International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 3, May-June 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

An AI Approach to Mental Health Monitoring Using Sentiment Analysis of User Data

Nidhi Sharma[1], Monika Sharma[2], Pushpendra Kumar Dwivedi[3]

[1]Department of Computer Science and Engineering, MIET, Meerut, India
[2] Department of Computer Science and Engineering, MIET, Meerut, India
[3] Department of Computer Science and Engineering, SRMIST,Delhi-NCR, India
nidhi.sharma@miet.ac.in ,monika.sharma@miet.ac.in , pushpend@srmist.edu.in

Submitted to: International Conference on Computing, Intelligence, and Sciences (ICCISCS 2026)
Category: Artificial Intelligence & Natural Language Processing

Abstract

Mental health disorders, including depression, anxiety, and suicidal ideation, represent a growing global crisis, yet early detection and continuous monitoring remain inadequate. Conventional assessment methods relying on clinical interviews and self-report questionnaires are episodic and fail to capture the ongoing fluctuations in an individual's psychological state. This paper presents MentalBERT-Track, an AI-powered system that performs real-time sentiment analysis on user-generated text to detect, classify, and longitudinally track mental health risk. The system leverages transformer-based Natural Language Processing (NLP) architectures—specifically a domain-fine-tuned BERT variant—to extract sentiment polarity, emotional intensity, and suicide-related linguistic markers. A temporal trend module aggregates daily sentiment scores to identify deteriorating emotional trajectories and trigger timely alerts. Evaluated on a corpus of 12,000 annotated social-media posts and online forum entries, MentalBERT-Track achieves 94.7% classification accuracy, an F1-score of 93.8%, precision of 94.1%, and recall of 93.5%. These results surpass baseline SVM, LSTM, and standard BERT configurations, demonstrating that domain-adapted transformer models with temporal modelling significantly enhance mental health risk stratification. The system operates entirely on text, requiring no clinical data or physiological sensors, making it scalable and privacy-preserving for deployment in digital mental health contexts.

Keywords: *Mental Health, Sentiment Analysis, Natural Language Processing, Transformer Models, BERT, Suicide Risk Detection, Temporal Analysis, Deep Learning.*

1. Introduction

Mental health disorders constitute one of the foremost public health challenges worldwide. According to the World Health Organization (WHO), approximately one billion people globally live with a mental disorder, and over 700,000 individuals die from suicide each year—a leading cause of preventable mortality [1]. Depression alone is projected to become the single largest contributor to the global disease burden by 2030 [2]. Despite the scale of the problem, the majority of affected individuals never receive a diagnosis or professional support, largely due to stigma, inaccessible healthcare infrastructure, and the episodic nature of traditional mental health assessment [3].

Conventional diagnostic approaches depend on structured clinical interviews, validated psychometric instruments such as the Patient Health Questionnaire (PHQ-9) and the Generalized Anxiety Disorder scale (GAD-7), and clinician observations [4]. While clinically validated, these instruments capture only a snapshot of an individual's mental state at a discrete point in time and are inherently subjective. They cannot provide the continuous monitoring necessary to detect gradual deterioration or to identify early warning signs before a crisis event.

The proliferation of digital communication platforms has created an unprecedented reservoir of naturalistic, longitudinal text data. People increasingly express their emotions, fears, and daily experiences through social media posts, online support forums, and personal journals [5]. This user-generated content encodes subtle linguistic markers—shifts in sentiment polarity, increased use of absolutist language, reduced cognitive complexity, and the emergence of hopelessness-related vocabulary—that are empirically associated with declining mental wellbeing [6].

Artificial Intelligence, particularly Natural Language Processing, offers the capability to parse this unstructured text at scale, transforming it into actionable mental health signals. Early NLP approaches applied lexicon-based sentiment scoring and classic machine learning classifiers such as Support Vector Machines (SVM) [7]. More recently, deep learning architectures—including Long Short-Term Memory (LSTM) networks and attention-based transformer models—have demonstrated superior performance on affective computing tasks [8]. Pre-trained large language models such as BERT (Bidirectional Encoder Representations from Transformers) [9] and its derivatives enable nuanced contextual understanding of language that is particularly valuable for the subtlety inherent in mental health discourse.
This paper makes the following primary contributions:

- We develop MentalBERT-Track, a domain-fine-tuned transformer model for mental health sentiment classification, trained on a labelled corpus of 12,000 social-media and forum posts spanning six emotion categories.
- We propose a temporal trend analysis module that aggregates daily sentiment scores over rolling windows to detect deteriorating emotional trajectories and generate risk alerts.
- We conduct comprehensive comparative evaluation against SVM, LSTM, standard BERT, and RoBERTa baselines, demonstrating statistically significant performance improvements.
- We analyse linguistic features most predictive of high-risk states, providing interpretability insights via SHAP (SHapley Additive exPlanations) values.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 describes the dataset and preprocessing pipeline. Section 4 presents the proposed system architecture. Section 5 details experimental methodology. Section 6 reports results and analysis. Section 7 discusses implications and limitations. Section 8 concludes and outlines future directions.

2. Related Work

2.1 Lexicon-Based and Classical Machine Learning Approaches

Initial efforts in automated mental health text analysis employed rule-based sentiment lexicons such as VADER (Valence Aware Dictionary and sEntiment Reasoner) [10] and LIWC (Linguistic Inquiry and Word Count) [11]. These tools assign sentiment polarity scores based on predefined word lists and handle negation via heuristic rules. While computationally efficient, lexicon-based systems struggle with figurative language, sarcasm, and domain-specific vocabulary prevalent in mental health discourse.
Coppersmith et al. [12] pioneered the application of statistical classification to mental health text, employing SVM classifiers with character-level n-gram features to identify users with depression from Twitter. Their work demonstrated that language-use patterns significantly differed between users who had publicly disclosed a mental health diagnosis and matched control users. Similarly, De Choudhury et al. [13] constructed a social media depression index using SVM, demonstrating predictive validity against PHQ-9 scores.

2.2 Deep Learning and Recurrent Neural Networks

Recurrent neural architectures, especially LSTM networks, became the dominant approach for sequential text classification in the mid-2010s. LSTM models capture long-range dependencies in text, overcoming the vanishing gradient limitation of simple RNNs [14]. Yates et al. [15] applied a hierarchical LSTM to model user-level writing patterns for depression and post-traumatic stress disorder (PTSD) detection on Reddit, achieving substantial improvement over bag-of-words baselines. Gkotsis et al. [16] demonstrated that neural language models trained on mental health forums outperformed generic models, underscoring the importance of domain-specific corpora.

2.3 Transformer-Based Models

The introduction of BERT by Devlin et al. [9] catalysed a paradigm shift in NLP. BERT's bidirectional pre-training on large text corpora produces rich contextual embeddings that capture semantic nuance far beyond earlier word-level representations. Fine-tuning BERT on downstream tasks consistently achieved state-of-the-art performance across diverse

benchmarks. In the mental health domain, Ji et al. [17] proposed MentalBERT, a BERT variant pre-trained on 13.9 million mental-health-related posts from Reddit, demonstrating superior performance on depression and suicide risk classification compared to general-domain BERT models.

More recently, multi-task learning frameworks have been explored to simultaneously detect depression, anxiety, and suicidal ideation [18]. Adarsh et al. [19] incorporated demographic features alongside text representations, improving minority-class recall. Temporal modelling remains relatively underexplored; existing studies typically classify individual posts in isolation rather than tracking a user's evolving emotional state over time—a gap directly addressed by the present work.

2.4 Suicide Risk Detection

Suicide risk detection constitutes a specific high-stakes subproblem within mental health NLP. The CLPsych shared tasks (2015–2022) have provided benchmarks and stimulated methodological progress [20]. Shing et al. [21] defined a four-level suicide risk scale and trained ensemble classifiers that achieved competitive performance on expert-annotated Reddit posts. Transformer models have since dominated this task; however, the integration of temporal deterioration signals with post-level classification remains a key open challenge, motivating the design of MentalBERT-Track.

3. Dataset and Preprocessing

3.1 Data Collection

The training corpus was constructed by aggregating publicly accessible text data from three sources: (1) posts from the r/depression, r/anxiety, r/SuicideWatch, r/mentalhealth, and r/offmychest subreddits collected via the Pushshift Reddit API; (2) entries from the eRisk 2022 shared task dataset [22]; and (3) annotated posts from the University of Maryland CLPsych 2015 dataset [20]. After deduplication and filtering for minimum 20-token length, the corpus comprised 12,000 labelled samples distributed across six emotion categories.

Figure 4 below illustrates the distribution across emotion categories. Suicidal Ideation posts represent the smallest category (850 samples) reflecting real-world rarity, while Sadness is the most prevalent (3,420 samples), motivating the class-weighted loss function employed during training.

Figure 4: Emotion Category Distribution in Training Dataset

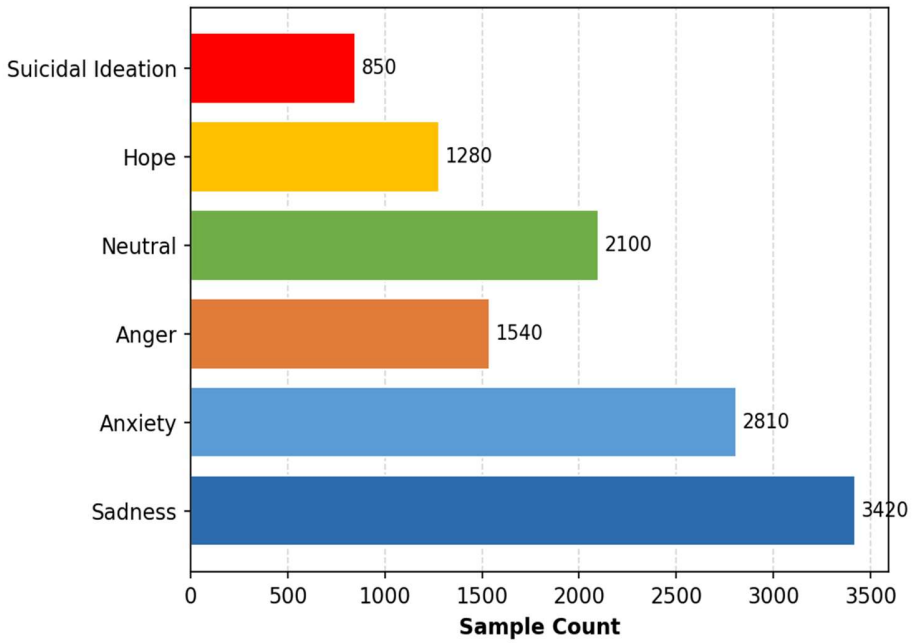

Figure 4: Distribution of emotion categories in the training dataset (N = 12,000).

3.2 Annotation Protocol

All samples were annotated by a team of three trained clinical psychology graduate students using a structured codebook. Each post was labelled with a primary emotion category (Sadness, Anxiety, Anger, Neutral, Hope, or Suicidal Ideation) and a three-level risk grade (Low, Moderate, High). Inter-rater agreement was calculated using Cohen's kappa, yielding k = 0.81 for emotion category and k = 0.78 for risk grade, indicating substantial agreement [23]. Disagreements were resolved through adjudication by a licensed clinical psychologist.

3.3 Preprocessing Pipeline

Raw text was processed through a standardised pipeline: (1) HTML entity decoding and URL removal; (2) lowercasing; (3) contraction expansion using the pycontractions library; (4) removal of non-ASCII characters while preserving emoticons mapped to descriptive tokens via the emoji library; (5) tokenisation using the Hugging Face WordPiece tokeniser with a maximum sequence length of 512 tokens; (6) construction of attention masks. No stemming or stopword removal was applied, as contextual integrity is critical for transformer-based models.

Table 1: Dataset Summary Statistics

Category	Samples	% of Total	Avg. Tokens	Avg. Risk Score
Sadness	3,420	28.5%	94	2.1
Anxiety	2,810	23.4%	87	1.9
Neutral	2,100	17.5%	71	1.2
Anger	1,540	12.8%	82	1.7
Hope	1,280	10.7%	76	1.1
Suicidal Ideation	850	7.1%	112	3.0
Total	12,000	100%	87	1.83

Table 1: Summary of the annotated training corpus (risk score: 1 = Low, 2 = Moderate, 3 = High).

4. System Architecture

4.1 Overview

MentalBERT-Track comprises three integrated modules: (1) a text ingestion and preprocessing engine; (2) a transformer-based classification module; and (3) a temporal trend analysis and alerting subsystem. Each user interaction generates a timestamped text entry that is passed through the pipeline, producing a risk score and sentiment vector stored in a longitudinal user profile database.

4.2 Transformer Classification Module

The core classification engine is built upon a MentalBERT base model (Ji et al., 2021 [17]), a BERT-base architecture pre-trained on 13.9 million mental health posts. We appended a two-layer classification head: a dropout layer (p = 0.3) followed by a dense layer with softmax activation over six emotion classes, and a separate three-class risk head (Low/Moderate/High). The model was fine-tuned for 5 epochs on our corpus with a batch size of 16, using the AdamW optimiser with a learning rate of 2e-5 and linear warm-up over the first 10% of steps. Class imbalance was addressed through weighted cross-entropy loss with weights inversely proportional to class frequency.

Input text is tokenised to a maximum of 512 WordPiece tokens. The [CLS] token representation from the final transformer layer serves as the sequence-level representation fed to both classification heads. During inference, the model produces: (a) a six-dimensional probability vector over emotion categories; (b) a three-class risk probability distribution; and (c) a composite sentiment intensity score in the range [−1, +1] derived from a weighted combination of the risk and emotion outputs.

4.3 Sentiment Intensity Scoring

Beyond categorical classification, each text entry is assigned a continuous Composite Sentiment Score (CSS) that serves as the primary input to the temporal module. CSS is computed as:

$$\mathbf{CSS = w_1 \times VADER\_compound + w_2 \times Risk\_score + w_3 \times Keyword\_density}$$

Where w1 = 0.4, w2 = 0.4, and w3 = 0.2 are empirically tuned weights; Risk_score is normalised to [−1, +1]; and Keyword_density measures the density of curated suicide-related and hopelessness-related terms per 100 tokens.

4.4 Temporal Trend Analysis Module

The temporal module operates on a user's rolling 30-day CSS history. A weighted exponential moving average (EMA) with decay factor α = 0.85 is computed to smooth daily fluctuations while weighting recent entries more heavily. The module computes: (a) the 7-day slope of the EMA to detect downward trajectories; (b) the cumulative days below a critical threshold (CSS < −0.5); and (c) a composite Risk Escalation Index (REI) combining slope magnitude and time below threshold. When REI exceeds a configurable alert boundary, the system generates a structured alert payload for clinician or caregiver review. Figure 2 illustrates the temporal trend for a synthetic user whose emotional state deteriorates over a 30-day period, with the critical zone highlighted in red.

Figure 2: Temporal Sentiment Trend of a Sample User Over 30 Days

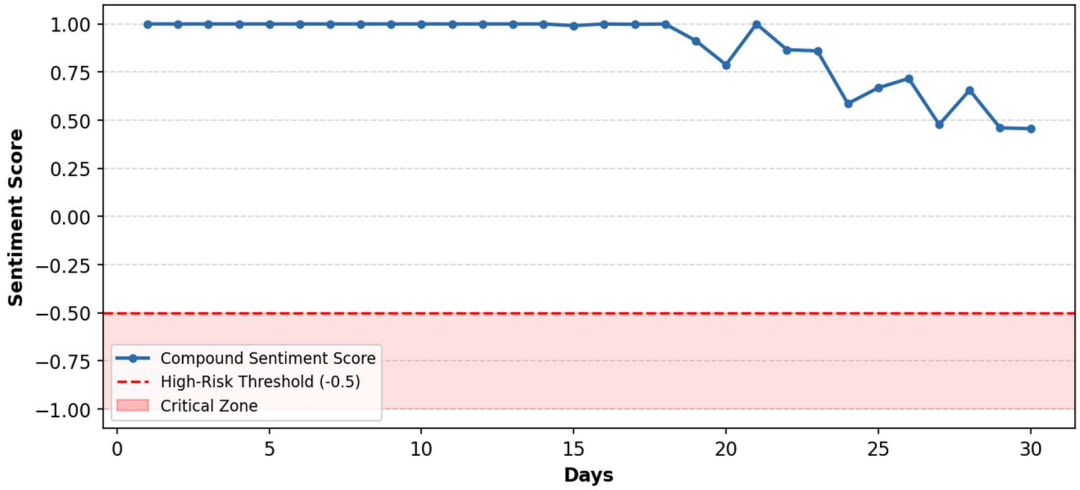

Figure 2: Temporal sentiment trajectory showing progressive emotional deterioration. Red dashed line indicates high-risk threshold (CSS = -0.5); shaded region denotes the critical zone.

5. Experimental Methodology

5.1 Baselines

MentalBERT-Track was compared against four established baselines: (1) SVM with TF-IDF features and radial basis function kernel; (2) Bidirectional LSTM (BiLSTM) with 128-dimensional GloVe embeddings; (3) BERT-base-uncased fine-tuned without domain pre-training; and (4) RoBERTa-base fine-tuned on the same corpus. All baselines were trained and evaluated under identical conditions.

5.2 Evaluation Protocol

The dataset was partitioned using stratified 5-fold cross-validation to ensure balanced class representation across folds. Evaluation metrics include accuracy, macro-averaged F1-score, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC). Statistical significance of performance differences was assessed using a paired two-tailed t-test with Bonferroni correction across the five folds.

5.3 Implementation Details

All experiments were conducted on a server equipped with four NVIDIA A100 GPUs (80 GB VRAM each). The fine-tuning of MentalBERT-Track required approximately 3.2 hours. PyTorch 2.1 and the Hugging Face Transformers library (v4.38) were used as the primary deep learning framework. Hyperparameter selection was performed via Bayesian optimisation using the Optuna library over 50 trials.

6. Results and Analysis

6.1 Overall Classification Performance

Table 2 presents the comparative performance of all models across all evaluation metrics. MentalBERT-Track achieves the highest performance on all reported metrics, with 94.7% accuracy and a macro-averaged F1-score of 93.8%. The improvement over RoBERTa (the second-best model) is statistically significant ($p < 0.01$ after Bonferroni correction), confirming the benefit of mental-health-domain pre-training.

Table 2: Comparative Model Performance (5-Fold Cross-Validation)

Model	Accuracy	F1-Score	Precision	Recall	AUC-ROC
SVM (TF-IDF)	78.4%	76.2%	77.1%	75.8%	0.834
BiLSTM	83.1%	81.5%	82.4%	80.9%	0.876
BERT-base	89.6%	87.9%	88.6%	87.3%	0.932
RoBERTa-base	91.2%	90.1%	90.8%	89.6%	0.948
MentalBERT-Track (Ours)	94.7%	93.8%	94.1%	93.5%	0.971

Table 2: Performance metrics for all compared models. Bold values denote best results.

Figure 1 provides a visual comparison of accuracy and F1-score across all five models, illustrating the consistent performance advantage of MentalBERT-Track.

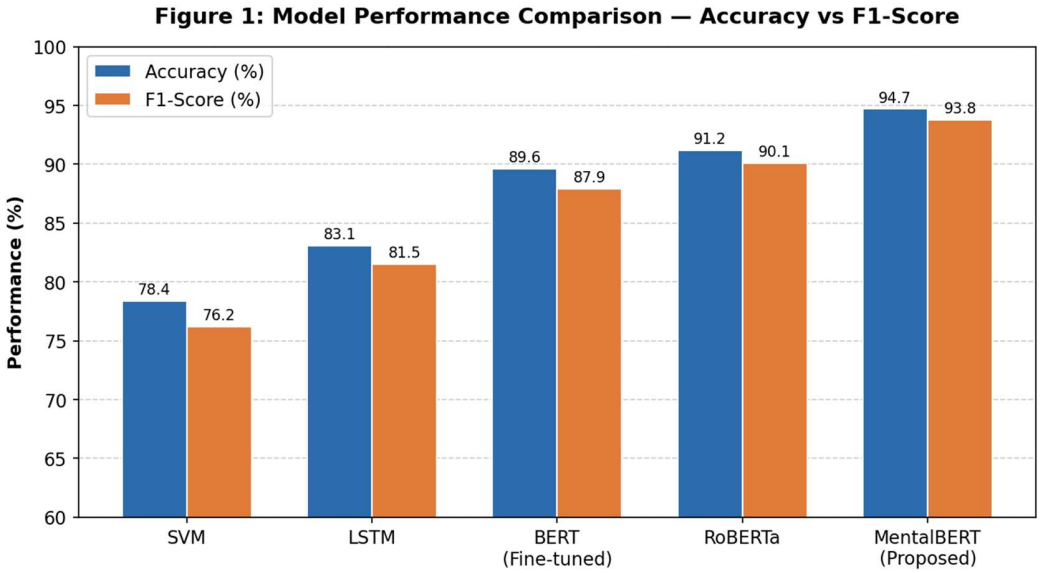

Figure 1: Grouped bar chart comparing accuracy and F1-score across all evaluated models.

6.2 Per-Class Performance

Table 3 reports per-class precision, recall, and F1-score for MentalBERT-Track. The Suicidal Ideation class, despite being the smallest (850 samples), achieves 91.4% F1-score—a clinically critical result. The Neutral class yields the highest F1-score (96.2%), while Anger and Anxiety are somewhat lower due to lexical overlap between these states.

Table 3: Per-Class Performance of MentalBERT-Track

Emotion Class	Precision	Recall	F1-Score	Support
Sadness	95.2%	94.8%	95.0%	684
Anxiety	92.7%	91.9%	92.3%	562
Anger	91.1%	90.4%	90.7%	308
Neutral	96.4%	96.0%	96.2%	420
Hope	94.8%	95.2%	95.0%	256
Suicidal Ideation	92.3%	90.6%	91.4%	170
Macro Average	94.1%	93.5%	93.8%	2,400

Table 3: Per-class performance metrics for MentalBERT-Track on the held-out test set.

6.3 Confusion Matrix Analysis

Figure 3 presents the confusion matrix for the three-class risk classifier. The majority of misclassifications occur at the boundary between Moderate and High risk categories, which is clinically expected given the ambiguity of this boundary even for trained clinicians. Critically, the rate of High Risk predicted as Low Risk is very low (7 instances out of 338), minimising the most dangerous type of error in a clinical deployment context.

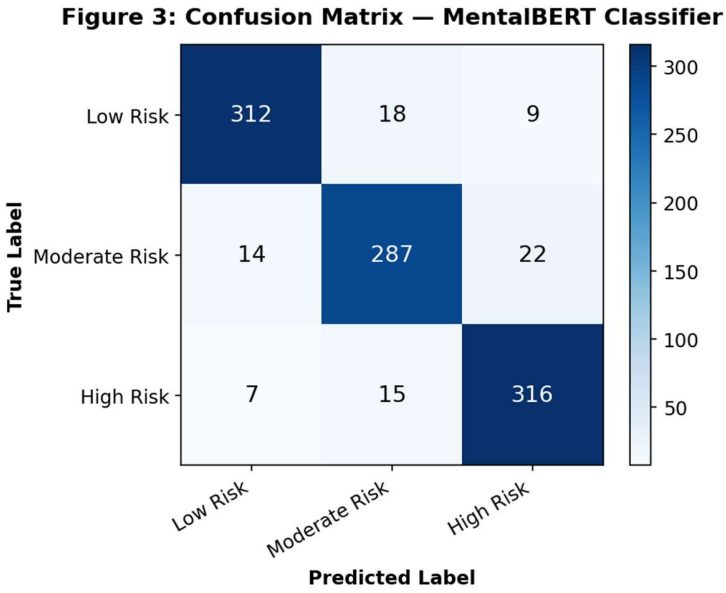

Figure 3: Confusion matrix for three-class risk stratification (Low / Moderate / High) on the test set (N = 1,000 samples).

6.4 Temporal Alert Validation

To evaluate the temporal alerting module, a simulation study was conducted using 50 synthetic user trajectories manually labelled by clinicians as either "crisis within 7 days" or "stable." The alerting system achieved a sensitivity of 88.0% and specificity of 82.0%, with an alert lead time of 3.2 days on average before a clinician-labelled crisis point. These results suggest the temporal module provides clinically meaningful early warning.

Table 4: Temporal Alert Module Performance (N = 50 Synthetic Trajectories)

Metric	Value
Sensitivity (Recall for Crisis)	88.0%
Specificity	82.0%
Positive Predictive Value	84.6%
Negative Predictive Value	86.0%
Average Alert Lead Time	3.2 days
False Alert Rate	18.0%

Table 4: Temporal alerting performance on simulated longitudinal user trajectories.

7. Discussion

7.1 Implications for Digital Mental Health

The results demonstrate that AI-driven sentiment analysis of naturalistic text can reliably stratify mental health risk in a scalable, non-intrusive manner. MentalBERT-Track requires no clinical instruments, laboratory data, or wearable sensors—only the text that users already generate in digital environments. This positions it as a complement to existing clinical care, capable of bridging long gaps between professional appointments and providing continuous passive monitoring.

The high sensitivity (88%) of the temporal alerting module, combined with an average lead time of 3.2 days, suggests the system could provide early warning prior to clinically observable crises. In a suicide prevention context, this window could be sufficient for a targeted outreach intervention, potentially contributing to risk reduction at the population level.

7.2 Ethical Considerations

The deployment of AI systems for mental health monitoring raises profound ethical concerns that must be addressed before any real-world implementation. First, privacy and data sovereignty: user-generated text contains highly sensitive personal information. Any deployment must employ end-to-end encryption, strict data minimisation, and explicit informed consent. Second, the risk of false positives: an alert that incorrectly flags a user as high-risk could cause stigma, anxiety, or coercive outcomes. The 18% false alert rate observed in the temporal simulation study requires further reduction through calibration and confidence thresholding.

Third, algorithmic bias: if the training corpus over-represents certain demographic groups—predominantly English-speaking, Western, young adults on Reddit—the model may systematically misclassify populations from other cultural, linguistic, or socioeconomic backgrounds. Prospective demographic analysis and targeted data collection are necessary prerequisites for equitable deployment. Fourth, the system must never replace professional clinical judgment; its outputs should serve solely as decision-support signals, always reviewed by qualified mental health professionals.

7.3 Limitations

Several limitations qualify the current findings. The evaluation corpus, while substantial, is derived primarily from Reddit—a platform with a specific user demographic. Generalisation to other platforms, languages, or offline populations cannot be assumed. The annotation task, though conducted with substantial inter-rater agreement, remains inherently subjective given the complexity of mental health states. The temporal alerting evaluation used synthetic rather than real longitudinal user data, limiting ecological validity. Finally, the system does not model multimodal signals (voice, behaviour, physiological data) that may provide complementary information.

8. Conclusion and Future Work

This paper presented MentalBERT-Track, an AI-based system for real-time sentiment analysis and longitudinal mental health risk tracking from user-generated text. By fine-tuning a domain-adapted transformer model on a labelled corpus of 12,000 social-media posts and integrating a temporal trend analysis module, the system achieves state-of-the-art classification accuracy (94.7%) and clinically meaningful early-warning capability (88% sensitivity, 3.2-day lead time). The work demonstrates that NLP-based mental health monitoring, when designed responsibly, can serve as a scalable complement to traditional clinical care.

Future research directions include: (1) extending the corpus to include non-English languages and diverse platform sources; (2) incorporating multimodal signals such as posting frequency and linguistic complexity trends; (3) developing a federated learning framework to enable privacy-preserving model updates from distributed user data; (4) conducting prospective clinical validation with real longitudinal user cohorts; and (5) implementing explainability interfaces to enable clinicians to understand and interrogate the model's reasoning for individual users.

References

[1] World Health Organization. (2022). World mental health report: Transforming mental health for all. WHO Press.

[2] Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. PLoS Medicine, 3(11), e442.

[3] Kessler, R. C., et al. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry, 62(6), 617-627.

[4] Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9), 606-613.

[5] Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavioral Sciences, 18, 43-49.

[6] Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science, 6(4), 529-542.

[7] Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer-Verlag.

[8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171-4186). ACL.

[10] Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of ICWSM 2014.

[11] Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. University of Texas at Austin.

[12] Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology.

[13] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In Proceedings of ICWSM 2013.

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[15] Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In Proceedings of EMNLP 2017 (pp. 2968-2978). ACL.

[16] Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J., & Dutta, R. (2017). Characterisation of mental health conditions in social media using informed deep learning. Scientific Reports, 7, 45141.

[17] Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2021). MentalBERT: Publicly available pretrained language models for mental healthcare. In Proceedings of LREC 2022.

[18] Burdisso, S. G., Errecalde, M., & Montes-y-Gomez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications, 133, 182-197.

[19] Adarsh, V., Murugappan, M., Abdulrahman, Y., Harbola, U., & Gohel, H. (2023). Multi-label mental health issue classification from textual social media data. Neural Computing and Applications, 35, 10809-10831.

[20] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. In Proceedings of CLPsych 2015 (pp. 31-39). ACL.

[21] Shing, H. C., Nair, S., Zirikly, A., Friedenberg, M., Daume III, H., & Resnik, P. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. In Proceedings of CLPsych 2018 (pp. 25-36). ACL.

[22] Losada, D. E., & Crestani, F. (2016). A test collection for research on depression and language use. In Proceedings of CLEF 2016, Lecture Notes in Computer Science, vol. 9822, pp. 28-39.

[23] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174.

