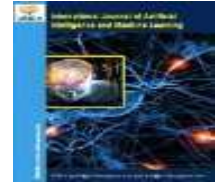




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open

Mental Health User Classification from Chronological Reddit Posts Using Transformer–Recurrent Hybrid Models

Refiloe Raselane¹, Satyananda Reddy²

¹Department of Computer Science and Systems Engineering, Andhra University College of Engineering Visakhapatnam, Andhra Pradesh, India – 530003. E-mail: 324206415019@andhrauniversity.edu.in

²Department of Computer Science and Systems Engineering, Andhra University College of Engineering Visakhapatnam, Andhra Pradesh, India – 530003. E-mail: satyanandau@yahoo.com

Abstract

Social media platforms contain valuable longitudinal information about users' linguistic patterns and behavioural changes, enabling automated mental health analysis. However, most existing approaches classify individual posts independently and fail to capture the chronological dependencies present in user-generated content. This study proposes a user-level mental health classification framework, MentalBERT–Temporal-BiLSTM, that fuses domain-specific contextual post embeddings with normalized inter-post time intervals and processes the resulting chronological representation with a bidirectional Long Short-Term Memory (BiLSTM) network. The framework is evaluated on 16,336 Reddit users across six classes—ADHD, Asperger's syndrome, depression, OCD, PTSD, and normal—achieving 90.64% accuracy and a 90.10% macro F1-score, and outperforming non-temporal and unidirectional temporal recurrent baselines. The results indicate that jointly modelling semantic content, posting-interval behaviour, and bidirectional sequential dependencies improves user-level classification over post-level and unidirectional approaches.

Keywords: Mental health classification; MentalBERT; temporal modelling; BiLSTM; user-level representation learning; social media mining.

I. Introduction

Mental disorders represent a serious global health burden: the World Health Organization estimated that nearly 15% of the world's population lived with a mental disorder in 2021, and a considerable share of them do not receive adequate treatment [1]. The growing use of social media has made large volumes of user-generated language available for studying mental health outside the conventional clinical setting.

Early computational approaches relied on traditional machine learning with handcrafted linguistic features such as LIWC-based lexical indicators [2], which offered useful insight but were limited by predefined feature sets and could not capture complex contextual variation [3]. Deep learning, and later transformer architectures such as BERT and its domain-adapted variants, enabled automatic learning of contextual representations from raw text and improved detection of mental-health-related signals [4],[5]. Nevertheless, many studies still treat individual posts as independent classification units, overlooking that mental-health-related expression evolves across a user's posting history rather than appearing in a single message [6]. Recent work has begun to incorporate temporal information [7],[8], but typically applies it after post-level prediction—e.g., through temporal windows or post hoc sentiment trajectories—rather than learning a single, unified, user-level representation end to end.

This paper proposes a domain-specific Transformer–Recurrent hybrid framework, MentalBERT–Temporal-BiLSTM, for user-level mental health classification from chronological Reddit posts. The framework generates contextual post representations with a mental-health-specific BERT model, fuses them with normalized inter-post time intervals, and applies bidirectional LSTM sequence modelling with global average pooling to learn longitudinal, user-level patterns. The main contributions are: (i) a user-level chronological modelling framework that classifies mental-health categories from sequences of posts rather than isolated posts; (ii) fusion of domain-specific BERT representations with explicit temporal information; (iii) bidirectional LSTM sequence modelling with global average pooling for a unified user representation; and (iv) evaluation on a six-class Reddit dataset covering ADHD, Asperger's syndrome, depression, OCD, PTSD, and normal users.

II. Related work

A. Traditional and Deep Learning Approaches

Early studies combined handcrafted linguistic features—most notably LIWC—with classifiers such as SVM, Logistic Regression, and Random Forests. Tausczik and Pennebaker [2] showed that word usage reflects psychological state, and De Choudhury et al. [16] and Coppersmith et al. [17] demonstrated that Twitter-derived linguistic and behavioural patterns can predict depression and PTSD before clinical diagnosis. Systematic reviews [18],[19] noted, however, that reliance on manual feature engineering limited these methods' ability to capture contextual semantics and evolving language, motivating deep learning approaches capable of automatic representation learning [15].

B. Transformer-Based and Domain-Adapted Models

Transformer self-attention overcomes the long-range dependency limitations of CNNs and RNNs and has been applied to detect depression and suicidal language [4],[5]. Because general-purpose transformers under-represent informal, disorder-specific vocabulary, domain adaptation has been explored in biomedical (BioBERT [10]) and clinical (ClinicalBERT [11]) text, and subsequently in mental health text through MentalBERT [9] and the double-domain-adapted DisorBERT [12]. Hybrid architectures such as the opinion-enhanced BERT+CNN+BiGRU model of Hossain et al. [13] enrich transformer output with sentiment and sequential components, but most of these models still analyse individual posts rather than a user's full chronological history.

C. Temporal Modelling

Villa-Pérez et al. [14] showed that the choice of historical window substantially affects depression-classification performance, and Shaina et al. [8] proposed post-level sentiment/thematic classification followed by temporal trajectory analysis. In both cases temporal information is applied after, rather than jointly with, representation learning, so a unified sequential user-level representation is not learned end to end.

D. Research Gap

Table I summarizes representative studies. Traditional models rely on handcrafted features with no temporal modelling; domain-specific transformers (BioBERT, ClinicalBERT, DisorBERT) improve contextual understanding but classify posts independently; and existing temporal approaches attach chronological analysis after post-level prediction rather than learning it jointly. Consequently, the relationships among post semantics, inter-post timing, and dependencies across a user's chronological history remain insufficiently explored—the gap this paper addresses through the MentalBERT–Temporal-BiLSTM framework.

Study	Representation	Level	Temporal
Coppersmith et al. [17]	Handcrafted (LIWC)	User	No
DisorBERT [12]	Domain transformer	Post	No
Hossain et al. [13]	BERT+CNN+BiGRU	Post	Limited
Villa-Pérez et al. [14]	Historical windows	User	Post hoc
Shaina et al. [8]	Sentiment trajectory	Timeline	Post hoc
Proposed	MentalBERT+BiLSTM	User	End-to-end

TABLE I. COMPARISON OF REPRESENTATIVE STUDIES

III. PROBLEM STATEMENT

Let $U = \{u_1, \dots, u_N\}$ denote the set of Reddit users, each assigned a label $y_i \in \{1, \dots, C\}$ with $C=6$ (ADHD, Asperger's syndrome, depression, normal, OCD, PTSD). For user u_i , let $S_i = \{p_1, \dots, p_T\}$ be the chronologically ordered sequence of posts, where p_t is the t^{th} post and T is the sequence length. Rather than treating each p_t as an independent observation, the goal is to learn a single user-level function $f(S_i) \rightarrow y_i$ that jointly captures (a) the semantic content of each post, (b) the time interval between consecutive posts, and (c) dependencies across the ordered sequence, addressing the tendency of post-level and post-hoc temporal methods to under-model these three sources jointly.

IV. Methodology

The proposed pipeline (Fig. 1) has seven stages: chronological user-post aggregation, MentalBERT contextual encoding, temporal feature projection, semantic-temporal fusion, bidirectional LSTM sequence modelling, global average pooling, and multi-class classification.

A. Dataset and Preprocessing

Disorder-specific posts (ADHD, Asperger's syndrome, depression, OCD, PTSD) were drawn from the Reddit Mental Health Posts dataset [20], with general AskReddit posts [21] used as a normal control group. Only users who posted more than once were retained, giving 16,336 users (Table II). Because roughly 95% of users posted 9 times or fewer, each user's sequence was fixed at $T=10$ posts: users with more than 10 posts were truncated to their earliest 10, and shorter sequences were zero-padded, preserving chronological order for batch-wise BiLSTM training.

Class	Users
ADHD	4,520
Asperger's syndrome	2,365
Depression	1,895
Normal	1,840
OCD	3,815
PTSD	1,901
Total	16,336

TABLE II. CLASS DISTRIBUTION OF THE USER-LEVEL DATASET

B. MentalBERT Contextual Embedding

Each post $p_{i,t}$ is WordPiece-tokenized ($L=128$ tokens, covering $\sim 90\%$ of posts) and passed through MentalBERT [9], a transformer pretrained on mental-health Reddit discourse. The final-layer [CLS] representation is taken as the post embedding:

$$\mathbf{e}_{i,t} = \text{MentalBERT}(p_{i,t}) \in \mathbb{R}^{768}$$

Embeddings are stacked chronologically per user, $E_i = \{\mathbf{e}_{i,1}, \dots, \mathbf{e}_{i,10}\} \in \mathbb{R}^{10 \times 768}$, giving a 10×768 user-level semantic tensor.

C. Temporal Feature Construction

For each user, inter-post intervals are computed as

$$\Delta t_{i,t} = \begin{cases} 0, & t = 1, \\ t_{i,t} - t_{i,t-1}, & t > 1. \end{cases}$$

then standardized with training-set Z-score normalization:

$$\tau_{i,t} = \frac{\Delta t_{i,t} - \mu_{\Delta t}}{\sigma_{\Delta t}}$$

giving a normalized temporal sequence $\tau_i = \{\tau_{i,1}, \dots, \tau_{i,10}\}$ aligned with E_i .

D. Semantic-Temporal Fusion and Sequence Modelling

Each scalar $\tau_{i,t}$ is projected to a 32-dimensional space and concatenated with the 768-dimensional embedding:

$$\mathbf{h}_{i,t}^{(\text{temp})} = \text{ReLU}(\mathbf{W}_\tau \tau_{i,t} + \mathbf{b}_\tau) \in \mathbb{R}^{32}$$

$$\mathbf{x}_{i,t} = [\mathbf{e}_{i,t} || \mathbf{h}_{i,t}^{(\text{temp})}] \in \mathbb{R}^{800}$$

The fused sequence $X_i = \{\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,10}\}$ is passed to a BiLSTM (128 hidden units per direction); forward and backward states are concatenated at each step:

$$\mathbf{h}_{i,t} = [\vec{\mathbf{h}}_{i,t} || \overleftarrow{\mathbf{h}}_{i,t}] \in \mathbb{R}^{256}$$

Global average pooling over the output sequence yields a fixed-length user representation $\mathbf{z} \in \mathbb{R}^{256}$, which is passed through a dense layer (256 units, ReLU), dropout ($p=0.3$), and a softmax layer over the six classes: $\hat{y}_i = \text{Softmax}(\mathbf{W}_o \cdot \mathbf{z} + \mathbf{b}_o)$

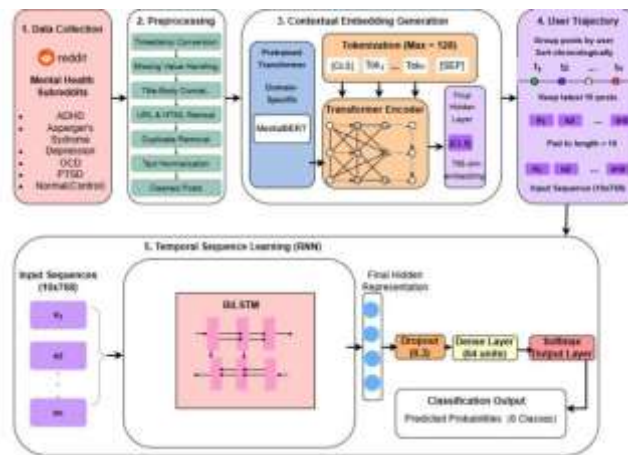


Fig. 1. Overall architecture of the proposed MentalBERT-Temporal-BiLSTM framework.

E. Training

The model was trained in TensorFlow with the Adam optimizer (initial learning rate 2×10^{-4}), sparse categorical cross-entropy loss, batch size 16, and a maximum of 15 epochs, using early stopping (patience 3 on validation loss) and learning-rate reduction on plateau. Class weights computed from the training set were incorporated into the loss to counter class imbalance. Data were split at the user level (stratified) into 72% training, 8% validation, and 20% test, using a fixed random seed (42) to prevent leakage across posts from the same user.

F. Evaluation Metrics

Performance was measured with overall accuracy, $\text{Acc} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$, and, because the dataset is class-imbalanced, macro-averaged $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$, $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$, and $\text{F1} = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall})$, each computed per class and averaged with equal weight. Confusion matrices were additionally examined to identify class-specific misclassification patterns.

V. EXPERIMENTAL SETUP

Experiments ran on Kaggle with two NVIDIA Tesla T4 GPUs; MentalBERT embeddings were generated once with PyTorch/Hugging Face Transformers and reused across all recurrent models, trained in TensorFlow/Keras, with Scikit-learn used for partitioning, class weighting, and evaluation. All preprocessing, splits, sequence length, and seeds were held constant across experiments for fair comparison.

Two baseline families were evaluated under identical MentalBERT embeddings and data partitions: (i) non-temporal recurrent models—LSTM, GRU, BiGRU—receiving only chronological MentalBERT embeddings; and (ii) temporal recurrent models—Temporal-LSTM, Temporal-GRU, Temporal-BiGRU—receiving the fused 800-dimensional semantic-temporal representation. The proposed Temporal-BiLSTM uses the same fused input with bidirectional processing and global average pooling. Performance was measured with accuracy, macro-averaged precision/recall/F1, and confusion-matrix analysis.

VI. Results and discussion

A. Overall Performance

On the independent test set, the proposed MentalBERT-Temporal-BiLSTM achieved 90.64% accuracy and a 90.10% macro F1-score. The close agreement between accuracy and macro F1 indicates relatively consistent performance across the six classes despite their unequal frequencies, supporting the value of chronological, user-level representation over independent post-level classification.

B. Comparison with Baselines

Model	Temporal	Acc. (%)	Macro F1 (%)
LSTM	No	88.98	88.26
GRU	No	80.02	79.86
BiGRU	No	89.93	89.31
Temporal-LSTM	Yes	90.15	89.80
Temporal-BiGRU	Yes	90.30	89.80

Model	Temporal	Acc. (%)	Macro F1 (%)
T-BiLSTM (Proposed)	Yes	90.64	90.10

TABLE III. COMPARISON BETWEEN RECURRENT BASELINES AND THE PROPOSED MODEL

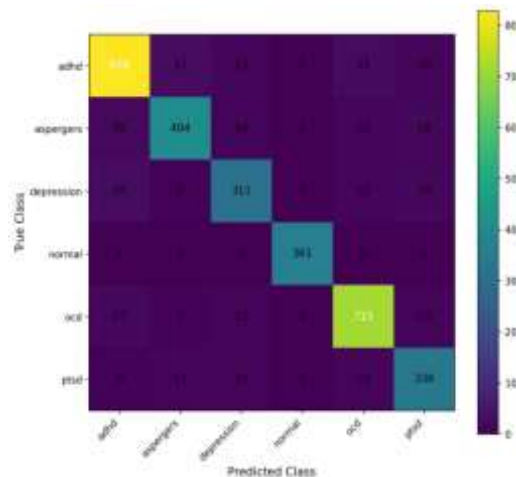
Adding temporal features consistently improved performance: Temporal-LSTM gained 1.17 accuracy and 1.54 macro-F1 points over LSTM, while Temporal-BiGRU gained 0.37 and 0.49 points over BiGRU, indicating that inter-post intervals carry information complementary to MentalBERT semantics, with the magnitude of benefit depending on the recurrent architecture. Bidirectional processing likewise outperformed unidirectional processing in both the temporal and non-temporal settings (e.g., BiGRU 89.93%/89.31% vs. LSTM 88.98%/88.26%), consistent with relevant classification evidence being distributed across multiple positions in a user's history rather than concentrated at the sequence end. The proposed Temporal-BiLSTM, combining both factors, achieved the best results overall—a 1.66-point accuracy and 1.84-point macro-F1 gain over the non-temporal LSTM baseline.

C. Class-Wise Performance

Class	Prec.	Rec.	F1	Supp.
ADHD	0.907	0.928	0.917	904
Asperger's	0.872	0.875	0.873	473
Depression	0.845	0.818	0.831	379
Normal	0.989	0.989	0.989	368
OCD	0.952	0.917	0.935	763
PTSD	0.842	0.879	0.860	381
Macro avg.	0.901	0.901	0.901	3268

TABLE IV. CLASS-WISE PERFORMANCE OF THE PROPOSED MODEL

The Normal class obtained the highest F1-score (0.989), showing that the model reliably separates general Reddit discussion from mental-health-related posting. OCD (0.935) and ADHD (0.917) followed, while Depression (0.831) and PTSD (0.860) were comparatively more difficult, plausibly because depressive and PTSD-related language overlaps with general expressions of sadness, stress, and loneliness that also appear in non-clinical posts.

**Fig. 2. Confusion matrix of the proposed MentalBERT–Temporal-BiLSTM model.**

Discussion

Three findings emerge from the comparative experiments. First, MentalBERT provides a strong domain-specific semantic foundation, capturing psychological terminology and informal expression common in mental-health discourse. Second, normalized inter-post intervals consistently add predictive value beyond semantics alone, though the size of the gain is architecture-dependent. Third, bidirectional sequence modelling with global average pooling outperforms unidirectional modelling, because relevant evidence

can appear at any point in a ten-post history rather than only near the sequence boundary. Together, these results support the central premise that jointly modelling what users write, when they write it, and how that information is distributed chronologically improves user-level classification relative to isolated post-level analysis.

The contribution of temporal information also varied by architecture: it improved LSTM more (+1.17 accuracy, +1.54 macro-F1) than BiGRU (+0.37, +0.49), suggesting that inter-post timing interacts differently with unidirectional versus gated bidirectional recurrence. Nevertheless, the proposed Temporal-BiLSTM achieved the best results overall, indicating that fusing temporal features with bidirectional LSTM processing is particularly well suited to modelling ten-post user histories, and that global average pooling—by aggregating information from every time step rather than only the final hidden state—helps the classifier draw on evidence distributed anywhere across a user's chronological posting history, including for users with shorter, zero-padded sequences.

Limitations include reliance on a single platform (Reddit), community-derived rather than clinically verified labels, and a fixed 10-post history that may under-represent highly active users. Results should therefore be read as computational classification evidence rather than clinical diagnosis.

VII. Conclusion

This paper presented MentalBERT–Temporal–BiLSTM, an end-to-end framework for user-level, multi-class mental health classification from chronological Reddit posts. By fusing MentalBERT contextual embeddings with normalized inter-post time intervals and processing the result with a bidirectional LSTM and global average pooling, the framework achieved 90.64% accuracy and a 90.10% macro F1-score on an independent test set, outperforming non-temporal and unidirectional temporal baselines. The comparative results confirm that both explicit temporal information and bidirectional sequence modelling contribute complementary gains, with class-wise analysis showing strongest performance on Normal, OCD, and ADHD and comparatively greater difficulty for Depression and PTSD. Future work will explore longer posting histories, additional behavioural and interaction features, cross-platform and multilingual evaluation, and explainable-AI techniques to identify the evidence underlying individual predictions, while continuing to treat the framework as a research tool rather than a diagnostic instrument.

References

- [1] World Health Organization, "Mental disorders," Fact sheet, 2026. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>
- [2] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, 2010.
- [3] A. Lee et al., "A comparative study of machine learning and language model approaches for classifying mental health disorders," in *Proc. IEEE ICHI*, 2026.
- [4] T. Zhang, A. M. Schoene, S. Ji, et al., "Natural language processing applied to mental illness detection: a narrative review," *npj Digital Medicine*, vol. 5, p. 46, 2022.
- [5] S. Arora et al., "Enhancing online well-being through transformer-based analysis of misinformation and mental health," *Discover Mental Health*, vol. 6, 2026.
- [6] H. S. Putra et al., "Early detection of mental health on social media using a hybrid Bi-LSTM–XGBoost model," *Sci. Rep.*, vol. 16, p. 20582, 2026.
- [7] I. Ali et al., "Overview of the CLPsych 2026 shared task: capturing mental health changes through social media timeline dynamics," in *Proc. CLPsych Workshop*, 2026, pp. 389–421.
- [8] S. U et al., "Sentiment trajectory modeling in mental health discussions via transformer-based analysis," *Front. Artif. Intell.*, vol. 9, p. 1810552, 2026.
- [9] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: publicly available pretrained language models for mental healthcare," in *Proc. LREC*, 2022, pp. 7184–7190.
- [10] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [11] K. Huang, J. Altsosaar, and R. Ranganath, "ClinicalBERT: modeling clinical notes and predicting hospital readmission," *CHIL Workshop*, 2020, arXiv:1904.05342.
- [12] M. E. Aragón et al., "DisorBERT: a double domain adaptation model for detecting signs of mental disorders in social media."
- [13] M. M. Hossain et al., "Multi task opinion enhanced hybrid BERT model for mental health analysis," *Sci. Rep.*, vol. 15, p. 3332, 2025.
- [14] M. E. Villa-Pérez et al., "A comparative and statistical analysis of depression classification in social media," *Soc. Netw. Anal. Min.*, vol. 16, p. 44, 2026.
- [15] J. Patel and P. Patel, "Deep learning for early detection of depression and anxiety," *Intelligent Hospital*, vol. 2, no. 2, p. 100066, 2026.

- [16] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. ICWSM, vol. 7, no. 1, pp. 128–137, 2013.
- [17] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in Proc. Workshop CLPsych, 2014, pp. 51–60.
- [18] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media," *Curr. Opin. Behav. Sci.*, vol. 18, pp. 43–49, 2017.
- [19] A. Wongkoblaph, M. A. Vadillo, and V. Curcin, "Researching mental health disorders in the era of social media," *J. Med. Internet Res.*, vol. 19, no. 6, e228, 2017.
- [20] solomonk, "Reddit mental health posts dataset," Hugging Face Dataset Repository, 2022.
- [21] Hylum, "AskReddit dataset," Hugging Face Dataset Repository, 2023.