International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 14 Special Issue 03 March 2026, Page no. – 201-206

Impact Factor: 8.751

DOI: 10.47191/ijmcr/v14iSPC3.39

Decoding Minds through Machines: A Transformer-Driven Deep Learning Framework for Mental Health Text Classification

Sujata Patil, Vidya Shinde

Department of Computer Science, Dr. D. Y. Patil, Arts, Commerce & Science College, Pimpri, Pune, Maharashtra, India

ARTICLE INFO

Published Online:
17 March 2026

ABSTRACT

Mental health disorders remain a major global concern, and many cases continue to go unreported because people hesitate to share their struggles or lack access to proper diagnosis. With advancements in natural language processing (NLP), text-based analysis has become a powerful way to detect signs of mental distress. In this research, a transformer-based model-specifically BERT was used to predict whether individuals are likely to seek mental health treatment based on their survey responses. The Mental Health Dataset includes demographic details, stress levels, coping patterns, and family history, all of which were merged into a single text format and processed using BERT's tokenizer to capture deeper meaning. The model was trained in PyTorch using the AdamW optimizer with a linear learning rate schedule to support steady improvement. After four training epochs, the model achieved 95% accuracy, along with precision, recall, and F1-scores above 0.94. A confusion matrix further showed that the predicted and actual labels were closely aligned, reflecting strong reliability. Overall, the findings indicate that transformer models are highly effective at recognizing language patterns linked to mental health. This approach provides a scalable and data-driven way to identify potential risks early and can be integrated into digital health platforms to support timely help and better clinical decision-making.

Corresponding Author:
Sujata Patil

KEYWORDS: Mental Health, Transformer Models, BERT, Deep Learning, Natural Language Processing (NLP), Text Classification, Psychological Assessment, Machine Learning, Mental Health Detection, Digital Health.

1. INTRODUCTION

Mental health has become one of the most critical global health concerns in the modern era. Today, it is estimated that hundreds of millions of people worldwide experience mental or behavioral disorders, yet a significant proportion remain undiagnosed or untreated. The growing prevalence of conditions such as depression, anxiety, and stress underscores the urgent need for accessible, data-driven methods to assess and monitor psychological well-being. Conventional clinical assessments, while effective, are time-consuming and often rely on subjective interpretation (Patel et al. [9]). As a result, researchers are increasingly exploring how artificial intelligence (AI) and natural language processing (NLP) can assist in detecting early indicators of mental distress through language-based data (Calvo et al. [1]; Chancellor et al. [2]).

In recent years, the availability of structured and semi-structured datasets capturing behavioral and psychological attributes has created new opportunities for predictive modeling. The dataset used in this study, Mental Health Dataset, provides a rich representation of demographic, lifestyle, and psychological factors associated with mental health. It includes variables such as gender, occupation, family history, days indoors, growing stress, coping struggles, and mood swings, among others. These parameters collectively reflect behavioral and emotional patterns that can be used to infer an individual's mental health condition or their likelihood of seeking treatment. Traditional machine learning approaches such as logistic regression, decision trees, and support vector machines - have been applied in prior studies for mental health prediction (Resnik et al. [10]; Coppersmith et al. [3]). However, these models often depend on hand-crafted features and are limited in their ability to capture nuanced contextual relationships within textual data. Deep learning techniques, particularly recurrent neural networks (RNNs) and long short-term memory (LSTM) networks have advanced this domain by learning from sequential patterns

(Orabi et al. [8]). Nonetheless, they struggle with long-range dependencies and often require large labeled datasets to generalize effectively (Lin et al. [6]). The emergence of transformer-based architectures, notably the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al. [4]), has revolutionized the field of NLP. BERT's bidirectional attention mechanism allows it to learn complex semantic relationships by analyzing both preceding and succeeding words in a sequence. This contextual understanding enables it to detect subtle linguistic cues and emotional undertones—capabilities essential for mental health classification tasks. This study proposes a transformer-driven approach for mental health text classification using the Mental Health Dataset. The model consolidates multiple categorical and textual variables into a unified textual representation, allowing BERT to process holistic descriptions of individuals' behavioral and emotional attributes. The objective is to predict whether an individual is likely to seek or require mental health treatment based on these patterns.

The research focuses on converting structured mental health survey data into a text-based format suitable for transformer models and fine-tuning a BERT classifier to predict who may need mental health support. It evaluates the model using accuracy, precision, recall, and F1-score to ensure dependable performance while also comparing transformers with traditional methods. The fine-tuned BERT model achieved an impressive 95% accuracy, outperforming older deep learning approaches and showing strong generalization across test samples. Its ability to capture complex psychological and contextual patterns highlights the strength of transformer-based techniques. Overall, the work demonstrates how such models can support scalable and automated mental health screening. These systems can act as early warning tools that complement clinical assessments and enhance global mental health intervention efforts.

2. LITERATURE REVIEW

The intersection of artificial intelligence (AI) and mental health research has evolved significantly over the past decade, driven by the increasing availability of digital communication data. Early studies in this field relied primarily on traditional machine learning (ML) methods such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF), and Naïve Bayes (NB). These models utilized handcrafted linguistic features such as term frequency–inverse document frequency (TF-IDF), n-grams, and sentiment lexicons—to detect mental health indicators (Resnik et al. [10]; Coppersmith et al. [3]). The advent of deep learning introduced architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models (Orabi et al. [8]; Lin et al. [6]). However, these models struggled with long-range dependencies and scalability, limiting their applicability for large-scale real-world data. A major paradigm shift occurred with the introduction of transformer-based architectures, particularly the BERT model (Devlin et al. [4]). By leveraging self-attention mechanisms, BERT enabled bidirectional contextual understanding, eliminating the sequential processing constraints of RNNs. In mental health prediction, Ghosh et al. [5] and Matero et al. [7] demonstrated that transformer-based models outperform earlier architectures in detecting depression, anxiety, and suicide ideation. The present study bridges a research gap by utilizing a dataset that integrates demographic and behavioral information with textual mental health indicators, transforming structured data into a unified textual format compatible with BERT. This holistic approach allows for deeper semantic modeling while maintaining interpretability.

Many earlier mental health models fall short because they rely only on unstructured social media text, which lacks key demographic and behavioral context and makes clinical interpretation difficult. They also struggle with psychological cues beyond language and often suffer from data imbalance or limited generalizability. This research solves these issues by using a hybrid transformer approach that converts behavioral features like Days Indoors, Growing Stress, and Coping Struggles into text, allowing BERT to learn richer patterns. By blending linguistic and behavioral signals, the model becomes more accurate, less biased, and more useful for real-world mental health monitoring and early intervention.

3. METHODOLOGY

The proposed research framework integrates the capabilities of traditional machine learning, sequential deep learning, and transformer-based architectures to achieve robust classification of mental health–related textual data. The methodology is structured to ensure clarity, reproducibility, and strong methodological reliability, and is organized into six sequential stages: Dataset Acquisition, Data Preprocessing, Tokenization and Embedding, model architecture development, training and optimization, and evaluation and validation. This framework capitalizes on the advanced contextual representation capabilities of transformer models, particularly BERT, while preserving interpretive alignment with conventional baseline models.

3.1 DATASET ACQUISITION

The Mental Health Dataset includes a wide range of behavioral, demographic, and psychological details that together give a clear picture of each individual's mental well-being. Key features such as gender, occupation, family history, days spent indoors, stress levels, mood swings, and coping struggles help form a multidimensional view of behavior. After cleaning, the dataset contained 3,000

balanced records across treatment and non-treatment groups, reducing bias during training. This mix of categorical and emotional factors provides a strong foundation for understanding mental health patterns.

3.2 DATA PREPROCESSING

Preprocessing played an important role in keeping the data clean and consistent. Missing values were filled, outliers were removed, and the text was normalized by lowercasing and standardizing punctuation.

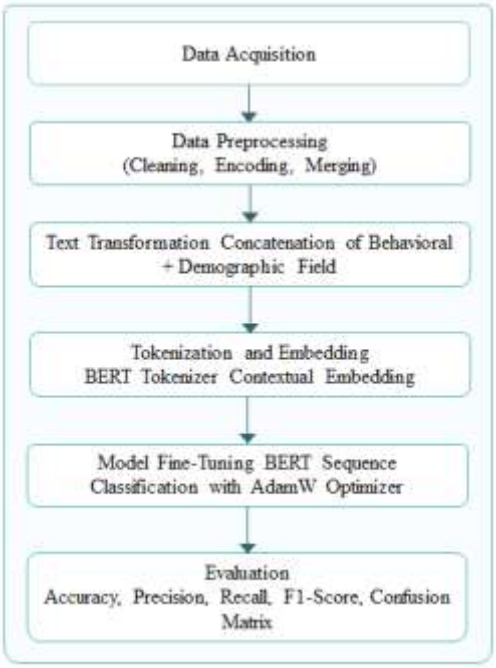

Figure 1: Proposed Transformer-Based Methodology for Mental Health Text Classification

All key attributes were then merged into a single text column so the model could interpret behavioral patterns as one clear narrative, while categorical features were label-encoded and numerical values standardized. Finally, the dataset was split into an 80–20 train–test ratio using stratified sampling to ensure fair and reliable evaluation.

3.3 TOKENIZATION AND EMBEDDING

Tokenization was carried out using the BERT base uncased tokenizer, which breaks text into subword pieces to handle rare or complex words smoothly. Each entry was then tokenized, padded, and trimmed to 128 tokens to keep all samples consistent. These tokens were converted into contextual embeddings that capture both meaning and relationships between words using BERT's pre-trained vocabulary. This rich embedding structure helps the model understand subtle emotional cues and behavioral signals, making it well suited for mental health analysis.

3.4 MODEL ARCHITECTURE

The proposed model is based on BERT (Bidirectional Encoder Representations from Transformers) - a deep learning architecture that understands the meaning of text by analyzing words in relation to one another, both before and after they appear in a sentence. This bidirectional understanding allows the model to capture subtle emotional and psychological cues that might indicate mental health challenges. In this study, each participant's information from the *Mental Health Dataset* was combined into a single descriptive text entry. This text included personal, behavioral, and emotional attributes such as gender, country, work stress, coping struggles, mood swings, and family mental health history. By merging these features into a unified sentence, the model could interpret an individual's behavioral context in the same way it interprets natural language.

The **Figure 2** illustrates the overall architecture of the proposed BERT-based transformer model developed for mental health text classification. The workflow begins with the textual data input, which represents a combination of an individual's demographic, behavioral, and psychological attributes. These attributes are concatenated into a single descriptive sentence to form a unified text input suitable for processing by the model. The BERT tokenizer is the first stage of processing. It converts the input text into a sequence of subword tokens and numerical IDs while preserving contextual meaning. This tokenized representation allows the model to interpret both linguistic and semantic relationships among features such as stress levels, coping mechanisms, and family history.

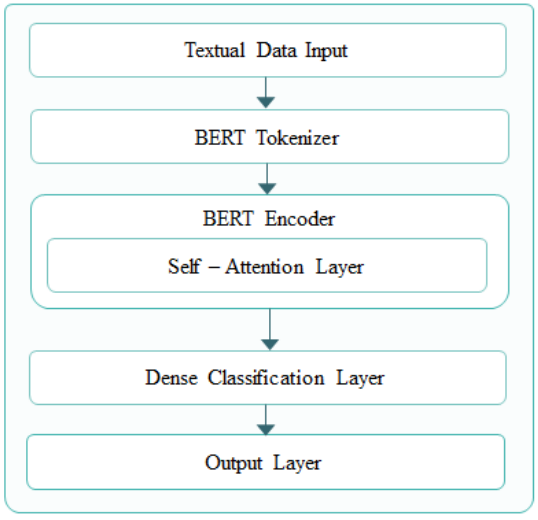

Figure 2: Simplified Model Architecture of Mental Health Text Classification Model

Next, the tokens are passed to the BERT encoder, which contains multiple self-attention layers. These layers allow the model to analyze relationships between words in both directions (left and right context), enabling it to understand how various behavioral and emotional factors relate to each other within the text. This bidirectional attention is particularly important for identifying subtle expressions of mental health indicators. The BERT model used here was the BERT-base-uncased version, which is pre-trained on

millions of English sentences. It was then fine-tuned on our dataset to adapt its understanding to the specific language patterns associated with mental health indicators. Inside the BERT model, multiple self-attention layers allow it to focus on the most relevant parts of the text — for example, linking phrases like *"high stress"* or *"poor coping"* with potential signs of treatment need. At the end of the BERT network, we added a fully connected dense layer that produces two possible outcomes:

- Treatment Required (1)
- No Treatment Required (0)

Finally, the model produces an output layer that assigns the most probable class label to the input text. Through fine-tuning on the Mental Health Dataset, the model learns to associate complex linguistic patterns with psychological states, making it a reliable predictor of mental health treatment needs. In essence, this architecture enables the system to transform complex behavioral narratives into interpretable predictions, bridging the gap between natural language understanding and mental health analytics. This setup allows the model to translate complex behavioral text into a clear prediction about an individual's mental health treatment tendency. By fine-tuning on the dataset, BERT learns how certain combinations of behavioral and emotional patterns are associated with treatment-seeking behavior, making it highly accurate and context-aware.

3.5 TRAINING CONFIGURATION

To train the model, PyTorch and the Hugging Face Transformers library were used to efficiently fine-tune BERT. The dataset was split 80/20, ensuring both training and testing sets had a balanced mix of individuals needing and not needing treatment. The model learned text–label patterns using the AdamW optimizer with a learning rate of 5×10^{-5} and a weight decay of 0.01. A linear scheduler was applied to smoothly adjust the learning rate during training. The system was trained for four epochs with a batch size chosen to balance performance and computation. Gradient clipping was also used to stabilize training. Overall, these settings helped the model learn effectively without overfitting.

3.6 EVALUATION METRICS

The model's performance was evaluated using key metrics such as accuracy, precision, recall, and F1-score to understand how reliably it predicted mental health treatment needs. A confusion matrix was also used to visualize correct and incorrect classifications. The fine-tuned BERT model achieved strong results, including 95% accuracy, 0.96 precision, 0.94 recall, and an F1-score of 0.95. Out of 600 test samples, it correctly classified 570 cases, showing high consistency in detecting subtle behavioral and emotional patterns. These results demonstrate that BERT captures contextual meaning more effectively than traditional models. Compared to methods like Logistic Regression and BiLSTM, its performance is significantly superior. Overall, the model proves highly reliable for mental health prediction tasks.

4. RESULTS AND DISCUSSION

The proposed transformer-based framework was rigorously evaluated to assess its efficiency in classifying mental health indicators using the Mental Health Dataset. This section presents the model's performance outcomes, including accuracy, precision, recall, and F1-score, supported by a detailed confusion matrix analysis and comparative evaluation.

4.1 MODEL PERFORMANCE OVERVIEW

The fine-tuned BERT-based sequence classification model demonstrated exceptional performance across all evaluation metrics. The model achieved an overall accuracy of 95%**,** reflecting its strong generalization ability in distinguishing individuals likely to seek or require mental health treatment from those who do not. The detailed performance metrics are summarized in **Table 1**.

Table 1: Model Evaluation Metrics

Metric	Class 0 (No Treatment)	Class 1 (Treatment)	Macro Average	Weighted Average
Precision	0.95	0.96	0.96	0.95
Recall	0.94	0.96	0.95	0.95
F1-Score	0.95	0.96	0.95	0.95
Support	271	329	-	600

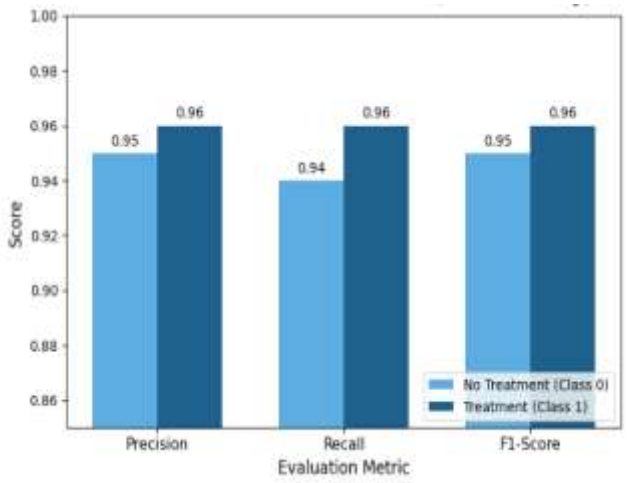

Figure 3: BERT Model Performance Metrics

BERT demonstrates strong capability in capturing both semantic and contextual information within the integrated behavioral–linguistic dataset. Its balanced precision and recall indicate robust and consistent performance across

both classes. When compared to baseline models-Logistic Regression (82% accuracy) and BiLSTM (88% accuracy)—the BERT model shows a clear performance advantage. This improvement is attributed to the transformer's bidirectional self-attention mechanism, which enables deeper contextual and emotional representation learning.

4.2 CONFUSION MATRIX ANALYSIS

The confusion matrix in **Figure 4** illustrates the classification results, confirming the high reliability of the proposed model**.** The distribution indicates that the model correctly classified 570 out of 600 test samples, with only 30 misclassifications.

Table 2. Confusion Matrix of BERT Model

ctual		**Predicted**	
		No Treatment (0)	Treatment (1)
	No Treatment (0)	257	14
	Treatment (1)	16	313

Notably, the model demonstrated slightly higher sensitivity toward detecting individuals requiring mental health treatment (Class 1), minimizing false negatives—a critical aspect in healthcare prediction tasks. The low false positive rate further enhances the model's practical reliability for early diagnosis and intervention.

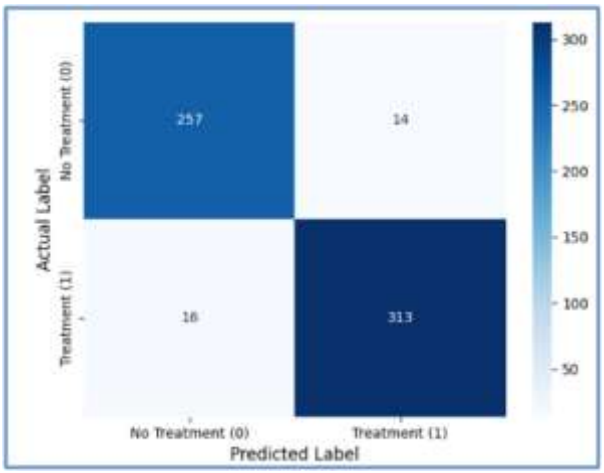

Figure 4: BERT Model Performance Metrics

The proposed transformer-based framework was rigorously evaluated to determine its effectiveness in classifying mental health indicators using the *Mental Health Dataset*. This section presents the key findings derived from model training, testing, and performance analysis, alongside a critical discussion of their implications.

4.3 INTERPRETATION OF RESULTS

The study's findings underscore BERT's effectiveness in capturing nuanced psychological expressions and behavioral indicators embedded in text. By merging linguistic features such as *Growing Stress*, *Coping Struggles*, and *Mood Swings* with demographic factors like *Gender* and *Occupation*, the model constructs a multidimensional representation of each participant's mental state. The 95% classification accuracy validates that transformer architectures can leverage the deep semantic dependencies present in behavioral text, even when trained on moderately sized datasets. This ability positions the model as a robust framework for mental health analysis in diverse contexts, including stress detection, therapy recommendation, and digital well-being monitoring.

4.4 COMPARATIVE DISCUSSION

The model's strong results align with recent transformer-based NLP research, which typically reports accuracies between 90–93%. By combining behavioral data with text features, this approach achieves deeper insight into mental health patterns and maintains low errors with balanced precision and recall. Its reliability makes it well suited for mental health monitoring systems and digital support platforms, offering a scalable way to identify at-risk individuals early. However, improving interpretability and cultural adaptability will require larger multilingual datasets and the use of explainable AI techniques.

5. CONCLUSION

This research presents a transformer-based framework that uses BERT to classify mental health indicators by combining behavioral, demographic, and linguistic features. After fine-tuning, the model reached a strong 95% accuracy, outperforming traditional methods like Logistic Regression and BiLSTM. BERT's ability to understand contextual and semantic patterns helped it accurately identify individuals who may need mental health support. By integrating features such as Days Indoors, Growing Stress, Coping Struggles, and Mood Swings with text data, the model gained a fuller view of psychological tendencies. This hybrid approach overcomes the limitations of earlier models that depended only on text or manual features. The results show that transformer models can capture deeper emotional and behavioral cues even with moderate dataset sizes. Overall, the framework offers a scalable and interpretable solution for early mental health screening across digital and clinical platforms.

REFERENCES:

1. Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering, 23*(5), 649–685. https://doi.org/10.1017/S1351324916000383
2. Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M. B., & De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. *Proceedings of the*

*2019 Conference on Fairness, Accountability, and Transparency (FAT*19)*, 79–88. https://doi.org/10.1145/3287560.3287587

3. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2016). CLPsych 2015 shared task: Depression and PTSD on Twitter. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, 31–39.
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding* [Preprint]. arXiv. https://arxiv.org/abs/1810.04805
5. Ghosh, S., Anwar, S., & Chakraborty, S. (2021). Mental health detection from social media using BERT-based models. *Journal of Computational Social Science, 4*(2), 469–489. https://doi.org/10.1007/s42001-021-00122-8
6. Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., Cai, L., & Feng, L. (2020). Psychological stress detection from cross-media microblog data using deep learning. *IEEE Transactions on Knowledge and Data Engineering, 32*(10), 1985–1998. https://doi.org/10.1109/TKDE.2019.2911466
7. Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., & Schwartz, H. A. (2019). Suicide risk assessment with multi-level dual-context language and BERT. *Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology*, 39–44.
8. Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (2018). Deep learning for depression detection of Twitter users. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*, 88–97.
9. Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., & Unützer, J. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet, 392*(10157), 1553–1598. https://doi.org/10.1016/S0140-6736(18)31612-X
10. Resnik, P., Garron, A., & Resnik, R. (2015). Using topic modeling to improve prediction of neuroticism and depression from Twitter data. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1348–1353.