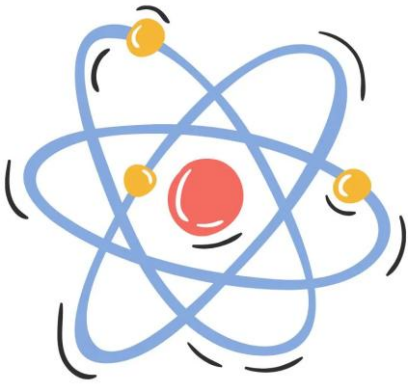

ISSN: 1672 - 6553

JOURNAL OF DYNAMICS AND CONTROL

VOLUME 9 ISSUE 5: 208 - 215

# DEEP LEARNING-BASED MENTAL HEALTH DETECTION USING FINE-TUNED BERT: A MULTICLASS TEXT CLASSIFICATION APPROACH

Urvashi[1], Syed Wajahat Abbas Rizvi[2], P.K. Dwivedi[3], Ashish Kumar Pandey[4], Sandhya[5]

[1,2]Amity University, Uttar Pradesh, India
[3,4]Dr. Rammanohar Lohia Awadh University, Uttar Pradesh, India
[5]Pathfinder, Lucknow, Uttar Pradesh, India

JOURNAL OF DYNAMICS AND CONTROL
DOI: https://doi.org/10.71058/jodac.v9i5018

# DEEP LEARNING-BASED MENTAL HEALTH DETECTION USING FINE-TUNED BERT: A MULTICLASS TEXT CLASSIFICATION APPROACH

**Urvashi[1], Syed Wajahat Abbas Rizvi[2], P.K. Dwivedi[3], Ashish Kumar Pandey[4], Sandhya[5]**

*[1,2]Amity University, Uttar Pradesh, India*
*[3,4]Dr. Rammanohar Lohia Awadh University, Uttar Pradesh, India*
*[5]Pathfinder, Lucknow, Uttar Pradesh, India*
*[1]dwivediurvashi9004@gmail.com, [2]swabbasrizvi@gmail.com, [3]drpkdwivedi@yahoo.co.in,*
*[4]ashishkumarpandey@rmlau.ac.in, [5]sandhya.psy27@yahoo.com*

***Abstract:*** *Mental health disorders, including anxiety, bipolar disorder, and suicidal tendencies, significantly affect individual well-being and necessitate timely detection for effective intervention. Traditional assessment methods, such as clinical evaluations and self-reported surveys, are often time-consuming and subjective. This paper introduces a deep learning-based approach utilizing a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model for multi-class mental health classification through textual analysis. The system classifies text into four categories—depression, anxiety, stress, and normal using advanced natural language processing (NLP) techniques. It features a real-time interface developed using Streamlit, offering an accessible and intuitive tool for clinicians and users. Evaluation on the DAIC-WOZ and Reddit-based datasets yields an accuracy of 93%, demonstrating the model's strong performance. A user-friendly Graphical User Interface (GUI) was developed to facilitate real-time classification, allowing users to input text and receive immediate feedback. Evaluation through classification metrics confirmed the model's effectiveness. The proposed system offers a scalable and automated approach to mental health assessment, supporting early intervention and reducing the burden on healthcare systems.*

***Keywords:*** **Mental Health, Deep Learning, BERT, Text Classification, Streamlit, NLP**

## I. Introduction

Mental health issues such as anxiety, and stress are prevalent in modern society and require timely identification for effective intervention [1]. Traditional diagnosis relies heavily on self-reported symptoms and clinical evaluations, which may be subject to delays or inaccuracies [2]. With the increasing availability of digital communication, textual data has become a valuable resource for analyzing mental health conditions.

Advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) have facilitated the development of automated systems that assess mental health using user-generated textual data [3]. Transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT), have demonstrated superior performance in contextual understanding and classification accuracy compared to traditional architectures like LSTM and CNN [4].

Recent studies have explored BERT and its variations in mental health prediction. Ji et al. introduced MentalBERT, a domain-specific model tailored for detecting psychological states in social media and clinical text [5]. Similarly, Murarka et al. utilized Mental RoBERT a, achieving improvements in depression and anxiety classification [6]. Hybrid architectures, such as Sentence-BERT with CNN, have achieved an 86% accuracy in depression detection from Reddit posts [7], while a BERT-based summarization model reached an F1-score of 0.81 on the DAIC-WOZ dataset [8].

More recently, Pourkeyvan et al. demonstrated that fine-tuned Hugging Face Transformers significantly outperform conventional models in multi-label classification tasks for mental health [9]. Liu and Su explored multitask BERT models to enhance detection accuracy across mental health conditions [10], and Yang et al. proposed MentaLLaMA, a framework combining large language models with interpretable social media analysis [11].

In social media-based detection, researchers have leveraged linguistic and emotional cues. Li et al. examined emotion signals on Weibo [12], and Skaik and Inkpen employed Twitter data for

DOI: https://doi.org/10.71058/jodac.v9i5018

depression detection using user features and linguistic patterns [13]. Ensemble BERT models have also been proposed to improve robustness in classification [14], while graph-based approaches like MGL-CNN have used hierarchical post aggregation for better context representation [15].

Multi-modal learning is also emerging as a promising direction. Beniwal and Saraswat used a hybrid BERT-CNN model with text and image inputs to detect depression [16], while Yang et al. integrated BERT and TabNet for psychiatric rehabilitation planning [17]. Despite progress, challenges remain in the generalizability, interpretability, and ethical deployment of such models [18]. Issues such as data privacy, cultural language differences, and the risk of false positives in mental health predictions are of critical concern [19]. Moreover, recent evaluations of large language models (LLMs) show inconsistent performance on comorbid mental health conditions, suggesting a need for more robust solutions [20]. This paper presents a fine-tuned BERT-based system integrated with a Streamlit GUI for real-time mental health classification.

## Problem Statement

The prevalence of mental health disorders including anxiety, depression, bipolar disorder, suicidal ideation, and personality disorders is increasing rapidly, posing serious challenges to individual well-being and exerting pressure on global healthcare infrastructure. Timely identification of these conditions is crucial for effective intervention, yet current diagnostic methods, such as clinical evaluations and patient-reported questionnaires, are often labor-intensive, subjective in nature, and not equipped to scale efficiently for widespread use.

In the digital age, people often express their mental states, emotions, and personal experiences through online platforms and social media. These textual expressions form a valuable dataset that, when analyzed properly, can serve as a basis for early mental health screening. However, extracting meaningful insights from unstructured and diverse text data remains a key obstacle—particularly in the context of developing systems that are not only accurate but also interpretable and accessible.

This study seeks to overcome these challenges by introducing an automated mental health classification framework based on deep learning. At the core of the system is a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model, trained to recognize and categorize various mental health conditions based on text inputs. To ensure ease of use and real-time feedback, the model is deployed through an interactive Graphical User Interface (GUI), offering a practical and scalable solution for mental health monitoring and support.

The remainder of this paper is organized as follows: Section II presents the architecture and methodology used to develop the fine-tuned BERT model for mental health classification. Section III discusses how the model generates predictions and classifies unseen textual inputs into predefined mental health categories. Section IV details the end-to-end implementation, including data collection, preprocessing, training, and evaluation. Section V outlines the results of the model's performance using accuracy, precision, recall, F1-score, and confusion matrix metrics, followed by a detailed discussion. Section VI concludes the study and discusses potential directions for future research.

## II. Methodology: Fine-Tuned BERT Model for Mental Health Text Classification

This study employs a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model for multi-class mental health text classification. The model architecture is built upon the bert-base-uncased variant provided by Hugging Face's Transformers library. It consists of three main components: a BERT encoder, a fully connected classification layer, and a softmax layer. The BERT encoder is responsible for extracting deep contextual representations from the input text. These embeddings are then passed through a fully connected layer to map the features to the target classes.

Finally, the softmax layer converts the output logits into probability distributions across the predefined mental health categories. For input processing, the model utilizes WordPiece tokenization to break down text into subword units, ensuring better handling of out-of-vocabulary terms. Special tokens such as [CLS] (classification token) and [SEP] (separator token) are added to maintain sentence structure and positional encoding, which preserves the sequence of words in the input.

DOI: https://doi.org/10.71058/jodac.v9i5018

The training process began with data collection from the DAIC-WOZ dataset, which includes interviews annotated for psychological states. The dataset was expanded to cover multiple categories including Anxiety, Bipolar Disorder, Normal, Personality Disorder, Stress, and Suicidal Thoughts. To enhance class balance and robustness, data augmentation techniques such as synonym replacement and back-translation were applied. During preprocessing, the text data was cleaned, tokenized, and converted into model-compatible formats. The model was optimized using the AdamW optimizer with a cross-entropy loss function. A learning rate scheduler with warmup and decay was employed to stabilize training, and gradient clipping was used to prevent exploding gradients. Training was conducted in mini-batches to improve efficiency and generalization.

Once training was complete, the model was capable of generating predictions for new, unseen textual inputs. The prediction process involves tokenizing and encoding the input text, extracting contextual features using the BERT encoder, computing class probabilities via the softmax layer, and selecting the class with the highest probability as the final prediction. This end-to-end framework allows for accurate and efficient classification of mental health states based on textual data.

## Pseudocode: Fine-Tuned BERT Model for Mental Health Text Classification

**Step 1: Load Pretrained BERT Model**

i. Load BERT-base model from Hugging Face Transformers
ii. Initialize classification head (fully connected layer + softmax)

**Step 2: Data Preparation**

i. Load dataset (e.g., DAIC-WOZ)
ii. Define target categories: ['Anxiety', 'Bipolar', 'Normal', 'Personality Disorder', 'Stress', 'Suicidal']

**Step 3: Data Augmentation**

For each text sample in dataset:
i. Apply synonym replacement
ii. Apply back-translation
iii. Add augmented sample to dataset

**Step 4: Preprocessing**

For each text sample:
i. Clean text (remove special characters, lowercase, etc.)
ii. Tokenize text using WordPiece tokenizer
iii. Add special tokens [CLS] and [SEP]
iv. Convert tokens to input IDs and attention masks

**Step 5: Model Training**

i. Set optimizer as AdamW
ii. Set loss function as CrossEntropyLoss
iii. Initialize learning rate scheduler with warmup and decay
iv. Set gradient clipping threshold
For each epoch:
For each batch in training data:
i. Forward pass through BERT encoder
ii. Compute logits using classification head
iii. Compute loss
iv. Backpropagate gradients
v. Clip gradients
vi. Update model parameters
vii. Update learning rate using scheduler

**Step 6: Model Evaluation**

i. Evaluate model on validation/test set
ii. Calculate metrics: Accuracy, Precision, Recall, F1-Score, Confusion Matrix

**Step 7: Save Trained Model**

i. Save fine-tuned BERT model and tokenizer

DOI: https://doi.org/10.71058/jodac.v9i5018

**Step 8: Inference (Model Predictions)**

i. Function PredictMentalHealthCategory(input_text):
ii. Clean and tokenize input_text
iii. Add special tokens [CLS], [SEP]
iv. Convert tokens to input IDs and attention mask
v. Load fine-tuned BERT model
vi. Pass input through BERT encoder
vii. Generate logits and apply softmax
viii. Return category with highest probability

## IV. Implementation setup

The proposed mental health classification system was implemented in Python 3.8 using key libraries such as Hugging Face Transformers for fine-tuning the bert-base-uncased model, and PyTorch as the core deep learning framework. Supporting tools included Scikit-learn, Pandas, NumPy, and Matplotlib for data handling, evaluation, and visualization. A user-friendly Streamlit GUI was developed to enable real-time predictions based on user-input text.

The primary dataset used was the DAIC-WOZ dataset, which contains annotated interview transcripts indicating mental health conditions like depression and anxiety. To improve class balance and diversity, data augmentation techniques (e.g., synonym replacement, back-translation) were applied. Additional data were incorporated from public sources, including Reddit-based mental health datasets on Kaggle.

The final dataset spanned seven classes: Normal, Anxiety, Suicidal Thoughts, Bipolar Disorder, Personality Disorder, and Stress. Model training involved fine-tuning BERT with hyperparameters such as the AdamW optimizer, learning rate scheduling, gradient clipping, and a cross-entropy loss function. Text preprocessing included tokenization and input formatting using special tokens ([CLS], [SEP]). The model was trained for four epochs with batch sizes of 16 (training) and 32 (evaluation), and predictions were made using softmax over the seven classes.

Evaluation showed a strong overall accuracy of 93%, with high precision and recall across all categories. The Streamlit interface enables real-world usability by providing instant predictions and confidence scores for any user-input text.

## V. Results and discussion

This research demonstrates the feasibility of using a fine-tuned BERT model for multi-class mental health classification with real-time analysis through a Streamlit GUI. The system achieves high accuracy and offers an intuitive interface for broader adoption in clinical and personal wellness applications. This model was evaluated on the labeled mental health dataset, and various classification metrics were computed to assess its accuracy. These metrics include accuracy, precision, recall, F1-score, and a confusion matrix to visualize the model's performance across different categories.

The confusion matrix for the fine-tuned BERT model is shown in table 1. It illustrates how well the model classifies each mental health condition and helps identify misclassifications across categories.

DOI: https://doi.org/10.71058/jodac.v9i5018

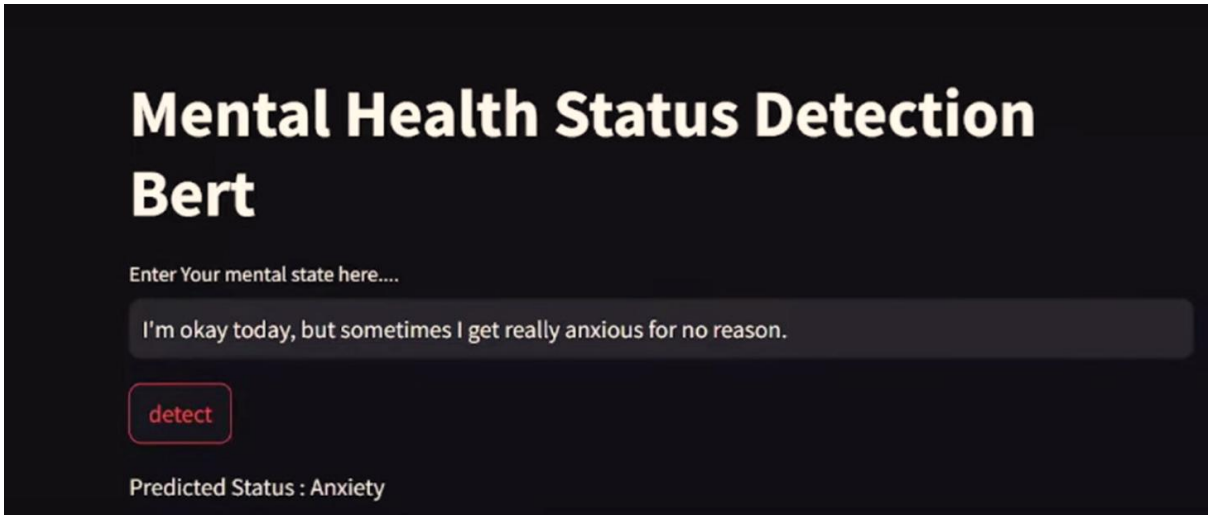


**Figure 1:** Fine-tuned BERT model

**Table1:** Confusion matrix for the fine-tuned BERT model

| Category | Anxiety | Bipolar | Depression | Normal | PD | Stress | Suicidal |
|---|---|---|---|---|---|---|---|
| Anxiety | 359 | 0 | 4 | 0 | 0 | 3 | 0 |
| Bipolar | 0 | 389 | 2 | 2 | 0 | 0 | 0 |
| Normal | 5 | 3 | 7 | 331 | 0 | 11 | 27 |
| Personality disorder (PD) | 0 | 0 | 0 | 0 | 393 | 0 | 0 |
| Stress | 0 | 0 | 0 | 0 | 2 | 419 | 0 |
| Suicidal | 0 | 0 | 23 | 7 | 0 | 0 | 358 |

Based on the confusion matrix table 2 presents the classification metrics achieved by the fine-tuned BERT model across different mental health categories and it is rerpesented in figure 1.

**Table 2:** Fine-tuned BERT model across different mental health categories

| Category | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Anxiety | 0.97 | 0.98 | 0.97 | 366 |
| Bipolar | 0.98 | 0.99 | 0.98 | 393 |
| Normal | 0.94 | 0.88 | 0.91 | 384 |
| Personality Disorder | 0.99 | 1.00 | 0.99 | 393 |
| Stress | 0.96 | 0.99 | 0.97 | 421 |
| Suicidal Thoughts | 0.93 | 0.91 | 0.92 | 388 |
| Overall Accuracy | 0.93 | N/A | N/A | 2690 |

DOI: https://doi.org/10.71058/jodac.v9i5018

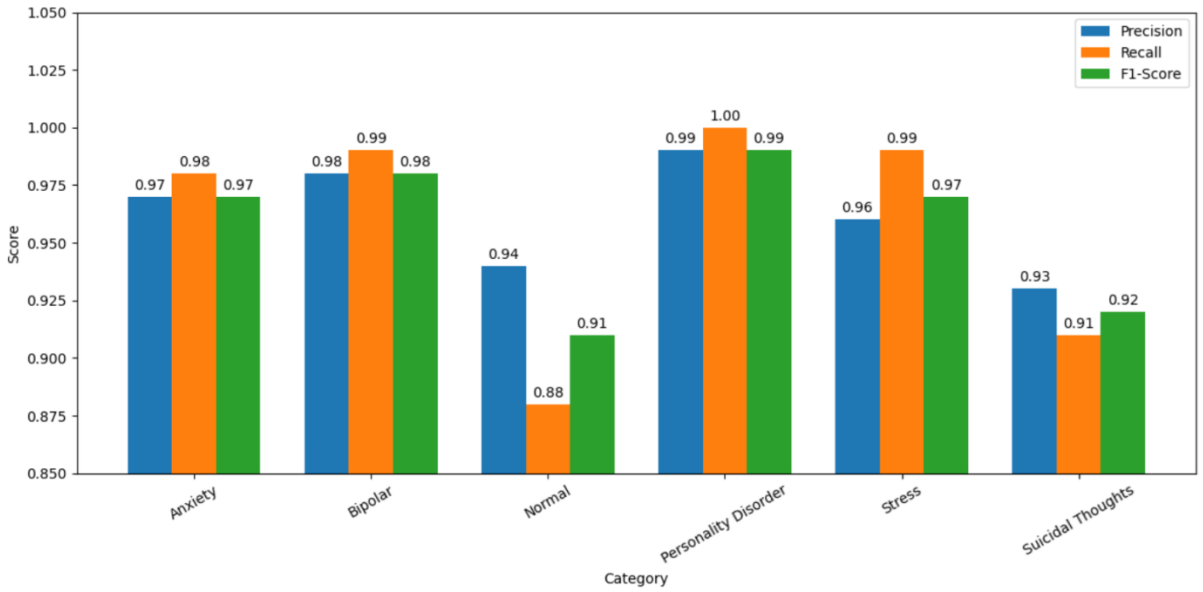


**Figure 1:** Precision, Recall, and F1-Score by Category

The overall perforance of model is demonstrated with a high overall accuracy of 93%, with Anxiety and Suicidal thoughts achieving particularly high F1-scores of 0.97 and 0.92, respectively.The model was most accurate in identifying Personality Disorder, Bipolar, and Stress categories, each with precision and recall above 95%.

## Discussion

The performance of the fine-tuned BERT model on the labeled mental health dataset demonstrates promising results, especially in the context of a highly sensitive and complex task such as mental health classification. The model achieved an overall accuracy of 93%, indicating robust general performance across the different mental health categories. The evaluation employed standard classification metrics precision, recall, F1-score, and a confusion matrix to provide a comprehensive understanding of the model's strengths and limitations.

The model exhibited excellent classification performance in detecting Personality Disorder, Bipolar, and Stress conditions. For each of these categories, both precision and recall exceeded 95%, resulting in F1-scores of 0.98 or higher. This indicates that the model not only accurately identifies these categories (high precision) but also consistently detects most of the true cases (high recall). These high scores suggest that the linguistic patterns associated with these categories may be more distinct, allowing the model to differentiate them with higher confidence.

Furthermore, Anxiety and Suicidal Thoughts also achieved strong F1-scores of 0.97 and 0.92, respectively. In the case of anxiety only a few instances (4 labeled as depression, 3 as stress), resulting in a precision of 0.97 and recall of 0.98. Suicidal thoughts also showed high overall performance, though with slightly more variability due to its overlap with depression.

## V. Conclusion and Future Scope

In this research, we developed an intelligent machine learning system capable of detecting potential mental health conditions—such as anxiety, stress, suicidal ideation, or a normal state based solely on written text. Through meticulous data preprocessing and addressing class imbalance with oversampling techniques, we ensured that the model could learn meaningful patterns across all mental health categories. By fine-tuning a pre-trained BERT model from the Hugging Face Transformers library on our curated dataset, we leveraged state-of-the-art natural language understanding to accurately interpret emotional cues in textual input.

DOI: https://doi.org/10.71058/jodac.v9i5018

Our evaluation demonstrated that the model performs well in identifying various forms of emotional distress, making it a promising tool for early detection and intervention in mental health contexts. The trained model, along with the accompanying prediction function, lays the groundwork for real-world applications such as chatbots, online mental health screening tools, or integration into digital health platforms. This work highlights the potential of combining advanced language models with accessible tools like Hugging Face to address critical challenges in mental health through automated text analysis. Future work may explore multimodal approaches and field testing in clinical environments.

## Acknowledgement

I would like to express my sincere gratitude to my supervisor, Dr. Syed Wajahat Abbas Rizvi, for his guidance and support throughout this research. His expertise and encouragement have been invaluable to the completion of this work. I also wish to extend my appreciation to Dr. P.K. Dwivedi, Dr. Ashish Kumar Pandey, and Dr. Sandhya, whose insights and feedback have greatly contributed to the improvement of this study. Special thanks to Amity University, Lucknow, and Dr. Rammanohar Lohia Avadh University, Ayodhya, for providing the necessary resources and assistance during the course of this research. Lastly, I am deeply grateful for the unwavering support of my family and friends, whose encouragement has been a constant source of strength.

## Author Contributions

All authors have equally contributed.

## Conflict of Interest

The authors declare that there is no conflict of interest regarding the study of this article.

## Funding

Not applicable.

## References

[1] World Health Organization, “Mental health: Strengthening our response”, *World Health Organization*, **(2021)**. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

[2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed., Arlington, VA: American Psychiatric Publishing, **(2013)**.

[3] S. Chancellor and M. De Choudhury, “Methods in predictive techniques for mental health status on social media: A critical review”, *NPJ Digit. Med.*, vol. 3, no. 1, **(2020)**, p. 43.

[4] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding”, in *Proc. NAACL-HLT*, **(2019)**, pp. 4171–4186.

[5] S. Ji, C. P. Yu, S. F. Fung, S. Pan and G. Long, “MentalBERT: Publicly available pre-trained language models for mental healthcare”, in *Proc. 60th Annu. Meet. Assoc. Comput. Linguistics (ACL)*, **(2022)** .

[6] A. Murarka and T. Mandl, “MentalRoBERTa: A transfer learning approach for mental health classification from social media”, in *Proc. CLEF Conf.*, **(2023)**.

[7] A. H. Orabi, P. Buddhitha, M. H. Orabi and D. Inkpen, “Deep learning for depression detection of Twitter users”, in *Proc. 5th Workshop Comput. Linguist. Clin. Psychol.: From Keyboard to Clinic*, **(2018)** , pp. 88–97.

[8] T. Al Hanai, M. M. Ghassemi and J. R. Glass, “Detecting depression with audio/text sequence modeling of interviews”, in *Interspeech*, **(2018)**, pp. 1716–1720.

[9] A. Pourkeyvan, M. Soleymani and G. Mohammadi, “Mental health multi-label classification with Hugging Face Transformers”, *J. Affect. Disord. Rep.*, vol. 12, **(2023)**, p. 100459.

[10] Y. Liu and Y. Su, “Multi-task learning for mental health prediction using BERT”, *IEEE J. Biomed. Health Inform.*, vol. 26, no. 3, **(2022)**, pp. 1012–1023.

[11] S. Yang, Y. Zhang and R. Xu, “MentaLLaMA: Explainable social media analysis using large language models for mental health monitoring”, in *Proc. EMNLP*, **(2023)**.

[12] Q. Li, M. Zhang, Y. Wang and Y. Liu, “Emotion-based user profiling for mental health detection on Weibo”, *Inf. Process. Manag.*, vol. 58, no. 5, **(2021)**, p. 102686.

DOI: https://doi.org/10.71058/jodac.v9i5018

[13] R. Skaik and D. Inkpen, "Predicting depression and emotions in social media posts using multimodal and ensemble approaches", in *Proc. 12th Lang. Resour. Eval. Conf. (LREC)*, **(2020)**.
[14] M. Aldarwish and H. F. Ahmad, "Ensemble BERT for early detection of depression from social media posts", in *Proc. 2021 Int. Conf. Data Min. Workshops (ICDMW)*, **(2021)**, pp. 1–8.
[15] H. He, X. Li and Y. Zhu, "MGL-CNN: Multigranular graph-based CNN for mental health detection on social media", *ACM Trans. Web (TWEB)*, vol. 16, no. 2, **(2022)**, pp. 1–20.
[16] D. Beniwal and M. Saraswat, "Multi-modal depression detection using hybrid BERT-CNN model", *Multimed. Tools Appl.*, vol. 82, **(2023)**, pp. 18901–18923.
[17] Z. Yang, J. Zhang, J. Chen and M. Wang, "Integrating BERT and TabNet for mental health assessment in psychiatric rehabilitation", *IEEE Access*, vol. 11, **(2023)**, pp. 12245–12257.
[18] W. Samek, T. Wiegand and K. R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models", *arXiv preprint arXiv:1708.08296*, **(2017)**.
[19] A. Benton, M. Mitchell and D. Hovy, "Multitask learning for mental health using social media text", in *Proc. EACL*, **(2017)**, pp. 152–162.
[20] T. Nguyen, L. Phan and D. Phung, "Evaluating large language models on comorbid mental health conditions: Challenges and future directions", in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 11, **(2023)**, pp. 12349–12357.