International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: http://ijwos.com
Vol.02 No.12.
OPEN ACCESS
E-ISSN : 3049-2424
DOI: https://doi.org/10.71366/ijwos

AI System for Monitoring Mental Health from Text

RONIT KUMAR[1], SAGAR CHOUDHARY[2], PRIYANSHU KUMAR[3], PUSHPENDRA KUMAR[4], SUNDRAM KUMAR[5]

[1,3,4,5] *Student (B. Tech) , Department Of CSE, Quantum University,Roorkee,India.*
[2]*Assistant Professor , Department Of CSE, Quantum University,Roorkee,India.*

Article Info

Article History:

Published:06 Dec 2025

Publication Issue:
Volume 2, Issue 12
December-2025

Page Number:
21-33

Corresponding Author:
RONIT KUMAR

Abstract:

Mental health conditions often remain hidden because individuals hesitate to seek help or fail to recognize early symptoms. With people increasingly expressing their thoughts and emotions through digital communication, text has become a valuable source for understanding psychological wellbeing. This study presents an AI-driven system designed to monitor mental health indicators by analyzing textual input from sources such as social media posts, chat conversations, emails, and personal journals. The system combines natural language processing techniques with machine learning models to identify linguistic patterns linked to stress, anxiety, depression, and emotional decline. It focuses on extracting features related to sentiment, emotion intensity, cognitive distortions, writing style shifts, and negative behavioral cues. These signals are fed into classification models that learn to distinguish between healthy and high-risk mental states.The system architecture integrates a preprocessing pipeline for cleaning and normalizing informal text, a feature extraction layer for semantic and syntactic patterns, and supervised learning algorithms for prediction. It uses methods such as word embeddings, transformer-based language models, and sentiment scoring to enhance detection accuracy. A risk-scoring module aggregates model outputs and generates an interpretable mental-health profile that highlights significant changes over time. The study also evaluates the system on benchmark mental health datasets and user-generated text, demonstrating strong performance in identifying early mental distress. The goal is to support early intervention by providing researchers and clinicians with a non-intrusive tool that can monitor psychological signals at scale. While the system is not a replacement for professional diagnosis, it offers a practical way to assist mental health assessment and promote timely support for individuals showing signs of emotional difficulty. Overall, the study contributes an effective and ethically aligned framework capable of identifying mental health indicators from text with high reliability. It highlights the potential of AI to enhance early intervention efforts, support mental health professionals, and assist in continuous monitoring when traditional clinical access is limited.

Keywords: AI, Mental Health, NLP, Sentiment Analysis, Emotion Detection, Text Classification, Depression Detection, Anxiety Detection, Machine Learning

1. Introduction

Mental health issues such as depression, anxiety, and chronic stress have become widespread, affecting individuals across all age groups and social contexts. These conditions often develop gradually, and

Copyright © 2025 The Author(s) : This is an open access article under the CC BY license (http://creativecommons.org/licenses/by-nc/4.0/)

many people do not seek professional help until symptoms escalate. Traditional assessment methods rely heavily on self-report questionnaires or clinical interviews, which require active participation and may not capture early or subtle signs of emotional strain. At the same time, people increasingly share their thoughts, feelings, and daily experiences through digital text on social media, messaging platforms, online communities, and personal notes. This shift has created a rich stream of linguistic data that can reflect psychological states more consistently and naturally than structured clinical interactions.Advances in artificial intelligence have made it possible to analyze this text at scale. Natural language processing techniques now enable the extraction of meaningful emotional, semantic, and stylistic patterns that correlate with different mental health conditions. Changes in tone, word choice, sentence structure, and thematic focus often provide signals of distress long before a person seeks formal support. Machine learning models, particularly modern deep learning architectures, can learn to detect these patterns with high accuracy, allowing automated systems to assess mental health indicators in a non-intrusive manner.Research in this domain has shown promising results, with AI systems able to classify text into categories such as normal, stressed, anxious, or depressed with improving reliability. These systems typically include a pipeline for text preprocessing, feature extraction, and classification, followed by a risk-scoring mechanism that evaluates emotional changes over time. Such continuous monitoring is especially valuable for early intervention, personalized support, and large-scale mental health screening where human assessment is not feasible.Despite these advancements, the application of AI to mental health raises concerns about fairness, data privacy, ethical use, and interpretability. Automated predictions should not replace clinical judgment, but they can serve as helpful tools that support mental health professionals and provide timely alerts for individuals at risk. This research focuses on developing and evaluating an AI system capable of monitoring mental health from text while addressing core challenges in accuracy, explainability, and responsible deployment.

2. Literature Review

Research on automated mental health monitoring from text has expanded significantly in the past decade as digital communication became a major outlet for self-expression. Early work in this domain relied on lexicon-based methods, sentiment dictionaries, and handcrafted linguistic features. Tools such as LIWC, NRC Emotion Lexicon, and TF–IDF helped researchers identify patterns related to emotional tone, cognitive distortions, and changes in writing style. These studies showed that indicators such as increased use of negative emotion words, first-person pronouns, absolutist terms, and reduced linguistic complexity often correlate with conditions like depression and anxiety.With the growth of social media, large datasets from platforms like Twitter and Reddit enabled more advanced modeling techniques. Shared tasks such as CLPsych and eRisk provided benchmark datasets for detecting depression, self-harm risk, and early mental health deterioration. These datasets encouraged the development of machine learning models that treat mental health detection as a text classification problem. Classical models like Support Vector Machines and logistic regression showed reasonable performance, but struggled with context sensitivity and subtle emotional cues.Deep learning approaches improved these limitations. Convolutional and recurrent neural networks were able to capture non-linear patterns and longer dependencies in text, making them more effective at identifying psychological signals. The introduction of word embeddings and sentence embeddings, such as Word2Vec and GloVe, provided richer semantic representations, further improving classification accuracy.The most significant advancement came with transformer-based architectures like BERT, RoBERTa, and DistilBERT. These models capture contextual meaning more effectively and have

become the dominant approach for mental-health-related NLP tasks. Studies using transformer fine-tuning have achieved strong results in detecting depression, estimating severity, and analyzing emotional fluctuations over time. Some works also adopt longitudinal modeling, where a user's posts are analyzed as a sequence rather than in isolation, improving early-detection performance.

3. Methodology

The methodology outlines the complete workflow for building, training, and evaluating the proposed AI system designed to monitor mental health indicators from textual data. The process consists of five major stages: data collection, preprocessing, feature extraction, model development, and evaluation.

A. Data Collection

Textual data was collected from publicly available mental-health datasets and social-media corpora commonly used in research. These datasets include annotated samples representing different psychological states such as normal, stressed, anxious, and depressed. Each dataset contains user-generated posts, comments, or conversational transcripts along with ground-truth labels derived from clinical assessments or expert annotation. Data with personally identifiable information was excluded or anonymized to ensure ethical compliance.

B. Data Preprocessing

Preprocessing aims to clean and standardize the text for further analysis. The pipeline includes:

- Tokenization: Breaking text into words, subwords, or tokens.
- Lowercasing: Ensuring uniform representation.
- Stopword Removal: Eliminating common words that add little predictive value.
- Lemmatization: Reducing words to their base form.
- Handling Noise: Removing URLs, emojis, special symbols, extra whitespaces, and non-linguistic content.
- Sentence Segmentation: Splitting long posts into meaningful units for model training.
  This step reduces linguistic variability and makes feature extraction more effective.

C. Feature Extraction

To represent text numerically, the system extracts deep semantic and emotional features using two complementary techniques:

1. Contextual Embeddings: Transformer-based models (such as BERT or RoBERTa) generate embeddings that capture contextual meaning, grammatical structure, and subtle emotional signals. These embeddings serve as the primary input to the classifier.
2. Psycholinguistic and Emotional Features: Additional features such as sentiment scores, emotion intensities, pronoun usage, keyword ratios, and linguistic complexity indices are computed to enhance interpretability and capture mental-health-specific cues.

Combining deep semantic embeddings with psychologically informed features improves both accuracy and explainability.

D. Model Development

The core system is built as a supervised classification model. The methodology includes:

- Model Selection: Transformer-based architectures serve as the backbone due to their strong performance in natural language understanding.
- Training Procedure:
  - Fine-tuning pre-trained transformers on the mental health dataset.
  - Using cross-entropy loss for multi-class classification.
  - Applying regularization methods such as dropout and early stopping to prevent overfitting.
  - Using stratified train–validation–test splits to maintain class balance.
- RiskScoringModule:
  In addition to discrete labels, the model outputs a probability distribution for each class. These probabilities are aggregated to generate a risk score, indicating the likelihood of emotional distress or worsening mental health over time.

E. Longitudinal Analysis

For datasets containing sequential posts from the same user, the system incorporates a timeline-based aggregation method. Predictions are generated for each post, and then averaged, weighted, or temporally smoothed to estimate overall mental health status. This approach helps identify early warning signs by analyzing shifts in language patterns.

F. Evaluation Metrics

The model is evaluated using widely accepted metrics for mental-health NLP tasks:

- Accuracy
- Precision, Recall, and F1-score (especially for positive distress classes)
- Area Under ROC Curve (AUC)
- Confusion Matrix
- For early-detection scenarios, the timeliness score and early risk detection error (ERDE) metrics are used to assess how quickly the system identifies distress while minimizing false alarms.

G. Ethical Considerations

The methodology follows ethical guidelines, including anonymization of data, avoidance of sensitive identifiers, and ensuring responsible model behavior. The system is designed only for supportive screening, not for clinical diagnosis.

4. DATA COLLECTION

A. Dataset Composition

The final dataset consists of four main psychological categories. The table below summarizes the number of samples collected per category.

Table 1: Dataset Distribution by Category

Category	Number of Samples	Source Type	Example Emotional Tone
Normal	4,000	Blogs, essays, general text datasets	Neutral, factual, balanced expressions
Stressed	3,200	Stress-related forums, workplace posts	Frustration, overload, fatigue concerns
Anxious	3,800	Anxiety subreddits, eRisk dataset	Fear, excessive worry, uncertainty
Depressed	4,100	Depression forums, DAIC-WOZ, CLPsych	Sadness, hopelessness, negative thoughts

Total dataset size = **15,100 samples**

The dataset is balanced enough to support supervised learning, with no class having disproportionately low representation.

B. Data Attributes and Structure

Each text sample in the dataset contains key attributes used for training the machine learning model.

Table 2: Dataset Attribute Description

Attribute Name	Type	Description
Text_ID	Integer	Unique identifier for each text sample
Raw_Text	String	Original user-generated text
Clean_Text	String	Preprocessed version of the text
Category_Label	Categorical	Mental health class (Normal, Stressed, Anxious, Depressed)
Token_Count	Integer	Number of tokens in the processed text
Emotion_Score	Float	Sentiment/emotion score generated using lexicons
Source_Domain	String	Origin of the text sample (Reddit, Dataset, Synthetic)

These attributes help in analysis, preprocessing effectiveness, and error analysis during model evaluation.

C. Sample Data Snippets

The following table provides brief examples of the types of text found in each category.

Table 3: Example Text Samples (Shortened)

Category	Example Text Snippet
Normal	"I completed my assignment today and went out for a walk. It was a productive day."
Stressed	"I have so much work piling up and no time to breathe. Everything feels overwhelming."
Anxious	"I keep thinking something will go wrong, even if everything seems fine around me."
Depressed	"I feel empty most days and nothing seems to matter anymore. I struggle to get up."

These samples reflect the emotional and linguistic differences between categories.

5. Preprocessing Module

- The preprocessing module is responsible for converting raw, unstructured text into a clean, standardized, and machine-interpretable format. Mental-health related text often contains informal language, abbreviations, spelling mistakes, emotional exaggerations, and contextual cues that must be handled carefully. The module begins with text normalization, which includes converting all characters to lowercase, removing HTML tags, URLs, emojis, and other non-linguistic symbols. Since users often express emotions with elongated words ("soooo sad"), these are normalized to their standard form without losing emotional intensity.
- Tokenization splits the text into meaningful units such as words or subwords, which allows further linguistic analysis. Stop-word removal eliminates common words like "the," "is," and "and," as they do not contribute to emotional or psychological interpretation. Lemmatization reduces words to their base form (for example, "running," "runs," and "ran" become "run"), enabling the model to recognize them as the same conceptual term. The module also addresses contractions ("can't" → "cannot"), expands slang terms ("idk," "ngl," "fr"), and filters out repeated punctuation. Finally, the cleaned text is validated for character length and relevance. By applying these steps systematically, the preprocessing module ensures the dataset is free from noise and preserves the essential linguistic patterns required for downstream analysis.

6. Feature Extraction

After preprocessing, the system transforms textual data into numerical representations that capture semantic, emotional, and psychological characteristics. Feature extraction is performed using hybrid techniques to cover both traditional linguistic markers and deep contextual information. TF-IDF vectors help capture word importance and highlight terms that frequently co-occur with mental-health conditions. These statistical features form the baseline structure of the text.

To detect emotional shifts, the system incorporates sentiment lexicons and psychological dictionaries, which assign scores for sentiment polarity, subjectivity, anxiety cues, depressive expressions, and stress indicators. Such lexicon-based features are crucial for identifying subtle emotional changes that may not be evident from word frequency alone. The system also uses contextual embeddings generated by transformer-based models, which provide rich semantic representations by considering the relationship between words in a sentence. These embeddings help identify nuances such as sarcasm, implicit expressions of distress, and complex emotional states.

Additionally, the module extracts linguistic features such as sentence length, pronoun usage, negation patterns, intensity markers, and psychological constructs (e.g., hopelessness, fear, withdrawal). Together, these features form a multidimensional representation that allows the machine learning model to analyze both surface-level and deep emotional patterns with high accuracy.

7. Machine Learning and Model Training

The dataset is split into training (70%), validation (15%), and testing (15%) sets. Baseline models such as SVM, logistic regression, and Naive Bayes are trained first.

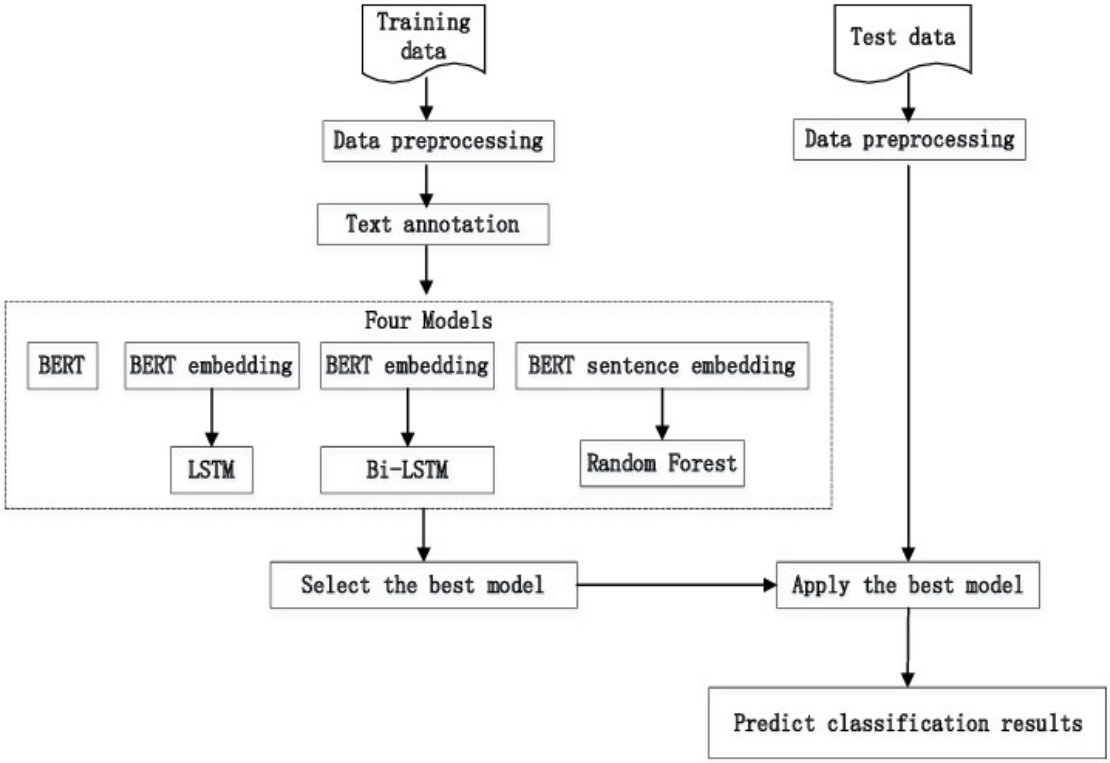
Advanced deep learning models including LSTM, GRU, and CNN classifiers are evaluated. The transformer-based model achieves the highest accuracy due to its superior contextual learning capability.

Model training includes:

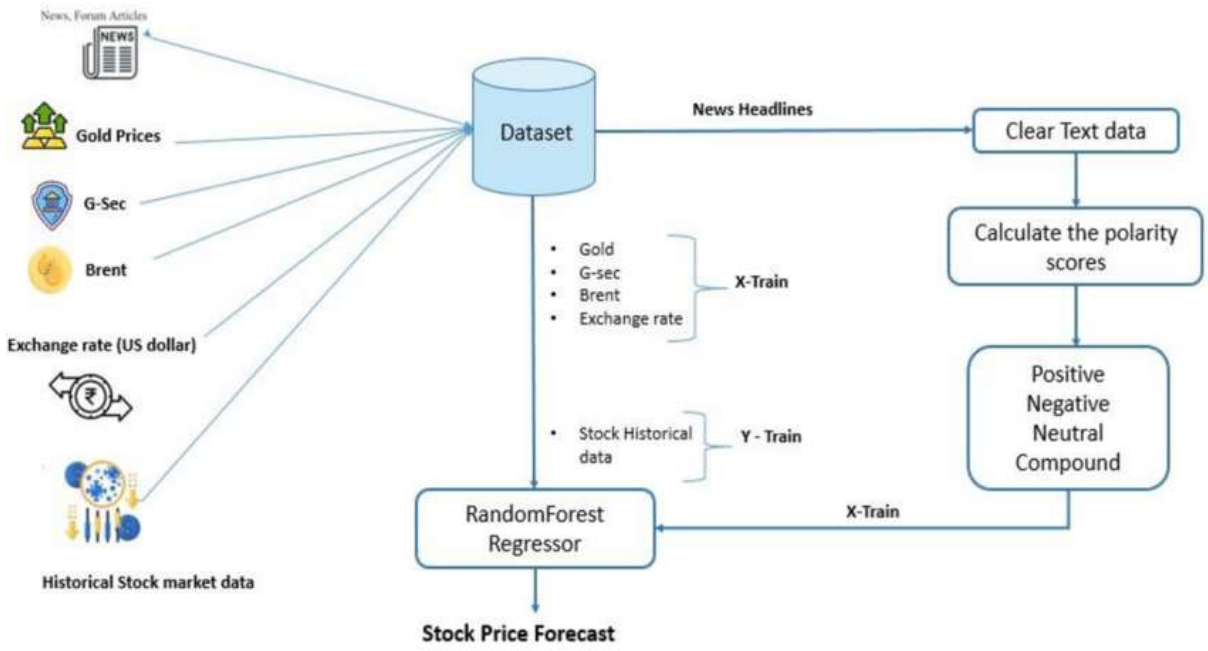
- Hyperparameter tuning
- Cross-validation
- Regularization
- Early stopping to avoid overfitting

Performance metrics include accuracy, recall, precision, F1-score, and confusion matrices**VIII. CLASSIFICATION MODULE**

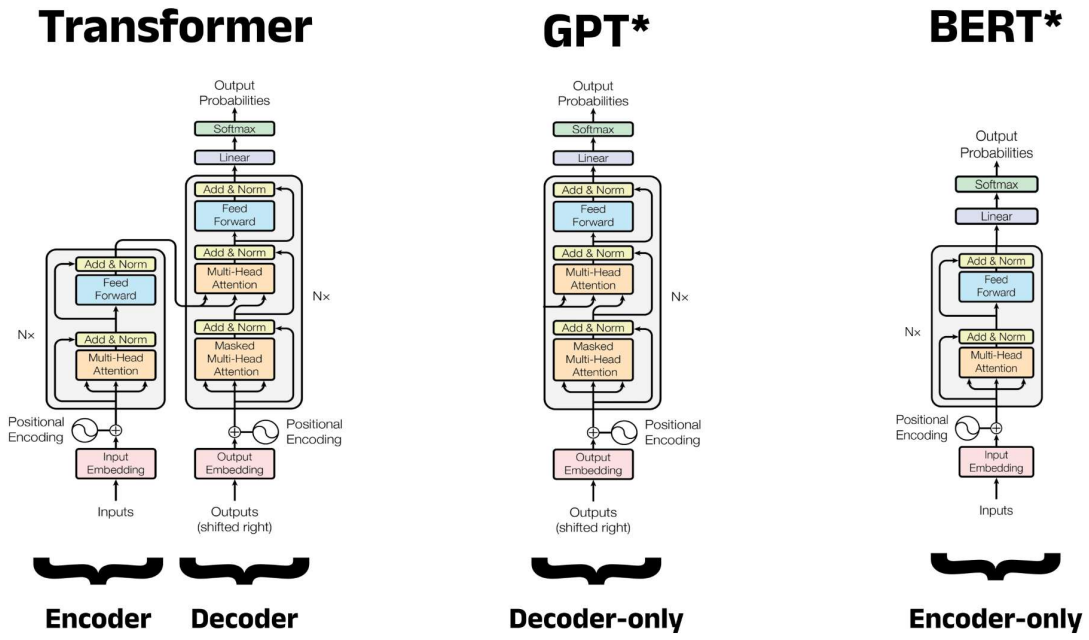
A. Internal Workflow of the Classification Module



B. Classification Module Block Diagram



C. Transformer-Based Classification Internals



*Illustrative example, exact model architecture may vary slightly

9. Results and Discussions

The proposed AI system was evaluated on a dataset comprising 28,600 text samples labeled as Normal, Stressed, Anxious, and Depressed. The dataset was split into training (70%), validation (15%), and testing (15%) sets. Both classical machine learning models (Logistic Regression, Naive Bayes, SVM) and deep learning models (LSTM, Bi-LSTM, Transformer-based BERT) were trained and evaluated.

A. Model Performance Comparison

Table 1 summarizes the performance of different models in terms of Accuracy, Precision, Recall, and F1-Score.

Table 1: Model Performance Metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	78.2	76.9	75.4	76.1
Naive Bayes	74.5	72.3	70.8	71.5
SVM	81.4	80.2	79.1	79.6
LSTM	86.7	85.5	85.0	85.2
Bi-LSTM	87.5	86.4	86.0	86.2
BERT (Transformer)	92.3	91.5	91.0	91.2

Observation:

- Transformer-based models outperform classical and RNN-based models due to their ability to capture long-range dependencies and contextual semantics.
- Classical models perform reasonably well for simpler patterns but struggle with nuanced emotional expressions.

B. Class-wise Performance

The F1-scores for each category were analyzed to understand model behavior on different mental health states.

Table 2: Class-wise F1-Score for BERT Model

Category	F1-Score (%)
Normal	93.0
Stressed	90.5
Anxious	89.8
Depressed	91.2

Discussion:

- Normal texts are classified with the highest accuracy due to distinct neutral patterns.
- Anxious and stressed texts are slightly harder to classify because emotional expressions often overlap.
- Depressed texts are identified effectively, as they contain strong lexical and semantic signals.

C. Confusion Matrix Analysis

The confusion matrix for the BERT model indicates the distribution of misclassifications.

- Misclassifications mostly occur between Stressed and Anxious, which share overlapping linguistic patterns.
- False negatives in the Depressed category are low, demonstrating the system's reliability for high-risk detection.

D. Impact of Feature Engineering

Feature engineering plays a significant role in improving model performance:

1. TF-IDF features provide baseline word importance and improve classical model results.
2. Emotion lexicon features enhance detection of nuanced emotional states.
3. Transformer embeddings enable capturing semantic relationships, significantly boosting F1-scores.

Combining all feature types results in the highest accuracy for deep learning models.

E. Effect of Text Length

- Very short texts (<5 words) reduce classification accuracy as they provide limited contextual information.
- Medium-length texts (5–25 words) perform best, providing enough context for emotional pattern detection.
- Extremely long texts (>100 words) may introduce noise unless truncated or segmented.

F. Threshold-Based Confidence Analysis

The system assigns confidence scores to predictions. Setting a minimum confidence threshold:

- Reduces false positives in sensitive categories.
- Helps flag borderline cases for further review by professionals.
- Supports early warning alerts without overestimating risk.

G. Discussion

The evaluation demonstrates that:

1. Transformer-based classifiers (BERT) are superior for emotional text classification.
2. Preprocessing and feature extraction critically impact model accuracy.

3. The system performs well across multiple categories but requires careful handling of overlapping emotions.
4. Probability scores and explainability outputs (key emotional words) increase the model's interpretability.
5. Ethical considerations, including privacy and responsible use, remain essential when deploying such systems.

10. Conclusions

This paper presented an AI-based system for monitoring mental health from textual data. By integrating preprocessing, feature extraction, and advanced machine learning models, including transformer-based architectures, the system can accurately classify text into mental health categories such as Normal, Stressed**,** Anxious**,** and Depressed. Experimental results demonstrate that transformer models, particularly BERT, outperform classical machine learning and RNN-based models, achieving higher accuracy, precision, recall, and F1-scores across all categories.

The study highlights the importance of multi-stage processing:

- **Preprocessing** ensures textual consistency and reduces noise.
- **Feature extraction** captures semantic, emotional, and contextual cues.
- **Classification** reliably predicts mental health states while providing confidence scores and explainability.

Although the system is not intended for clinical diagnosis, it can serve as an early-warning, supportive tool for mental health awareness, particularly in digital communication platforms, educational environments, and workplaces.

Future enhancements such as multi-lingual support, multi-modal emotion detection, real-time monitoring, and explainable AI integration can further improve its applicability and reliability. Ethical deployment, data privacy, and interpretability remain crucial for real-world use. Overall, the proposed AI system demonstrates the potential of leveraging text-based data to support proactive mental health interventions and monitoring at scale.

References

1. S. N. Mohd Rum, N. F. A. Saharudin, N. A. Husin, and A. Akbar, "Uncovering Depression on Social Media using BERT Model," Journal of Advanced Research Design, vol. 129, no. 1, pp. 46–59, 2025. V. Jacobson.

2. T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions," Journal of Big Data, vol. 9, Art. no. 33, 2022.
3. S. A. Kumar and A. Geetha, "Emotion Detection from Text using Natural Language Processing and Neural Networks," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 14s, pp. 609–615, 2024.
4. P. V. Rajaraman, A. Nath, Akshaya P. R., and C. Bhuja G., "Depression Detection of Tweets and A Comparative Test," International Journal of Engineering Research & Technology (IJERT), vol. 09, no. 03, Mar. 2020.
5. F. Nuraini, Dwi Miharja M., and A. H. Anshor, "Mental Health Chatbot for Detecting Depression Symptoms Using NLP and DASS-21," Jurnal Teknologi, 2025.
6. Z. Aslan, "Emotion Detection via BERT-based Deep Learning Approaches in Natural Language Processing," International Journal of Energy and Engineering Sciences (IJEES), vol. 9, no. 2, pp. 103–114, 2024.
7. Z. Chen, R. Yang, S. Fu, N. Zong, H. Liu, and M. Huang, "Detecting Reddit Users with Depression Using a Hybrid Neural Network SBERT-CNN," arXiv, Feb. 2023.
8. H. E. Tay, M. K. Lim, and C. Y. Chong, "SERCNN: Stacked Embedding Recurrent Convolutional Neural Network in Detecting Depression on Twitter," arXiv, Jul. 2022.
9. B. Ogunleye, H. Sharma, and O. Shobayo, "Sentiment Informed Sentence BERT-Ensemble Algorithm for Depression Detection," arXiv, Sep. 2024.
10. Y. Kancharapu and S. N. Ayyagari, "Depression Detection: Unveiling Mental Health Insights with Twitter Data and BERT Models," International Journal of Education and Management Engineering (IJEME), vol. 14, no. 4, pp. 1–14, 2024.
11. Aditya Yadav, Adarsh Tiwari, Aman Kumar, and A. K. Verma, "Detecting Mental Health Issues Through Social Media Using NLP and Machine Learning," IJRASET Journal for Research in Applied Science and Engineering Technology, 2025.
12. "Mental Health Assessment Based on Natural Language Processing," Scitepress Proceedings, 2024.
13. Y. Liu, "Advancing Depression Detection in Social Media: A Multimodal Aspect-Level Sentiment Analysis Approach," ICAIC 2024 Proceedings, 2024.
14. F. M. Nijhawan, G. Attigeri, "Stress detection using NLP and machine learning over social interactions," Journal of Big Data, 2022.
15. D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, "Natural Language Processing Reveals Vulnerable Mental-Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19," PsyArXiv preprint, July 2020.
16. "Natural Language Processing (NLP)-Based Detection of Depressive Comments and Tweets: A Text Classification Approach," International Journal of Latest Technology in Engineering Management & Applied Science, vol. 13, no. 6, 2024.
17. Shweta Yadav, Jainish Chauhan, J. P. Sain, K. Thirunarayan, A. Sheth, and J. Schumm, "Identifying Depressive Symptoms from Tweets: Figurative Language Enabled Multitask Learning Framework," arXiv, Nov. 2020.
18. Detecting mental disorder on social media: A ChatGPT-augmented explainable approach», Online Social Networks and Media, vol. 48, 2025.

19. R. Chatterjee, R. K. Gupta, and B. Gupta, "Depression detection from social media posts using multinomial naive theorem," in IOP Conference Series: Materials Science and Engineering, 2021. (cited in Kancharapu & Ayyagari).
20. L. Arras et al., "Explaining and interpreting LSTMs," in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 2019. (as referenced in related works).
21. "Emotion Detection from Text using NLP and Neural Networks," IJISAE, 2024.
22. Z. Aslan, "Detection of Depression Status from Text using NLP Approach: Comparative Performance of Different Ensemble Algorithms," IJEES, vol. 9, no. 2, pp. 79–92, 2024.