# Leveraging Emotions for Enhanced Mental Health Prediction Using Machine Learning Algorithms

Aishwarya Daga
*School of Computer Science and Enginnering*
*Vellore Institute of Technology*
Vellore, India
aishwarya.daga2021@vitstudent.ac.in

Prachi Jain
*School of Computer Science and Enginnering*
*Vellore Institute of Technology*
Vellore, India
prachi.jain2021@vitstudent.ac.in

N. Kopperundevi
*School of Computer Science and Enginnering*
*Vellore Institute of Technology*
Vellore, India
kopperundevi.n@vit.ac.in

2025 11th International Conference on Communication and Signal Processing (ICCSP) | 978-1-6654-5736-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCSP64183.2025.11088443

***Abstract*— Mental health disorders are a growing global concern, yet early detection remains difficult due to stigma and limited awareness. This study investigates the prediction of mental health issues using a text-based dataset of user-generated content labeled as poisonous or non-poisonous. Classical machine learning models, including Random Forest, Naïve Bayes, and LSTM, were evaluated on the baseline dataset. To improve predictive performance, an emotion-augmented dataset was introduced, incorporating selected emotion features to capture nuanced sentiment variations. A fine-tuned RoBERTa model was then developed, leveraging these emotion features to enhance classification accuracy. The proposed model achieved an F1-Score of 0.92, significantly outperforming both baseline models and traditional machine learning techniques. These findings highlight the potential of integrating emotion-aware features to improve mental health detection and offer a scalable, data-driven approach for early intervention, particularly in resource-constrained settings. By leveraging natural language processing and deep learning, this work contributes to advancing automated mental health assessment, aiding in timely and effective support for individuals at risk.**

***Keywords*— Emotion-aware classification, machine learning, mental health detection, RoBERTa, text analysis.***

## I. Introduction

Across various demographics mental health disorders have become a global concern, with millions of people being affected. Despite their prevalence, early detection remains a significant challenge, especially due to stigmas, cultural barriers and limited resources in many parts of the world. Traditional methods for detecting mental health issues often rely on clinical assessments, which can be resource-intensive and inaccessible in many regions. While text-based analysis using machine learning has shown promise, it struggles to capture the subtleties of emotional expression, leading to suboptimal predictive performance.

Developing scalable, accurate, and efficient systems for early detection of mental health issues can facilitate timely interventions, improving outcomes for individuals and reducing the societal burden of untreated mental health disorders. This work proposes leveraging emotion detection to enhance text-based machine learning models for mental health prediction. By augmenting a dataset of user-generated comments labelled as poisonous or non-poisonous with emotion attributes, models are enabled to capture nuanced sentiment variations, improving predictive accuracy. The performance of classical machine learning models and a fine-tuned RoBERTa model evaluated, demonstrating the effectiveness of incorporating emotion features.

The primary contributions of this work include: (1) augmenting a text-based dataset with emotion features for enriched analysis, (2) evaluating the performance of classical models on the basic dataset, and (3) demonstrating the superior performance of an emotion-aware RoBERTa model, achieving an F1-Score of 0.92. These findings highlight the potential of integrating emotion features for scalable and effective mental health prediction.

## II. Related Works

Mental health is a critical aspect of overall well-being, and the rise of digital platforms has emphasized the need for effective methods to detect mental health issues. Recent advancements in machine learning (ML) have provided novel approaches to this challenge, utilizing data from various sources, including textual content. This survey reviews relevant works to contextualize the proposed method of using emotion-based features to fine-tune the RoBERTa model for detecting poisonous comments as indicators of potential mental health concerns.

R. Abd Rahman et al. [1] provided a systematic review of ML methods in mental health detection, highlighting the utility of diverse algorithms and its performance on different datasets. The method underscores the importance of feature engineering and the role of textual data in understanding mental health indicators, establishing a foundation for to focus on comment-based detection.

Chung and Teo [2] proposed a taxonomy of ML techniques applied to mental health prediction, identifying key challenges such as the scarcity of annotated datasets and the need for explainable AI, emphasis to overcome these challenges aligns with the approach of augmenting existing models like RoBERTa with auxiliary features like emotions to enhance interpretability and accuracy.

Iyortsuun et al. [3] reviewed ML and deep learning approaches for mental health diagnosis, demonstrating the potential of deep learning architectures in handling complex data patterns, noted that emotion recognition from text could serve as a valuable signal in understanding mental health states, directly motivating the proposed inclusion of emotion as an auxiliary feature.

Vaishnavi et al. [4] explored the use of ML algorithms for predicting mental health illnesses, focusing on structured data. The work highlights the growing interest in combining multiple data modalities. Tate et al. [5] employed ML techniques to predict mental health problems in adolescents using longitudinal data. Xu et al. [6] introduced Mental-LLM, which leverages large language models (LLMs) for mental health prediction using online text data. The innovative use of LLMs illustrates the potential of advanced natural language processing (NLP) models.

Chaithra and Samanvaya [7] investigated deep learning for text-based emotion recognition, presenting techniques to identify emotional states from textual data. The work validates the feasibility of incorporating emotion recognition into NLP models.

Acheampong et al. [8] reviewed transformer models, particularly BERT-based approaches, for text-based emotion detection. The analysis advantages the transformer architectures in capturing contextual and semantic nuances. Hence the existing body of research highlights the efficacy of ML methods in mental health detection, with a growing focus on text-based approaches. By combining emotion recognition with transformer-based NLP models, the proposed method addresses key gaps, such as enhancing feature richness and improving predictive accuracy. This integration aims to provide a robust framework for identifying poisonous comments as indicators of potential mental health issues, contributing a novel perspective to the field.

Zhao et al. [9] conducted a comprehensive study on user trust in LLM-based mental health applications, emphasizing factors such as response accuracy, perceived empathy, and algorithmic transparency. Through expert interviews, user surveys, and literature reviews, the study underscored the pivotal role of user trust in the acceptance and effectiveness of AI-driven mental health tools. This aligns closely with the goals of the proposed emotion-aware RoBERTa framework, which integrates emotional cues to enhance user engagement and model interpretability. Similarly, Lai et al. [10] introduced Psy-LLM, an LLM-based psychological support system trained on expert Q&A and mental health literature. Designed to operate both independently and as a clinical aid, it demonstrates scalability and responsiveness, underscoring the potential of LLMs to address therapist shortages and reduce user reluctance through emotionally informed dialogue.

Radwan et al. [11] proposed a hybrid framework combining traditional machine learning classifiers with GPT-3-based embeddings for detecting stress-related disorders from social media, achieving 83% accuracy. Their results highlighted the importance of semantically rich embeddings and model explainability, reinforcing the case for integrating emotion-enhanced features in transformer models like RoBERTa. In a related study, Lamichhane [12] evaluated ChatGPT's zero-shot capability in identifying suicidality, stress, and depression, where it outperformed baseline models despite some limitations in multi-class scenarios. These findings further support the notion that incorporating emotion signals into LLMs enhances the detection of nuanced psychological states in complex mental health contexts.

## III. Proposed Methodology

The proposed methodology seeks to integrate emotion detection as an auxiliary feature to enhance machine learning models for detecting poisonous comments, which often signal underlying mental health issues. Poisonous comments, often characterized by hostile, aggressive, or emotionally charged language, can serve as indirect indicators of underlying mental health struggles such as anxiety, anger, or depression. Individuals experiencing psychological distress may express their emotions through toxic language as a coping mechanism or cry for help. Analyzing such comments can help identify at-risk individuals and patterns of emotional dysregulation.

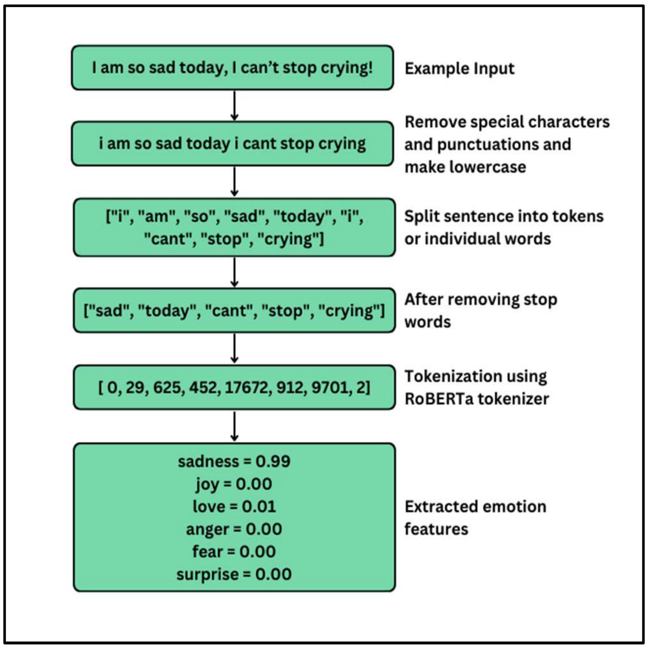
I am so sad today, I can't stop crying!
Example Input
i am so sad today i cant stop crying
Remove special characters and punctuations and make lowercase
["i", "am", "so", "sad", "today", "i", "cant", "stop", "crying"]
Split sentence into tokens or individual words
["sad", "today", "cant", "stop", "crying"]
After removing stop words
[ 0, 29, 625, 452, 17672, 912, 9701, 2]
Tokenization using RoBERTa tokenizer
sadness = 0.99
joy = 0.00
love = 0.01
anger = 0.00
fear = 0.00
surprise = 0.00
Extracted emotion features

Fig. 1. Data Preprocessing and Emotion Extraction

The methodology involves several stages, including data collection, preprocessing, emotion feature extraction, model fine-tuning, and evaluation. The goal is to develop a robust classification model that can accurately identify harmful or toxic comments by considering both textual content and emotional undertones.

The dataset used in this work is the Mental Health Corpus from Kaggle, which contains labeled text comments from individuals dealing with mental health issues such as anxiety, depression, and other related conditions. The corpus has two main columns: one containing the text comments and the other containing labels that indicate whether a comment is poisonous or not. There are 27292 unique comments in the dataset, with 50.5% being non-poisonous and 49.5% being poisonous comments. Initially, the dataset does not include emotion features, which are extracted later as part of the preprocessing phase.

A pre-trained emotion model is used to extract the emotion features due to the significant advantages it offers in terms of efficiency, accuracy, and scalability. Developing an emotion model from scratch requires a large annotated dataset, extensive computational resources, and substantial time to train and fine-tune the model effectively. Leveraging a pre-trained model allows the utilization of the expertise and training embedded in the model, typically fine-tuned on diverse and expansive datasets. This achieves high accuracy with minimal effort and focuses on integrating emotion detection as a feature rather than spending resources on model development. Moreover, pre-trained models are well-optimized to capture nuanced emotional cues and linguistic patterns, making them ideal for the context where precise emotion detection significantly enhances the classification of unhealthy comments.

As shown in Fig. 1, the preprocessing phase begins with basic text cleaning and normalization. Basic text cleaning is done by removing special characters and punctuation, standardizing text case and removing stop words. The text comments are tokenized into individual words or sub-words using a tokenizer compatible with the pre-trained BERT-

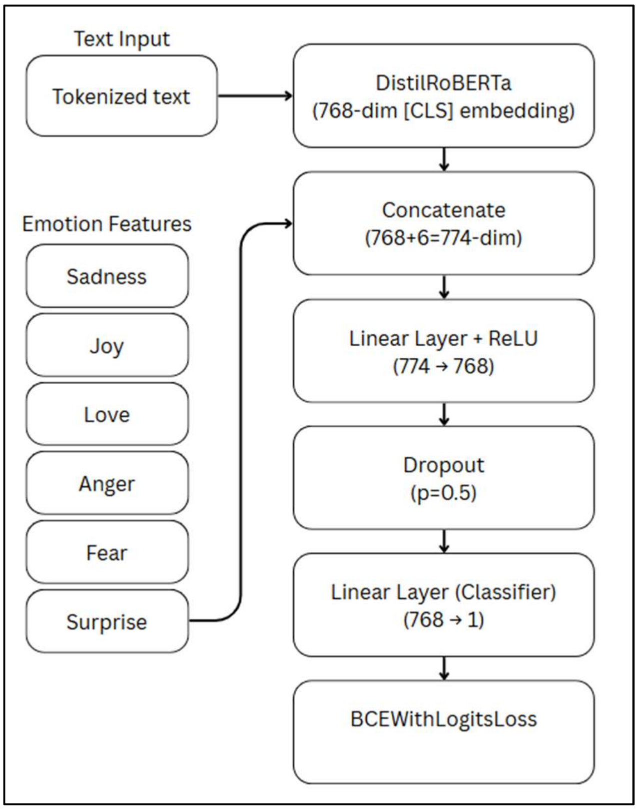

Fig. 2. Fine-tuned RoBERTa Model

based model. After tokenization, the text is vectorized using the pre-trained RoBERTa tokenizer, which converts the tokens into embeddings that represent the semantic meaning of the text. The tokenization process is mathematically represented as:

$$\boldsymbol{T_i = Distil - RoBERTa\,Tokenizer(x_i)} \qquad (1)$$

Where $x_i$ represents the i-th comment in the dataset, and $T_i$ is the tokenized representation of the comment.

In parallel, emotion features are extracted from the text using a pre-trained emotion classification pipeline, nateraw/bert-base-uncased-emotion, available on Hugging Face's model hub. This model is specifically fine-tuned for emotion recognition and predicts six distinct emotional categories: Anger, Fear, Joy, Love, Sadness, and Surprise. The model processes each comment, and the scores for these six emotions are extracted, forming a feature vector $E_i = (e_1, e_2, e_3, e_4, e_5, e_6)$, where each $e_i$ represents the score for one of the six emotions.

These emotion features are then appended to the tokenized text embeddings, resulting in a final feature vector $X_i$ for each comment, which combines both the text and the emotion features:

$$\boldsymbol{X_i \; = \; [T_i\,, E_i]} \qquad (2)$$

Where $T_i$ is the tokenized text embedding for the *i*-th comment and $E_i$ is the emotion feature vector for the same comment, consisting of the six emotion feature scores.

This combined feature vector is used as input to the classification model, which aims to predict whether the comment is Poisonous (indicating potential mental health issues) or Not Poisonous (indicating the absence of such issues).

The classification model is based on DistilRoBERTa, a smaller and more efficient variant of RoBERTa. DistilRoBERTa retains most of the original model's performance while reducing the number of parameters, making it suitable for applications where computational resources may be limited or real-time performance is needed.

The model is fine-tuned by integrating both the tokenized text embeddings and the emotion features. Specifically, the emotion scores are concatenated with the tokenized text embeddings, allowing the model to consider both the textual and emotional context of the comments during training. Fig. 2 represents the architecture of the fine-tuned RoBERTa model. The objective function used for training the model is binary cross-entropy loss, which is suitable for binary classification tasks:

$$\boldsymbol{L = -\frac{1}{N}\sum_{i=1}^{N}[y_i(log\hat{y}_i) + (1 - y_i)\,log(1 - \hat{y}_i)]} \qquad (3)$$

Where *N* is the number of comments in the dataset, $y_i$ is the true label for the *i*-th comment (1 for poisonous, 0 for non-poisonous) and $\hat{y}_i$ is the predicted probability that the *i*-th comment is Poisonous.

For training the model, the AdamW optimizer is employed, which is an enhanced version of the Adam optimizer and includes weight decay to prevent overfitting. The optimizer's update rule is:

$$\boldsymbol{\theta_t \; = \; \theta_{t-1} - \eta\frac{m_t}{\sqrt{v_t}+\epsilon} + \eta\lambda\theta_{t-1}} \qquad (4)$$

Where:

- $\theta_t$ represents the model parameters at time step $t$.
- $m_t$ and $v_t$ are the first and second moment estimates of the gradient for each parameter.
- $\eta$ is the learning rate, which is set to 1.5e-6.
- λ is the weight decay parameter, set to 0.2.

The training hyperparameters include:

- Learning Rate: 1.5e-6, chosen to prevent disrupting the pre-trained model's learned knowledge.
- Weight Decay: 0.2, applied to improve generalization by penalizing large weights.
- Batch Size: 128, chosen based on available computational resources.
- Epochs: 20 epochs, with early stopping implemented. Training stops if the validation loss does not improve for 5 consecutive epochs.

Model performance is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score.

To assess the impact of emotion features, the proposed model is compared with traditional machine learning models such as Naive Bayes, Random Forest, and Long Short-Term Memory (LSTM) networks, which are trained on the same dataset but without the inclusion of emotion features. This methodology combines emotion detection with machine learning techniques to offer a more comprehensive approach to detecting potential mental health issues from textual data. By integrating emotion features, the model is expected to gain a deeper understanding of the emotional context within the text, improving its ability to detect toxic or harmful comments more effectively.

## IV. Results

This section presents a comprehensive evaluation of the predictive performance of several machine learning models employed for the detection of poisonous substances. The models under consideration include Naive Bayes (NB), Random Forest (RF), Long Short-Term Memory (LSTM) networks, and fine-tuned RoBERTa. The evaluation focuses on key classification metrics—namely, accuracy, precision, recall, and F1-score—while also considering the computational and financial costs associated with the training and deployment of these models.

The models were assessed using a standard test dataset, and the results reveal that advanced models, such as LSTM, and fine-tuned LLMs, consistently outperformed traditional machine learning approaches like Naive Bayes and Random Forest in terms of predictive accuracy and class detection capabilities.

Tables I and II summarize the key performance metrics for each model. Table III depicts the confusion matrix for fine-tuned distil-RoBERTa, indicating a high recall for the poisonous class and overall strong classification performance.

Among the evaluated models, the fine-tuned DistilRoBERTa demonstrated the highest performance, achieving an accuracy of 92%. This model exhibited a remarkable recall of 0.95 for the "poisonous" class, highlighting its capacity to accurately identify harmful comments. Its F1-score of 0.92 reflects a well-balanced performance across both precision and recall. However, the fine-tuning process for LLMs required significant computational resources, with training times ranging from several hours to several days depending on the dataset size. High-performance GPUs were essential to accommodate the complexity of the model and ensure its effectiveness.

The LSTM model, while slightly trailing behind the fine-tuned LLM in accuracy at 90%, exhibited commendable performance in capturing sequential dependencies within the data. This model was particularly effective in precision for the "poisonous" class. However, despite its strong performance, LSTM models are known for high computational requirements, especially for large datasets, which limits the practicality in resource-constrained environments.

Random Forest achieved an accuracy of 89.03%, performing well in handling feature interactions and non-linear relationships. However, it faced challenges with noisy or imbalanced data, resulting in a lower recall for the "poisonous" class when compared to more advanced models. While it performs adequately in less complex datasets, its performance could degrade in the presence of unbalanced or noisy data.

The Naive Bayes model, while performing the weakest among the models evaluated with an accuracy of 89.01%, still demonstrated an ability to identify poisonous comments with a relatively high recall of 0.96. This result highlights the model's effectiveness in identifying harmful comments, although its assumption of feature independence limits its ability to capture complex relationships between words, affecting its overall performance on more intricate datasets.

TABLE I.
PERFORMANCE METRIC ANALYSIS

| Model | Accuracy (%) | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Naive Bayes | 89.01 | 0.90 | 0.89 | 0.89 |
| Random Forest | 89.03 | 0.89 | 0.89 | 0.89 |
| LSTM | 90.0 | 0.90 | 0.90 | 0.90 |
| Fine-tuned RoBERTa | 91.8 | 0.92 | 0.92 | 0.92 |

TABLE II
PARAMETER ANALYSIS

| Model | Training Time | Inference Time | Computational Cost |
|---|---|---|---|
| Naive Bayes | Low | Low | Low |
| Random Forest | Moderate | Moderate | Moderate |
| LSTM | High | High | High |
| Fine-tuned RoBERTa | Very High | Moderate | Very High |

TABLE III
CONFUSION MATRIX FOR FINE-TUNED MODEL

| | Non-Poisonous | Poisonous |
|---|---|---|
| Non-Poisonous | 1905 | 233 |
| Poisonous | 111 | 1948 |

The fine-tuned LLM (DistilRoBERTa) emerged as the top-performing model, particularly excelling in terms of recall and F1-score. Its ability to capture the context of words and the relationships within sequences allowed it to outperform the other models in terms of identifying subtle and complex toxic content. Despite its high computational and financial costs, its performance makes it the ideal choice for high-stakes applications requiring the detection of harmful or poisonous comments with high precision and recall.

The LSTM model, while slightly less accurate than the fine-tuned LLM, proved valuable in capturing long-term dependencies within sequential data. This capability makes it a strong contender for tasks that require an understanding of context across multiple tokens or sequences. relationships between features places them at a disadvantage in more challenging classification tasks. However, lower computational cost and simpler deployment make them viable for less demanding use cases or when rapid deployment is essential.

In contrast, Naive Bayes and Random Forest performed adequately but were less suited for handling more complex and noisy datasets. The lack of capacity to capture intricate

## V. Conclusion

The proposed algorithm successfully demonstrated the potential of integrating emotion detection with machine learning models to improve the early detection of mental health issues. By augmenting a text-based dataset with emotion features such as sadness, joy, love, anger, fear, and surprise, the models provide a richer understanding of the underlying emotional states expressed in user-generated content. The fine-tuned RoBERTa model, which incorporated these emotion features, achieved an impressive F1-score of 0.92, significantly outperforming traditional machine learning models such as Naïve Bayes and Random Forest. This result highlights the importance of considering emotional context when analyzing textual data for mental health prediction.

The findings suggest that emotion-aware models can offer more accurate and nuanced insights into individuals' mental health, potentially enabling earlier intervention in resource-constrained settings. Although fine-tuning transformer models like RoBERTa demands considerable computational resources, the benefits in predictive performance make it a promising tool for detecting mental health issues through text analysis. Furthermore, the method emphasizes the importance of addressing privacy and ethical concerns when deploying these models in real-world applications.

## VI. Limitations and Future Work

Despite achieving high performance, the proposed approach has limitations that must be acknowledged. One significant limitation is the computational cost associated with fine-tuning large language models such as DistilRoBERTa. Training these models requires high-performance GPUs and substantial memory, which may not be accessible in resource-constrained environments. This restricts the scalability and real-time applicability of the solution, particularly in mobile or low-resource settings.

While emotion features were integrated to enhance model performance, their effectiveness depends heavily on the accuracy of the pre-trained emotion detection model. Inaccurate or noisy emotion labels can inadvertently introduce errors and reduce the overall robustness of the classifier.

Looking forward, future work can explore the comparative impact of emotion features by conducting an ablation study. This would involve training a version of the fine-tuned DistilRoBERTa model without the additional emotion features to quantify the performance gains attributed specifically to emotional context. Further, optimizing the model for real-time or edge deployment through techniques such as model quantization, pruning, or knowledge distillation could improve accessibility and reduce computational overhead. Expanding the scope of the study to include multilingual datasets and culturally diverse comment samples would also improve the generalizability of the findings. Lastly, integrating this model into practical applications such as mental health support systems or social media moderation tools, with appropriate privacy safeguards and ethical considerations, remains a promising direction for real-world impact.

## References

[1] R. Abd Rahman, K. Omar, S. A. M. Noah, M. S. N. M. Danuri, and M. A. Al-Garadi, "Application of machine learning methods in mental health detection: A systematic review," IEEE Access, vol. 8, pp. 183952-183964, 2020.

[2] J. Chung and J. Teo, "Mental health prediction using machine learning: Taxonomy, applications, and challenges," Applied Computational Intelligence and Soft Computing, vol. 2022, no. 1, p. 9970363, 2022.

[3] N. K. Iyortsuun, S. H. Kim, M. Jhon, H. J. Yang, and S. Pant, "A review of machine learning and deep learning approaches on mental health diagnosis," Healthcare, vol. 11, no. 3, p. 285, Jan. 2023.

[4] K. Vaishnavi, U. N. Kamath, B. A. Rao, and N. S. Reddy, "Predicting mental health illness using machine learning algorithms," in Journal of Physics: Conference Series, vol. 2161, no. 1, p. 012021, 2022.

[5] E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola, "Predicting mental health problems in adolescence using machine learning techniques," PloS One, vol. 15, no. 4, p. e0230389, 2020.

[6] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, "Mental-LLM: Leveraging large language models for mental health prediction via online text data," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 8, no. 1, pp. 1-32, 2024.

[7] V. Chaithra and K. J. Samanvaya, "Text-Based Emotion Recognition Using Deep Learning," in 2024 Second International Conference on Advances in Information Technology (ICAIT), vol. 1, pp. 1-7, IEEE, July 2024.

[8] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," Artificial Intelligence Review, vol. 54, no. 8, pp. 5789-5829, 2021.

[9] Y. Zhao, J. Wu, P. Qu, B. Zhang, and H. Yan, "Assessing user trust in LLM-based mental health applications: Perceptions of reliability and effectiveness," *J. Comput. Technol. Appl. Math.*, vol. 1, no. 2, pp. 19–26, 2024.

[10] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang, "Psy-LLM: Scaling up global mental health psychological services with AI-based large language models," *arXiv preprint arXiv:2307.11991*, 2023.

[11] A. Radwan, M. Amarneh, H. Alawneh, H. I. Ashqar, A. AlSobeh, and A. A. A. R. Magableh, "Predictive analytics in mental health leveraging LLM embeddings and machine learning models for social media analysis," *Int. J. Web Serv. Res.*, vol. 21, no. 1, pp. 1–22, 2024.

[12] B. Lamichhane, "Evaluation of ChatGPT for NLP-based mental health applications," *arXiv preprint arXiv:2303.15727*, 2023.