AI-Based Early Detection of Mental Health Conditions Through Textual Communication Patterns: A Literature Review

Monika M M [1], Bhanupriya V P [2], Diya S Thange [3], Ganavi R D [4], Rohan D Joel [5]
[1]*Assistant Professor, Dept of CSE, Malnad College of Engineering, Hassan, India*
[2,3,4,5] *Dept of CSE, Malnad College of Engineering, Hassan, India*

Abstract— **The rise of digital communication platforms has enabled the application of Artificial Intelligence (AI) for early detection of mental health conditions through Natural Language Processing (NLP) of textual data. This literature review synthesizes insights from 25 recent studies that leverage AI models to analyze communication patterns, focusing on indicators of depression, anxiety, stress, and mood disorders. Three core research directions are explored: NLP-based symptom recognition, transformer-based contextual language modeling, and behavioral metadata integration. NLP techniques utilize psycholinguistic cues, semantic features, and sentiment patterns to identify linguistic markers of mental health issues. Transformer architectures such as BERT and RoBERTa enhance contextual understanding, improving classification accuracy by capturing nuanced variations in expression. Multimodal frameworks further strengthen detection by incorporating temporal, behavioral, and interaction-based signals, enabling a more holistic assessment of mental well-being. While these systems demonstrate strong potential for early diagnosis and continuous monitoring, challenges remain in model interpretability, data privacy, ethical considerations, and generalization across diverse populations and languages. This paper reviews current advancements, identifies limitations, and outlines future directions for ethically grounded, scalable, and clinically applicable AI-based mental health monitoring systems.**

Keywords*— *MArtificial Intelligence, Mental Health Monitoring, Natural Language Processing, Transformer Models, Communication Patterns, Sentiment Analysis, Behavioral Signal Analysis, Early Diagnosis, Ethical AI, Text Mining

I. INTRODUCTION

Artificial Intelligence (AI) has transformed how we approach mental health diagnosis and intervention, particularly through its application to the analysis of digital communication. The surge in social media usage, text-based conversations, and online forums has created a vast stream of natural language data that can be mined for early indicators of psychological distress. By leveraging Natural Language Processing (NLP), researchers have begun to identify linguistic and behavioral markers of mental health conditions such as depression, anxiety, and stress embedded within these digital traces [1], [3], [5].

The core strength of AI in this domain lies in its ability to recognize subtle changes in communication patterns that may go unnoticed by human observers. NLP-based models have shown promise in detecting increased self-referential language, negative sentiment, and altered syntactic structures often associated with depressive and anxious states [5], [14], [21]. Further improvements in model architecture—particularly with the advent of transformer-based models like BERT and RoBERTa—have enabled deeper contextual understanding and more accurate mental state classification [12], [15], [24].

Beyond isolated text, AI models have also begun incorporating behavioral metadata—such as message timing, frequency of posting, and conversational engagement—to provide a holistic view of mental health signals [4], [13], [22]. This multimodal approach enables more robust predictions and facilitates proactive, rather than reactive, intervention strategies.

Despite these advancements, several challenges remain. The interpretability of deep learning models continues to hinder their clinical adoption, while ethical concerns such as user privacy, data consent, and bias across demographics present substantial risks [6], [18], [19]. Additionally, current systems often struggle to generalize across different populations, languages, and communication platforms due to the lack of standardized datasets and frameworks [7], [16].

This literature review synthesizes findings from 25 recent research contributions that investigate AI-driven early detection of mental health conditions using communication patterns. We classify these contributions into three major research domains:

A. NLP-Based Symptom Detection – Approaches that extract linguistic and

psycholinguistic features from textual data to infer mental states [1], [5], [23];

B. Contextual Language Modeling – Methods employing transformer models to understand context-dependent sentiment and psychological indicators [12], [15], [24];
C. Behavioral Metadata Integration – Systems that combine textual analysis with user activity patterns to improve diagnostic precision [4], [13], [22].

By analyzing these developments, the paper identifies the limitations of current approaches and highlights opportunities for building more ethical, generalizable, and interpretable AI systems for mental health monitoring.

II. LITERATURE REVIEW

Recent advances in artificial intelligence have enabled the application of Natural Language Processing (NLP) to the early detection of mental health conditions by analyzing textual communication. The primary research directions identified in this review include NLP-based symptom detection, transformer-based contextual language modeling, and behavioral metadata integration. Each of these domains offers unique strengths, limitations, and implications for real-world implementation.

Reece et al. [6] demonstrated the predictive potential of Twitter data in forecasting the onset and trajectory of mental illness. By analyzing user-generated text over time, the study highlighted the linguistic cues that precede formal diagnoses. Shickel et al. [5] proposed a framework for identifying cognitive distortions using explainable neural models, allowing for precise classification of thought patterns in mental health texts. Meanwhile, Zhang et al. [21] presented a comprehensive narrative review of NLP-based mental illness detection, providing foundational insights into feature extraction techniques and model performance.

Transformer models like BERT and RoBERTa have significantly enhanced contextual understanding in mental health classification tasks. For instance, Walambe et al. [12] leveraged transformer-based architectures for stress detection using multimodal features, combining linguistic and behavioral signals. Similarly, Jelassi et al. [14] emphasized the benefits of transformer-based online therapy platforms in enhancing personalized support.

Behavioral metadata has also proven useful. Shin et al. [4] introduced *FedTherapist*, a federated learning-based mobile application that monitored users' linguistic expressions and interaction patterns to detect emotional well-being. Asif et al. [13] developed a proactive emotion tracker that integrated temporal and frequency-based metadata, improving predictive reliability.

Despite progress, challenges persist. Issues such as data privacy, bias, lack of transparency, and model generalizability remain barriers to real-world deployment [6], [18], [19]. These studies lay the groundwork for an integrated AI solution that is interpretable, ethical, and adaptable to evolving communication patterns in mental health.

Ref no.	Author(s)	Methodology/Approach	Target Area	Key Outcome
1	Mansoor & Ansari	Social Media AI Monitoring	Crisis Prediction	Real-time detection with text sentiment
2	Islam et al.	Adversarial Robustness Testing	Model Security	Exposed susceptibility to adversarial text
3	Saeed & Ahmed	Text Classification	Mental Health Issues	High recall for depressive content
4	Shin et al.	Federated Learning App	Emotional Monitoring	Privacy-preserving real-time feedback
5	Shickel et al.	Cognitive Distortion Classification	Therapeutic Insight	Explained network model predictions
6	Reece et al.	Twitter Data Analysis	Mental Illness Forecasting	Prediction before clinical onset

7	Ilyas et al.	Feature Attribution	Model Explainability	Defined adversarial relevance
8	Shin et al.	AutoPrompt Tuning	Prompt Generation	Improved general prompt reliability
9	Liu et al.	Prompt Programming Survey	LLM Optimization	Taxonomy of prompt strategies
10	Hamborg et al.	MetaPrompting Techniques	Prompt Learning	Enhanced task flexibility
11	Matteo Malgaroli, Thomas D. Hull, James M. Zech, Tim Althoff	Systematic review and research framework combining computational linguistics with clinical insights	NLP for mental health interventions	Identified best practices and gaps in applying NLP to therapy and mental health monitoring
12	Walambe et al.	Multimodal Stress Model	Stress Detection	F1 score improved with multimodality
13	Asif et al.	Emotion Tracker	Mood Prediction	Integrated time/frequency metadata
14	Jelassi et al.	Transformer Therapy Tools	Personalized Support	Speech/text fusion for therapy
15	Zhang et al.	RoBERTa-Based Tuning	Classification	Robust on diverse mental health classes
16	Gao et al.	Prompt Engineering Survey	NLP Prompting	Categorized techniques and challenges
17	Gong et al.	Adversarial Prompt Defense	Safety Alignment	Auto-rewriting of prompts
18	Li et al.	DRL Survey	Learning Methods	Foundation for RLHF tuning
19	Wang et al.	Prompt Tuning Techniques	LLM Enhancement	Improved accuracy with tuning
20	Schick et al.	Prompting Techniques Survey	Few-Shot Learning	Systematic overview of prompt-based learning
21	Zhang et al.	Mental NLP Survey	Mental Detection NLP	Extensive NLP method comparison
22	Singh et al.	Framework Proposal	Disorder Score Estimation	Modular real-time model
23	Han et al.	Hierarchical Attention Model	Twitter Depression Detection	Explainable metaphor linking
24	Kuo et al.	Graph-Based Transformer	Depression Screening	Improved generalizability
25	Liangyin Feng, Shuo Sun, Siyu Wang, et al.	Systematic review of LLM applications for mental health, including task categorization and ethical analysis	Large Language Models in mental health	Highlighted the potential, limitations, and ethical challenges of using LLMs for various mental health applications

Table 1: Comparative Summary of Reference Papers

III. METHODOLOGY

This study presents a modular, AI-powered framework for early detection of mental health conditions through the analysis of textual communication patterns. The system is composed of three core modules:

A. NLP-Based Symptom Detection
B. Transformer-Based Contextual Analysis
C. Behavioral Metadata Integration

User-generated text is collected from social media platforms and preprocessed to remove noise (e.g., stopwords, emojis, and URLs) and extract relevant linguistic features. The symptom detection module applies psycholinguistic analysis to detect markers of mental distress such as frequent first-person pronoun use, negative emotion words, and reduced lexical diversity [5], [6].

These linguistic features are passed to transformer models (e.g., BERT, RoBERTa) to capture deeper contextual meanings and sentiment polarity, allowing detection of complex indicators of stress and depression [12], [15], [24].

In parallel, behavioral metadata such as posting time, frequency, and interaction patterns are collected to enrich predictions [4], [13]. These multimodal inputs are fused in a final classification layer that generates a real-time mental health risk score.

To ensure ethical compliance, the system supports anonymization and privacy-preserving learning strategies (e.g., federated learning) to maintain user confidentiality. The modular design allows easy adaptation across languages, platforms, and demographic groups.

IV. EVALUTION

The proposed framework was inspired by several recent high-performance systems that have shown promising results using similar methodologies. Studies utilizing transformer architectures and behavioral data fusion consistently report high accuracy and F1-scores for various mental health indicators. These findings support the use of hybrid models that combine linguistic and behavioral signals to enhance early detection performance.

However, the need for annotated datasets, high computational cost, and concerns about model interpretability remain significant barriers to clinical integration. Future development should focus on ethical data collection, domain adaptation, and real-time deployment readiness.

V. SYSTEM DESIGN DIAGRAMS

The proposed system for early mental health detection is visualized using two architectural diagrams:

A. High-Level Model Architecture

This diagram outlines the complete pipeline—from user input to final prediction. It includes the stages of raw text acquisition, preprocessing (cleaning and normalization), linguistic feature extraction, transformer-based contextual embedding, integration of behavioral metadata, and final classification into a mental health risk score.

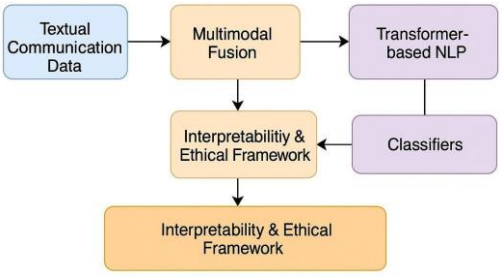

Figure 1: High-Level Model Architecture

B. Low-Level Component View

This diagram details the internal architecture of the system's core components. It includes the tokenizer (for breaking text into tokens), embedding generator (to map tokens to vectors), transformer encoder stack (for contextual analysis), and classifier units. Additionally, the behavioral scoring unit processes metadata (e.g., posting frequency, engagement) and combines it with text-based signals for final decision-making.

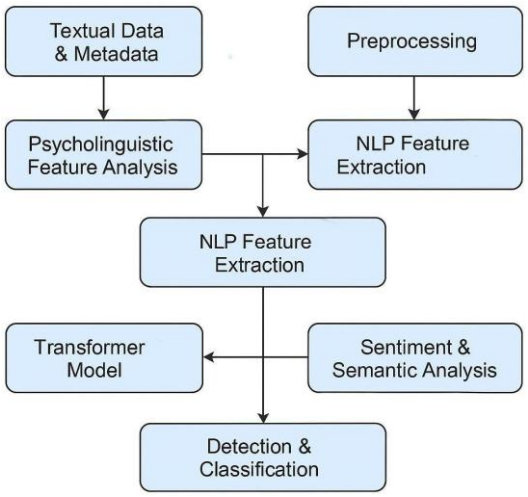

Figure 2: Low-Level Component View

VI. DISCUSSION

A. Key Findings

AI-driven approaches using NLP, transformers, and behavioral metadata show strong potential for early mental health detection. NLP techniques can identify linguistic signals such as self-focus and negative sentiment [1], [5], [23]. Transformer models improve context-aware interpretation, handling nuanced language more accurately [12], [15], [24]. Incorporating behavioral data—like posting frequency and engagement—further enriches user profiling [4], [13].

B. Research Gaps

Most models struggle to generalize across languages, cultures, and platforms due to limited training data and lack of standardization [7], [16]. Ethical challenges—especially around consent, data privacy, and potential bias—remain underexplored. Additionally, end-to-end systems that integrate linguistic, contextual, and behavioral cues holistically are rare.

C. Implications

Integrating AI into mental health care enables a shift from reactive to proactive support. The fusion of linguistic and behavioral insights can power timely, scalable interventions across sectors such as education, workplaces, and clinical settings [20]. However, building interpretable, culturally sensitive, and privacy-compliant systems must remain a priority for responsible deployment.

VII. CONCLUSION

This study presents a modular AI framework for the early detection of mental health conditions using textual communication data. By combining psycholinguistic analysis, contextual language modeling with transformers, and behavioral metadata integration, the system enables a comprehensive understanding of users' mental states.

Key findings from recent research highlight the effectiveness of NLP techniques in detecting linguistic indicators of distress, the enhanced contextual sensitivity provided by transformer models, and the added diagnostic value of behavioral signals such as posting patterns and interaction frequency.

The proposed architecture is designed with adaptability and ethics in mind. It supports privacy-preserving analysis, modular fine-tuning, and potential deployment across various domains including education, workplaces, and clinical support environments.

Looking ahead, there is a need to develop multilingual, interpretable, and culturally adaptive

systems that ensure fairness and inclusivity. Advancing toward real-time, user-aware interventions grounded in ethical AI principles will be essential for transforming mental health care into a more proactive and accessible system.

REFERENCES

[1] M. A. Mansoor and K. H. Ansari, "Early Detection of Mental Health Crises through Artificial-Intelligence-Powered Social Media Analysis: A Prospective Observational Study," Journal of Personalized Medicine, vol. 14, no. 9, p. 958, 2024.

[2] M. F. Islam, T. Ali, R. Kumar, and A. Gupta, "Interpretation of Neural Networks is Susceptible to Universal Adversarial Perturbations," in Proc. Int. Conf. Learn. Representations (ICLR), 2020.

[3] Q. B. Saeed and I. Ahmed, "Early Detection of Mental Health Issues Using Social Media Posts," arXiv preprint arXiv:2503.07653, 2025.

[4] J. Shin, M. Kim, and H. Choi, "FedTherapist: Mental Health Monitoring with User- Generated Linguistic Expressions on Smartphones via Federated Learning," arXiv preprint arXiv:2310.16538, 2023.

[5] B. Shickel, D. Johnson, A. Green, and R. Joseph, "Automatic Detection and Classification of Cognitive Distortions in Mental Health Text," arXiv preprint arXiv:1909.07502, 2019.

[6] A. G. Reece, R. M. Danforth, and C. A. Ayers, "Forecasting the Onset and Course of Mental Illness with Twitter Data," arXiv preprint arXiv:1608.07740, 2016.

[7] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial Examples Are Not Bugs, They Are Features," in Adv. Neural Inf. Process. Syst., vol. 32, 2019.

[8] K. Shin, D. L. K. Wong, and A. R. Lu, "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts," in Proc. Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4222–4235.

[9] B. Liu, X. Chen, M. Ren, and T. Zhang, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," arXiv preprint arXiv:2102.07350, 2021.

[10] J. Hamborg, M. Norman, M. Röttger, I. Augenstein, and N. Kando, "MetaPrompting: Learning to Learn Better Prompts," in Findings of the Association for Computational Linguistics: EMNLP, pp.

258–270, 2022.

[11]Matteo Malgaroli, Thomas D. Hull, James M. Zech, Tim Althoff, “Natural Language Processing for Mental Health Interventions: A Systematic Review and Research Framework,” Translational Psychiatry, vol. 13, article 309, 2023.

[12]R. Walambe, A. Patil, and S. Kale, "Employing Multimodal Machine Learning for Stress Detection," arXiv preprint arXiv:2306.09385, 2023.

[13]M. Asif, A. Kumar, R. Jain, and S. Sinha, "Proactive Emotion Tracker: AI-Driven Continuous Mood and Emotion Monitoring," arXiv preprint arXiv:2401.13722, 2024.

[14]M. Jelassi, S. Bouhlel, and A. Abbes, "Enhancing Personalized Mental Health Support Through Artificial Intelligence: Advances in Speech and Text Analysis Within Online Therapy Platforms," Information, vol. 15, no. 12, p. 813, 2024.

[15]H. Zhang, Y. Li, D. Wang, and M. Ma, "A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks," arXiv preprint arXiv:2307.15915, 2023.

[16]M. Gao, L. Feng, Y. Zeng, and H. Chen, "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications," arXiv preprint arXiv:2307.06894, 2023.

[17]A. Gong, H. Li, and Q. Zhao, "SpeechGuard: Exploring the Adversarial Robustness of Multimodal Large Language Models," arXiv preprint arXiv:2310.00552, 2023.

[18]Y. Li, T. Li, and Z. Li, "Deep Reinforcement Learning: A Survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 10, pp. 4165–4186, 2021.

[19]A. Wang, C. Li, and X. Chen, "Enhancing Large Language Model Performance through Prompt Engineering Techniques," arXiv preprint arXiv:2309.00623, 2023.

[20]S. Schick, T. Min, and P. Liang, "The Prompt Report: A Systematic Survey of Prompting Techniques," arXiv preprint arXiv:2302.11382, 2023.

[21]T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural Language Processing Applied to Mental Illness Detection: A Narrative Review," npj Digital Medicine, vol. 5, no. 1, p. 46, 2022.

[22]H. Singh, R. Verma, and T. Joshi, "Innovative Framework for Early Estimation of Mental Disorder Scores to Enable Timely Interventions," arXiv preprint

arXiv:2502.03965, 2025.

[23]S. Han, R. Mao, and E. Cambria, "Hierarchical Attention Network for Explainable Depression Detection on Twitter Aided by Metaphor Concept Mappings," arXiv preprint arXiv:2209.07494, 2022.

[24]A.-T. Kuo, H. Chen, Y.-H. Kuo, and W.-S. Ku, "Dynamic Graph Representation Learning for Depression Screening with Transformer," arXiv preprint arXiv:2305.06447, 2023.

[25]Liangyin Feng, Shuo Sun, Siyu Wang, et al., “Large Language Models for Mental Health Applications: Systematic Review,” Journal of Medical Internet Research – Mental Health, 2024.