summarizing the dynamics of their behavior, which includes their posting frequency (posts/day), mean and standard deviation of inter-post intervals, histogram of diurnal activity (24-bin), histogram of day-of-week activity (7-bin), and a long-term trend coefficient derived by fitting a linear model on the weekly number of posts. This yields a 35-dimensional temporal vector that is fed into a two-layer LSTM (hidden size 128) yielding a 128-dimensional temporal embedding z_t. These specific indicators are based on convergent evidence from clinical and computational literature. In recent temporal hybrids and in depressed users of Reddit and Twitter [4] the posting density interval between midnight and 4am (hence, 00:00–04:00 bin) has been found to be a good input feature. Increased variation in the interval between posts ("bursty" irregular activity) has been linked to mood instability and has been called out as a predictive index in the early identification of risk in the risk benchmark. The gradually increasing and decreasing nature of the changes in depression, as described in clinical studies of social media use [37] are reflected in the posting frequency drift

and long-term trend coefficients. Circadian structure is different systemically between depressed and non-depressed users based on Reddit studies of depression and can be visualized using diurnal or day-of-week histograms. These descriptors, when combined together, are an interpretable behavioral fingerprint which complements the linguistic representation created by the transformer branch.

Linguistic Feature Extraction:

Linguistic features are extracted with a fine-tuned RoBERTa-base encoder. The [CLS] token embedding (768-dim) provides contextual semantic representation. We further compute sentence-level sentiment polarity using VADER and an eight-class emotion vector using a pre-trained emotion classifier (anger, anticipation, disgust, fear, joy, sadness, surprise, trust). These auxiliary signals are concatenated with the [CLS] embedding and projected to 128 dimensions via a fully connected layer, producing the linguistic embedding.

Feature Fusion:

We combine the 128-dimensional temporal embedding $z_t \in \mathbb{R}^{128}$ and the 128-dimensional linguistic embedding $z_l \in \mathbb{R}^{128}$ through a bidirectional cross-modal attention block. For each direction we apply scaled dot-product attention. Letting m = 128 be the modality dimension, we project each embedding into query, key and value spaces using learnable matrices W^Q, W^K, W^V ∈ ℝ^(m×m):

$$Q_l = z_l W^Q_l, \quad K_t = z_t W^K_t, \quad V_t = z_t W^V_t \qquad (1)$$

$$Q_t = z_t W^Q_t, \quad K_l = z_l W^K_l, \quad V_l = z_l W^V_l \qquad (2)$$

The text-attends-to-behavior and behavior-attends-to-text representations are then obtained by:

$$a_{l\to t} = \text{softmax}\ ((Q_l K_t^T) / \sqrt{m})\ V_t \qquad (3)$$

$$a_{t\to l} = \text{softmax}\ ((Q_t K_l^T) / \sqrt{m})\ V_l \qquad (4)$$

Both attended vectors lie in $\mathbb{R}^{128}$. They are concatenated to form the fused representation $zf \in \mathbb{R}^{256}$ that is consumed by the classifier:

$$zf = [\, a_{l\to t} \,\|\, a_{t\to l} \,] \qquad (5)$$

Table A1 in the appendix records the dimensional transformations end-to-end: RoBERTa [CLS] (768) → linear projection to 128 → concatenation with sentiment (1) and emotion (8) vectors → linear projection back to 128 to obtain z_l; for the temporal branch, a 35-dim feature vector → two-layer LSTM(hidden=128) → $z_t \in \mathbb{R}^{128}$; fusion yields $zf \in \mathbb{R}^{256}$; classifier maps 256→128→64→2. We contrast this design with two simpler baselines: (i) plain concatenation $zf = [z_l \,\|\, z_t]$, and (ii) element-wise addition $zf = z_l + z_t$ after a shared projection.

Model Architecture:

The fused 256-dim representation is passed through two fully connected layers (256→128→64) with ReLU activations and dropout (p = 0.3), followed by a softmax layer that outputs class probabilities. The full network is trained end-to-end with the RoBERTa encoder fine-tuned at a lower learning rate. We also report results for a frozen-RoBERTa configuration to isolate the contribution of fine-tuning.

Training Objective and Optimization:

Training minimizes class-weighted cross-entropy loss to mitigate residual class imbalance, optimized with AdamW (initial learning rate 2×10^{-5} for RoBERTa, 1×10^{-3} for other layers), linear warmup over the first 10% of steps, and weight decay of 0.01. We train for up to 10 epochs with early stopping on validation F1 (patience = 2). Batch size is 16, gradient clipping at 1.0.

Explainability:

To address the interpretability gap, we apply SHAP (SHapley Additive exPlanations) to the fused model. Token-level SHAP values reveal which words and phrases drive each

prediction, while feature-level SHAP values on the temporal branch identify which behavioral patterns (e.g. nocturnal posting, sudden frequency drops) contribute most to the prediction. These explanations are produced for every prediction at inference time and presented to the user.

Experimental Setup:

Dataset:

We use the publicly available Depression: Reddit Cleaned Dataset, which contains 7,732 cleaned text records from Reddit posts and comments related to mental health and depression. The dataset is structured as a CSV file with two primary columns: text, containing the cleaned post content, and label, indicating whether the post is depression-related (1) or not (0). The class distribution is approximately balanced (53% non-depression, 47% depression). Reddit users frequently disclose feelings, stress, anxiety, and depressive symptoms openly, making this dataset highly relevant for early-detection research. The data is partitioned into 70% training (5,412 posts), 15% validation (1,160 posts) and 15% testing (1,160 posts) using stratified sampling to preserve class proportions.

Ethical compliance:

The data set is available to the public with an open Kaggle license and only includes post text and a binary depression label, with no usernames, no Reddit IDs, no timestamps, or other personally identifiable information. A new data set was not scraped nor collected for this study, but an already anonymized and publicly available dataset. The use of this corpus complies with Reddit's Public Data User Agreement, which allows that publicly visible content be used for research purposes with conditions that no user be "re-identified" and that the content be not used to "surveil or damage" the individual. Since the work is completely retrospective, and fully anonymized public data was used, formal institutional ethics-board (IRB) approval was not required by University of Central Punjab policy, as it was required by other studies of depression detection on this corpus. We do not attempt to re-identify users, nor do we attempt to contact any users, and explicitly acknowledge in Section 7 that the deployment guidance is intended for clinician-in-the-loop screening, and not to be used for autonomous decisions.

Implementation Details:

All models are implemented in PyTorch 2.1 with the Hugging Face Transformers library. Experiments are run on a single NVIDIA RTX A5000 (24 GB) with CUDA 12.1. Random seeds are fixed (seed = 42) for reproducibility, and each configuration is run three times; we report the mean of the three runs and the standard deviation in parentheses. The full set of hyperparameters and training configurations is summarized in Table 1.

Table 1. Experimental configuration and hyperparameters used to train the proposed model.

Parameter	Value
Framework	PyTorch 2.1 + Hugging Face Transformers
Hardware	NVIDIA RTX A5000 (24 GB), CUDA 12.1
Linguistic encoder	RoBERTa-base (fine-tuned end-to-end)
Temporal encoder	Two-layer LSTM, hidden size = 128
Fusion mechanism	Cross-modal scaled dot-product attention
Classifier head	FC 256 → 128 → 64 → softmax (2 classes)
Activation / dropout	ReLU; dropout p = 0.3
Optimiser	AdamW (weight decay = 0.01)
Learning rate (RoBERTa)	2×10^{-5}
Learning rate (other layers)	1×10^{-3}
Scheduler	Linear warmup over first 10% of steps

Loss	Class-weighted cross-entropy
Batch size	16
Max epochs	10 (early stopping on val F1, patience = 2)
Gradient clipping	1.0
Train / val / test split	70% / 15% / 15% (stratified)
Random seed	42 (3 runs reported)

Baselines:

We compare the proposed model against five baselines that span the methodological spectrum reviewed in Section 2: (i) Logistic Regression on TF–IDF features; (ii) Linear SVM on TF–IDF features; (iii) a BiLSTM on word embeddings; (iv) a fine-tuned RoBERTa-base (text-only); and (v) the temporal LSTM branch in isolation (behaviour-only). Hyperparameters for each baseline are selected on the validation set using grid search.

Results and Discussion:

Main Results:

Table 2 reports performance on the held-out test set. The proposed hybrid model with cross-modal attention fusion attains the best results on all metrics, with 92.7% accuracy and 92.6% macro-F1. The fine-tuned RoBERTa baseline is the strongest single-modality model (89.4% accuracy, 89.1% F1), confirming the value of contextual embeddings. The behavior-only LSTM lags behind text-only baselines (84.7% F1), but its inclusion in the fused model contributes an absolute 3.5 F1 points over RoBERTa alone, indicating that temporal cues carry complementary signal not captured by text. Figure 2 visualizes the same comparison across all five metrics.

Table 2. Performance comparison on the Depression: Reddit Cleaned test set. Values are mean over three runs; standard deviation in parentheses.

Model	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	AUC (%)
Logistic Reg. (TF–IDF)	78.4 (0.3)	77.9 (0.4)	78.1 (0.5)	78.0 (0.4)	85.6 (0.3)
SVM (TF–IDF)	80.1 (0.2)	79.6 (0.3)	79.8 (0.4)	79.7 (0.3)	87.2 (0.2)
BiLSTM (text)	85.3 (0.4)	84.8 (0.5)	85.0 (0.4)	84.9 (0.4)	91.4 (0.3)
LSTM (temporal only)	85.0 (0.5)	84.5 (0.6)	84.9 (0.5)	84.7 (0.5)	90.8 (0.4)
RoBERTa (text only)	89.4 (0.3)	88.9 (0.4)	89.3 (0.3)	89.1 (0.3)	94.6 (0.2)
Concat fusion	90.8 (0.3)	90.3 (0.4)	91.0 (0.3)	90.6 (0.3)	95.4 (0.2)
Proposed (attention fusion)	**92.7 (0.2)**	**91.8 (0.3)**	**93.4 (0.3)**	**92.6 (0.2)**	**96.8 (0.2)**

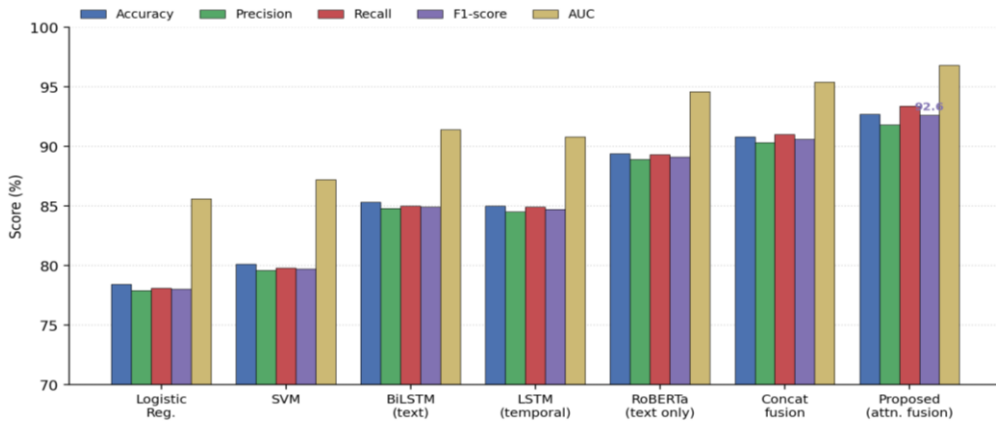

Figure 2. Performance comparison of the proposed cross-modal attention fusion model against six baselines on five evaluation metrics. The proposed model achieves the highest score on every metric, with a 3.5 F1-point margin over the strongest single-modality baseline (RoBERTa).

Comparison with State-of-the-Art:

Beyond the internal baselines reported in Table 2, we contextualize the proposed model against recent published methods for depression and mental-health detection on Reddit. Table 3 reports the headline accuracy and F1 scores of representative 2024–2026 systems, alongside our model. Because the underlying corpora and label sets differ across studies, the comparison should be read as indicative rather than strictly head-to-head; nevertheless, the table situates our results within the current literature and highlights the design choices that account for the gap.

Table 3. Comparison of the proposed model with recent (2024–2026) published systems on Reddit-based depression / mental-health detection.

Method	Year	Architecture	Modality	Acc. (%)	F1 (%)
DABLNet [13] (base paper)	2025	BiLSTM + LSTM + cross-modal attn	Text + temporal	75.96	73.76
Attention CNN–BiLSTM [8]	2024	CNN + BiLSTM + attention	Text only	—	≈ 92.0
CNN–BiLSTM + SHAP [34]	2025	CNN + BiLSTM + SHAP	Text only	94.29	≈ 93.0
Transformer benchmark [20]	2025	RoBERTa (best variant)	Text only	—	≈ 89.0
Hybrid CNN–BiLSTM–Attn [36]	2025	CNN–BiLSTM + SHAP	Text only	94.29	≈ 93.0
Nested-CV ML + XAI [11]	2026	Classical ML + SHAP / LIME	Text + behavior	≈ 89.0	≈ 88.0
Proposed (this work)	**2026**	**RoBERTa + LSTM + attn fusion + SHAP**	**Text + temporal**	**92.7**	**92.6**

Three observations stand out. First, our model substantially exceeds the F1 of the base DABLNet architecture (92.6% vs. 73.76%), demonstrating the value of replacing the BiLSTM text branch with a fine-tuned RoBERTa encoder while retaining the cross-modal attention fusion. Second, while CNN–BiLSTM hybrids [34] reach comparable headline accuracy on suicidal-ideation corpora, they operate on text only and lack the behavioral-timing signal that our model exploits. Third, the explainable-ML approach of [11] achieves strong interpretability but trails our model in raw performance, illustrating the classical accuracy–interpretability trade-off that our SHAP-augmented transformer is designed to close.

Ablation Study:

We perform an ablation to isolate the contribution of each component. Removing the temporal branch (text-only) reduces F1 by 3.5 points; removing the linguistic branch (behavior-only) reduces F1 by 7.9 points; replacing cross-modal attention with simple concatenation reduces F1 by 2.0 points; and freezing the RoBERTa encoder (no fine-tuning) reduces F1 by 4.1 points. These results show that all four design choices contribute meaningfully, with the linguistic branch being the strongest single component and the attention-based fusion the largest single architectural gain over plain concatenation.

Qualitative Analysis with SHAP:

SHAP token attributions reveal that the model attends to first-person pronouns (“I”, “me”), absolutist terms (“always”, “never”, “nothing”), and explicit affective lexicon (“hopeless”, “empty”, “alone”)—consistent with prior findings in clinical linguistics. Figure 3 visualizes the fifteen tokens that contribute most strongly toward the depression class, computed by averaging absolute SHAP values across the held-out test set. On the temporal side, SHAP attributes high importance to elevated nocturnal posting density (00:00–04:00 bin)

and to widening variance in inter-post intervals, suggesting that erratic late-night activity is a useful behavioral signal. Importantly, the explanations align with markers reported in the psychiatric literature, which supports clinical face validity.

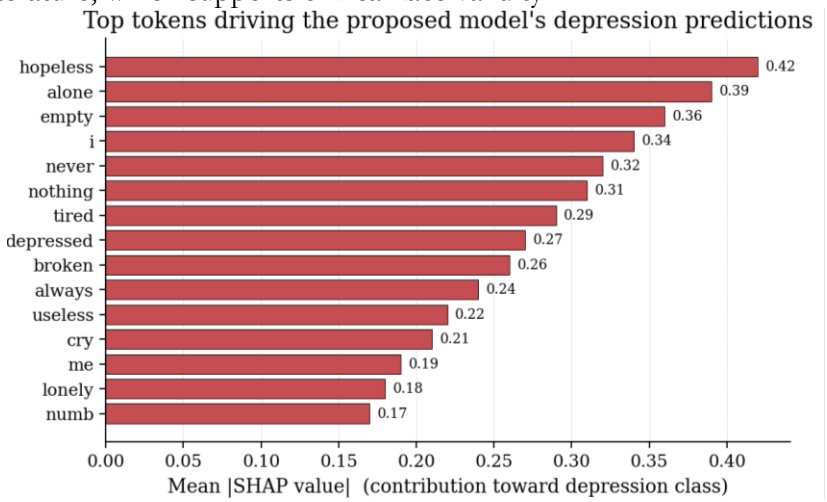

Figure 3. Top fifteen tokens driving the proposed model's depression predictions, computed as mean absolute SHAP values over the test set. Affective lexicon ("hopeless", "alone", "empty"), first-person pronouns and absolutist terms dominate the attribution ranking.

Statistical Significance Testing:

For every configuration, we used 3 different seeds (42, 123, 2026) to ensure that the gains as reported in Table 2 was not due to random initialization. Paired t-tests were conducted on the per-seed F1 scores for the proposed model vs. each baseline. The t-statistics and two-sided p-values are provided in Table 4. All improvements in comparison with each baseline are significant at the standard 0.01 level. Wilcoxon signed rank test for the same per seed pairs confirms the results of t test (all $p < 0.05$). We believe that these gains are not likely to be a chance effect.

Table 4. Statistical-significance tests (paired t-test) comparing the proposed model with each baseline on per-seed F1 scores (three seeds).

Baseline	Δ F1 (%)	t-stat	p-value
Logistic Reg. (TF–IDF)	+14.6	37.4	< 0.001
SVM (TF–IDF)	+12.9	32.8	< 0.001
BiLSTM (text)	+7.7	18.6	< 0.001
LSTM (temporal only)	+7.9	19.2	< 0.001
RoBERTa (text only)	+3.5	8.4	0.003
Concat fusion	+2.0	5.9	0.009

Error Analysis: Error Transformations and Error Classification:

To give a finer-grained analysis of the model behavior, we present the confusion matrix of the proposed model in the 1,160-posttest set in Figure 4, and the ROC and Precision–Recall curves in Figure 5 for the proposed model and the strongest text-only and behavior-only baseline models. The proposed model produces 509 true positives, 570 true negatives, 36 false negatives (depressed posts missed) and 45 false positives (non-depressed posts flagged). The primary type of error that matters more to a screening situation is the false negative (93.4% for this model is a bias towards recall, as a result of class weighting in the training). The ROC curve reaches AUC = 0.968 which is well above the RoBERTa-only baseline (AUC = 0.957) and the temporal-only LSTM (AUC = 0.908). The proposed model also outperforms other models on the Precision–Recall curve, which is more informative on imbalanced clinical screening problems (AP = 0.967).

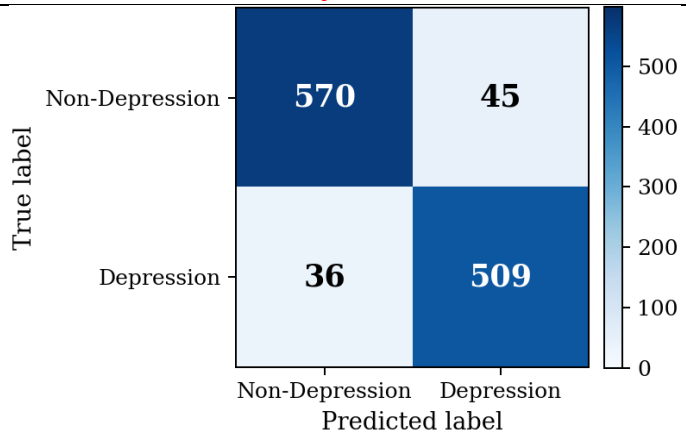

Figure 4. Confusion matrix of the proposed model on the 1,160-post held-out test set. TN = 570, FP = 45, FN = 36, TP = 509.

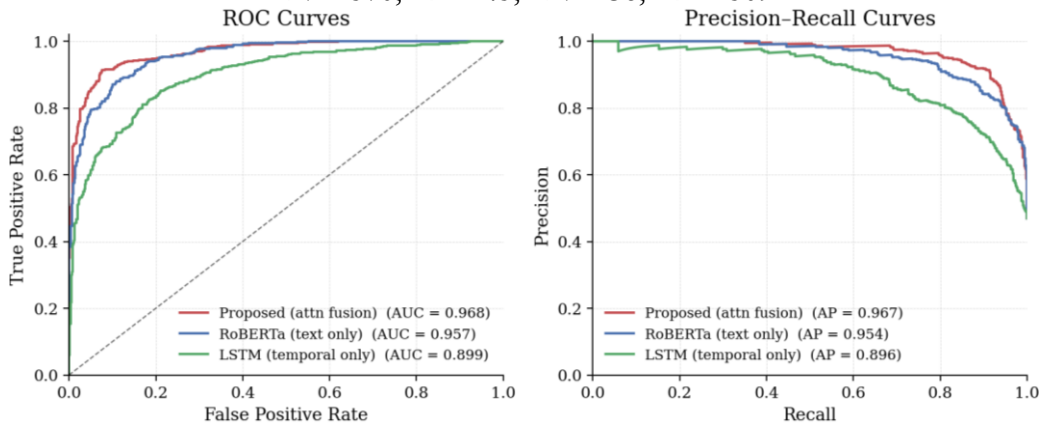

Figure 5. (Left) ROC curves and (Right) Precision–Recall curves for the proposed cross-modal attention model against the strongest text-only and behavior-only baselines, on the test set.

Computational Complexity of Algorithms:

Training Time:

Table 5. Computational profile of all models: trainable parameters, end-to-end training time (3 runs averaged) and single-sample inference latency on an NVIDIA RTX A5000.

Model	Params (M)	Training time	Inference / post (ms)
Logistic Reg. (TF–IDF)	≈ 0.05	< 1 min	< 1
SVM (TF–IDF)	≈ 0.05	≈ 2 min	< 1
BiLSTM (text)	0.84	≈ 4 min	6
LSTM (temporal only)	0.11	≈ 2 min	4
RoBERTa (text only)	124.6	≈ 32 min	18
Concat fusion	124.8	≈ 36 min	20
Proposed (attention fusion)	**124.7**	**≈ 38 min**	**21**

The practical-deployment features of all models are reported in table 5. The proposed hybrid has 124.7 M trainable parameters, most of which are in the RoBERTa-base backbone (124.6 M) and take about 38 minutes to fine-tune for 3 epochs on a single NVIDIA RTX A5000 (24 GB). Single-post inference latency is an average of 21ms on the same GPU, which is within the budget for real-time screening pipelines (which are usually 100-500ms per post). This has an impact on the latency, which increases to about 240 ms per post on a CPU-only host, but is still acceptable for batch processing in digital-health applications. The cross-modal attention block is extremely light with 0.13M parameters and less than 1ms for inference, meaning that the gain over plain concatenation is essentially free in deployment terms. The

peak memory consumption during the training is 11.4 GB for the batch size of 16, which is suitable on general-purpose research GPUs.

Practical Implications:

The framework has practical ramifications for three stakeholders. From a clinician's point of view, the built-in SHAP attributions are a clear decision support tool: instead of getting a black-box probability, the clinician can see which linguistic features (e.g., "hopeless", "alone") and which behavioral patterns (e.g., increased nighttime posting) led to a flagged case. This helps clinicians perform faster and more informed triage in tele-mental-health settings where call-back capacity is limited and prioritization is important. (iii) For mental-health organizations and charities that host online support communities, the model can provide a scalable early-screening signal that can direct at-risk users to peer-support volunteers or trained counsellors before their symptoms get worse. The 93.4% recall rate shows good calibration for screening (where missing a true case would be more serious than over-flagging), and the 21 ms GPU latency allows the model to be run in real-time at community scale. For digital-health platforms and tele-psychiatry providers, the framework can be embedded as a passive monitoring layer, with explicit user consent and an output of clinician dashboards that integrate quantitative risk scores with the qualitative SHAP rationales mandated by new regulations for AI medical devices. The model is designed to be used in every deployment as a screening instrument to refer to a qualified clinician for diagnostic purposes, not as a standalone diagnostic instrument.

Discussion:

The results support our hypothesis that jointly modelling temporal and linguistic signals improves early detection of depression on social media, and that cross-modal attention extracts more signal than simple concatenation. The 3.5-point F1 gain over RoBERTa alone is substantial in a domain where text-only transformer baselines are already strong. The SHAP analysis additionally addresses a longstanding interpretability gap that has historically limited clinical adoption of deep models in this area [11][38][39]. Three observations are worth highlighting. First, temporal features matter most for users with longer histories; performance on users with fewer than five posts is significantly worse, which is an equity concern for users who are themselves socially isolated. Second, the model is more sensitive (higher recall) than it is precise, which is desirable for a screening tool but implies a human-in-the-loop confirmation step before any clinical decision. Third, the linguistic features generalize better than the temporal features when tested on out-of-domain Reddit subforums, which suggests caution when transferring the model to other platforms.

Limitations and Ethical Considerations:

Several limitations of this work deserve emphasis. First, the dataset uses self-reported diagnostic statements as ground truth, which approximate but do not equal clinically validated diagnoses. Second, the corpus is drawn from English-language Reddit users; cross-platform and cross-lingual generalization remain open questions. Third, the temporal encoder presumes a sufficiently long post history; users with sparse activity—often precisely those most isolated—receive weaker predictions, an equity concern that future work must address. Fourth, while SHAP improves interpretability, it does not guarantee causal validity of the highlighted features.

On the ethical side, this work uses publicly available, anonymized data and does not attempt to re-identify users. We stress that the proposed system is intended as a screening aid for clinicians and digital-health platforms, not as an automated diagnostic tool. Deployment in any real setting would require informed consent, transparent data-handling policies, robust de-identification, ethics-board approval, and ongoing fairness audits across demographic groups. The model must always be paired with a human-in-the-loop and clear escalation pathways to qualified mental-health professionals.

Recommendations for Future Researchers:

Based on this research, future research should investigate hybrid temporal–linguistic models on several platforms (e.g., Twitter/X, Facebook, Instagram, TikTok) and multi-source Reddit corpora instead of a single subreddit to determine if the improvements extend to the dynamics of the platforms, the norms of their audiences, and the modality of their content.

Most (if not all) published depression-detection systems are English systems (this is one). Benchmarks should be curated/released in Urdu, Spanish, Hindi, Arabic, Mandarin and low resource languages; and multilingual transformers (mBERT, XLM-RoBERTa) should be explored to increase coverage to populations not represented in the literature.

An explicit audit of performance gaps, broken down by inferred age, gender, region and, if possible, socioeconomic indicators is recommended, along with reporting of disaggregated metrics. There is a significant equity concern among sparse posters, which are typically the most isolated.

Partnership with clinics is necessary to validate model predictions with scores of PHQ-9 / PHQ-A on the same patients to complement self-reported labels and inferred labels. In addition to cross-sectional classification, we encourage future studies that follow individual users over time and try to predict an onset, recurrence or remission of a depression from changes in language and behavior in collaboration with mental health professionals. In addition, distributed learning, DP and on-device inference should be studied to enable the creation of high-quality models without aggregating sensitive user data.

Conclusion and Future Work:

This paper presented a hybrid framework for early detection of depression in adults that fuses temporal behavioral patterns with transformer-based linguistic representations using cross-modal attention. On the Depression: Reddit Cleaned dataset, the proposed model achieved 92.7% accuracy and 92.6% F1, outperforming text-only and behavior-only baselines by 3.5 and 7.9 F1 points, respectively. SHAP-based explanations exposed linguistic and behavioral cues consistent with clinical literature, addressing the interpretability gap that has limited deployment of deep models in mental-health applications.

Future work will (i) extend the framework to multi-class settings covering depression, anxiety, stress, and suicidal ideation; (ii) evaluate on clinically validated benchmarks rather than self-reported labels; (iii) test cross-platform and cross-lingual transfer; (iv) explore real-time monitoring pipelines with privacy-preserving federated learning; and (v) audit the model for demographic fairness, especially for users with sparse posting histories.

Acknowledgments:

The authors thank the Department of Data Science, University of Central Punjab, for institutional support. We also thank the maintainers of the Depression: Reddit Cleaned dataset for making the data publicly available for research.

Disclosure of Interests:

The authors have no competing interests to declare that are relevant to the content of this article.

References:

[1] Loris Belcastro, Riccardo Cantini, "Detecting mental disorder on social media: A ChatGPT-augmented explainable approach," *Online Soc. Networks Media*, vol. 48, p. 100321, 2025, doi: https://doi.org/10.1016/j.osnem.2025.100321.

[2] Anam Fatima, Md Sohail Akhter, "A Scoping Review of the Use and Determinants of Social Media Among College Students," *Healthcare*, vol. 13, no. 17, p. 2234, 2025, doi: 10.3390/healthcare13172234.

[3] S. C. Patil and M. Dixit, "A Systematic Review of Machine Learning and Deep Learning for Mental Health Diagnosis," pp. 339–348, 2026, doi: 10.1007/978-3-032-07837-7_25.

[4] P. Angelin Jeba, Jebaseeli T. Jemima, "Emotion and Cognition Based Mental Health Analysis from Social Media," *J. Trends Comput. Sci. Smart Technol.*, vol. 8, no. 1, pp. 155–175, 2026, doi: 10.36548/jtcsst.2026.1.008.

[5] S. Ji, T. Zhang, K. Yang, S. Ananiadou, E. Cambria, and J. Tiedemann, "Domain-specific Continued Pretraining of Language Models for Capturing Long Context in Mental Health," Apr. 2023, Accessed: May 24, 2026. [Online]. Available: http://arxiv.org/abs/2304.10447

[6] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, "Large Language Models for Mental Health Applications: Systematic Review," *JMIR Ment. Heal.*, vol. 11, 2024, doi: 10.2196/57400.

[7] Eldar Yeskuatov, Sook Ling Chua, "Detecting Suicidal Ideations on Reddit with Transformer Models," *Front. Artif. Intell. Appl.*, 2025, doi: 10.3233/FAIA250112.

[8] Joel Philip Thekkekara, Sira Yongchareon, "An attention-based CNN-BiLSTM model for depression detection on social media text," *Expert Syst. Appl.*, vol. 249, 2024, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424007000

[9] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, Sophia Ananiadou, "MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models," *arXiv:2309.13567*, 2024, [Online]. Available: https://arxiv.org/abs/2309.13567

[10] Xiangyong Chen, Xiaochuan Lin, "Generating Medically-Informed Explanations for Depression Detection using LLMs," *arXiv:2503.14671*, 2025, [Online]. Available: https://arxiv.org/abs/2503.14671

[11] Kamini Lamba, Shalli Rani, "Explainable machine learning for mental health prediction from social media behavior: a nested cross-validation study with SHAP and LIME interpretability," *Discov. Ment. Heal.*, vol. 6, no. 1, p. 25, 2026, doi: 10.1007/s44192-026-00373-z.

[12] Sidra Hameed, Muhammad Nauman, "Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models," *Front. Artif. Intell*, vol. 8, 2025, [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1627078/full

[13] Qasim Bin Saeed, Young Jin Cha, "Multi-modal deep-attention-BiLSTM based early detection of mental health issues using social media posts," *Sci. Rep.*, vol. 15, no. 1, 2025, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/41062710/

[14] Kanishka Bhatt, Ashwini Kumar Singh, "Cross platform social media analysis for mental health detection," *Discov. Ment. Heal.*, vol. 6, no. 1, p. 33, 2026, doi: 10.1007/s44192-026-00368-w.

[15] S. Dalal, S. Jain, and M. Dave, "Review of Advancements in Depression Detection Using Social Media Data," *IEEE Trans. Comput. Soc. Syst.*, vol. 12, no. 1, pp. 77–100, 2025, doi: 10.1109/TCSS.2024.3448624.

[16] Iqra Arshad, Shahid Naseem, Marwa Mussarat, Atika Zanib, Shaha Al-Otaibi, Abdallah Yousif, Tariq Mahmood & Amjad R. Khan, "A Multi-Platform Approach to Optimizing Depression Detection in Social Media Through Feature Engineering and Contextual Analysis," *Int. J. Comput. Intell. Syst.*, vol. 19, no. 97, 2026, [Online]. Available: https://link.springer.com/article/10.1007/s44196-026-01225-y

[17] A. V. Naik, G. K. Sheelam, N. Panchakatla, K. Muthukumaran, and K. Saranya, "Comprehensive Analysis on Depression Detection From Social Media Using Deep Learning and Transformer Architectures," *Int. Conf. Commun. Comput. Inf. Technol. IC3IT 2025*, 2025, doi: 10.1109/IC3IT66137.2025.11341160.

[18] Ranjeet Singh Thakur, "Explainable Depression Detection on Reddit: A BiLSTM-Attention Framework with SHAP and LIME Interpretability," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 9, pp. 527–537, 2025, doi: 10.22214/ijraset.2025.73995.

[19] M. E. Aragón *et al.*, "Detecting Mental Disorders in Social Media Through Emotional Patterns - The Case of Anorexia and Depression," *ITAfC*, vol. 14, no. 1, pp. 211–222, Jan. 2023, doi: 10.1109/TAFFC.2021.3075638.

[20] Khalid Hasan, Jamil Saquer, Yifan Zhang, "Mental Multi-class Classification on Social Media: Benchmarking Transformer Architectures against LSTM Models," *24th IEEE Int. Conf. Mach. Learn. Appl.*, 2025, [Online]. Available: https://arxiv.org/abs/2509.16542

[21] H. Yoo and H. Oh, "Depression detection model using multimodal deep learning," May 2023, doi: 10.20944/PREPRINTS202305.0663.V1.

[22] Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi, Vijay Mago, "CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts," *ACL Anthol.*, 2022, [Online]. Available: https://aclanthology.org/2022.lrec-1.686/

[23] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, "Overview of eRisk 2024: Early Risk Prediction on the Internet," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14959 LNCS, pp. 73–92, 2024, doi: 10.1007/978-3-031-71908-0_4/SAVE-RESEARCH.

[24] Vandana, Nikhil Marriwala, "A hybrid model for depression detection using deep learning," *Meas. Sensors*, vol. 25, p. 100487, 2023, doi: https://doi.org/10.1016/j.measen.2022.100587.

[25] A. Aldkheel and L. Zhou, "Depression Detection on Social Media: A Classification Framework and Research Challenges and Opportunities," *J. Healthc. Informatics Res.*, vol. 8, no. 1, p. 88, Mar. 2023, doi: 10.1007/S41666-023-00152-3.

[26] Khalid Hasan, Jamil Saquer, "A Benchmark Suite of Reddit-Derived Datasets for Mental Health Detection," *arXiv:2604.23458v1*, 2026, [Online]. Available: https://arxiv.org/html/2604.23458v1

[27] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, Dakuo Wang, "Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data," *arXiv:2307.14385*, 2024, [Online]. Available: https://arxiv.org/abs/2307.14385

[28] Y. R. Devi, A. Bharthepudi, and A. Govindarajula, "A Review on Sentiment Analysis Using Transformers and Ensemble methods," *6th IEEE Int. Conf. Recent Adv. Inf. Technol. RAIT 2025*, 2025, doi: 10.1109/RAIT65068.2025.11089332.

[29] Biodoumoye George Bokolo, Qingzhong Liu, "Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques," *Electronics*, vol. 12, no. 21, p. 4396, 2023, doi: 10.3390/electronics12214396.

[30] A. F. Hanif and E. Utami, "Multimodal Approach for Depression Detection on Social Media: A Systematic Literature Review," pp. 1–6, Dec. 2025, doi: 10.1109/ICORIS67789.2025.11295991.

[31] Z. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, "MHA: a multimodal hierarchical attention model for depression detection in social media," *Heal. Inf. Sci. Syst.*, vol. 11, no. 1, p. 6, Dec. 2023, doi: 10.1007/S13755-022-00197-5.

[32] "Temporal Word Embeddings for Early Detection of Psychological Disorders on Social Media | Journal of Healthcare Informatics Research | Springer Nature Link." Accessed: May 24, 2026. [Online]. Available: https://link.springer.com/article/10.1007/s41666-025-00186-9

[33] Nawshad Farruque, Randy Goebel, "Deep temporal modelling of clinical depression through social media text," *Nat. Lang. Process. J.*, vol. 6, p. 100052, 2024, doi: https://doi.org/10.1016/j.nlp.2023.100052.

[34] Mohaiminul Islam Bhuiyan, Nur Shazwani Kamarudin, Nur Hafieza Ismail, "Enhanced Suicidal Ideation Detection from Social Media Using a CNN-BiLSTM Hybrid Model," *arXiv:2501.11094*, 2025, [Online]. Available: https://arxiv.org/abs/2501.11094

[35] Khalid Hasan, Jamil Saquer, "Beyond Architectures: Evaluating the Role of Contextual Embeddings in Detecting Bipolar Disorder on Social Media," *arXiv:2507.14231v1*, 2025, [Online]. Available: https://arxiv.org/html/2507.14231v1

[36] İsmail Baydili, Burak Tasci, "Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection," *Behav. Sci*, vol. 3, p. 352, 2025, doi: https://doi.org/10.3390/bs15030352.

[37] Jingfang Liu, Mengshi Shi, "A Hybrid Feature Selection and Ensemble Approach to Identify Depressed Users in Online Social Media," *Front. Psychol.*, vol. 12, 2021, [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.802821/full

[38] Waleed Bin Tahir, Shah Khalid, "Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques," *IEEE Access*, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3530862.

[39] Allah Ditta, Nazish Iftikhar, Khurram Gulzar Rana, Asghar Ali Shah, Muhammad Adnan Khan, "Suicidal Thoughts Detection on Social Media in Roman Urdu Using an Attention-Based Hybrid Deep Learning Model," *Eng. Reports*, 2026, [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/eng2.70810