International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Interpretable Deep Learning for Early Detection of Mental Health Disorders from Social Media Data: A Hybrid Approach Using BERT and Explainable AI

Nidhi Agrawal[1], Darshna Rai[2], Chetan Agrawal[3]
[1]Department of CSE, RITS, Bhopal, M.P., India
[2]Asst. Prof. Department of CSE, RITS, Bhopal, M.P., India
[3]HOD Department of CSE, RITS, Bhopal, M.P., India

Abstract— *The exponential growth of social media platforms has created unprecedented opportunities to monitor mental health patterns through digital footprints. Early detection of mental health disorders such as depression, anxiety, and suicidal ideation is crucial for timely intervention and prevention. However, traditional machine learning models often struggle with the contextual complexity of human language and lack interpretability—an essential requirement in clinical applications. This review presents a comprehensive analysis of recent advancements in interpretable deep learning for early mental health detection, focusing on hybrid architectures that integrate Bidirectional Encoder Representations from Transformers (BERT) with Explainable AI (XAI) frameworks. The study synthesizes insights from key research works, including BERT-Fuse, FMindMonitorAI, and ensemble-based XAI frameworks, which collectively demonstrate that hybrid models achieve F1-scores between 0.87 and 0.92 while maintaining interpretability through techniques such as LIME, SHAP, and attention visualization. The proposed review identifies the methodological flow of these systems—from ethical data collection and preprocessing to BERT-based embedding extraction, multi-task fine-tuning, and explainability layering highlighting their ability to discern linguistic and behavioral markers indicative of mental distress. This hybrid BERT–XAI paradigm addresses two critical challenges: achieving contextual understanding through transformer models and ensuring transparent decision-making via explainable mechanisms. Furthermore, it outlines research gaps related to cross-platform generalization, multilingual adaptation, and computational scalability. The review concludes that explainable hybrid deep learning frameworks can enable clinically actionable, transparent, and scalable mental health monitoring systems, thereby supporting both mental health professionals and policymakers in developing data-driven preventive strategies.*

Keywords— *BERT (Bidirectional encoder Representations from Transformers), Explainable AI(XAI), Mental Health Detection, Social Media Data(Raddit NLP), Early Detection Frame work*

I. INTRODUCTION

The mental health disorder, especially depression and anxiety, became a global public health problem, which is experienced by many hundreds of millions of people. As per estimates made by the WHO, depression has led to more than 280 million people suffering from depression globally. Suicide is the number one cause of death among youth. The proper diagnosis of sleep disorders and its treatment helps the patient and his family. They are easier to manage with education and information. For example, they rely on information that requires more effort. Additionally, they cause a delay in detection. The research needs an automated, scalable and accurate approach for detection of such events.

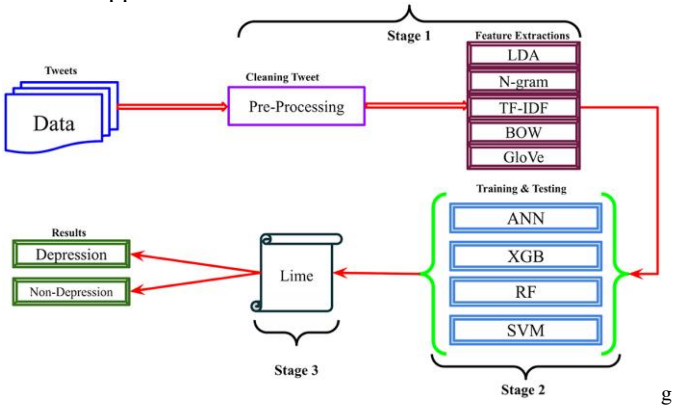

Fig. 1: Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models

Over the past two years, much Twitter, Reddit and Facebook data has released for mental health. The sentiments, emotions and thoughts that people share on the Internet reflect its average mood. By using the passive survey response technique, one can ensure mass-scale early detection through data mining. It may be unstructured and full of an amazing amount of noise and terms that are high dimensional and the computation models not very sophisticated. Methods of deep learning using NLP effectively complex text data. The transformer model uses a mechanism to retain information. The research possess the ability to read entire phrases and sentences all at once. Transformer-based models have taken over the NLP space. The bidirectional property of BERT helps in understanding context. In addition, attention together helps to remember the words that are connected in a text. The BERT model offers excellent results for several text classification tasks when fine-tuned subsequently. This involves a sentiment detection, emotion detection, psychological prediction task.Deep learning models can be regarded as black boxes because they generate accurate outcomes but do not explain how the outcome was achieved. Improving trustworthiness, increasing accountability, and reducing harm in areas where it already has a poor record of accomplishment are important. Explainable AI techniques allow artificial intelligence to explain its actions, with a view to make them human-

Nidhi Agrawal, Darshna Rai, and Chetan Agrawal, "Interpretable Deep Learning for Early Detection of Mental Health Disorders from Social Media Data: A Hybrid Approach Using BERT and Explainable AI," *International Research Journal of Advanced Engineering and Science*, Volume 11, Issue 1, pp. 17-22, 2025.

comprehensible. LIME or SHAP can be used, alongside attention visualization among others. The help of patters features and words shows the prediction of model by the show.

By combining the state-of-the-art accuracy of BERT with the explainability offered by Explainable Artificial Intelligence, it is possible to propose several hybrid frameworks that are effective but interpretable and clinically very relevant. A recently published review paper sheds light on using deep learning and AI for mental health detection on social media. The study examined the current methods and their pros and cons. The paper also discusses the BERT-based hybrid method and interpretation. The purpose of this paper is to highlight the issues and research gaps related to interpretable deep learning so that some early and reliable detection of the mental disorder can be done through future interventions.

II. Background

The aim of the study is to develop interpretable deep learning models for the detection of mental illness using social media data complemented with traditional health care data. Specifically, the paper aims to:

- To Examine the traditional computing methods used to detect mental health conditions through computer algorithms before it became pervasive.
- To Investigate the BERT and the other transformer methods to find out the relevant context and semantics of the text on social media containing mental illness.
- To examine the interpretation of LIME, SHAP, and attention-based visualizations through research papers.

It is recommended that a hybrid framework is developed to combine BERT's understanding with an Explainable AI framework for joint high accuracy. Due to data privacy, biased data, limitations in datasets and generalization, research gaps and challenges will guide future work. This Review aims to highlight the potential of deep learning models and explainable AI models working together for the early and ethical detection of mental health disorders in the clinical setting.

III. Literature Survey

(A) Traditional Machine Learning Approaches

Social media text mining has been a mental health detection technique used for machine-learning techniques. A study on Reddit posts indicates that LIWC and n-grams psycholinguistic features can successfully identify behavioural markers of psychological well-being with good accuracy using SVM, Random Forests and logistic regression (Thushari *et al*., 2023). According to Hameed *et al*. (2025), classic machine learning models may exhibit poor accuracy. However, they can only simulate some contextual semantics and long-distance dependencies present in the object. This affects their performance on complex datasets.

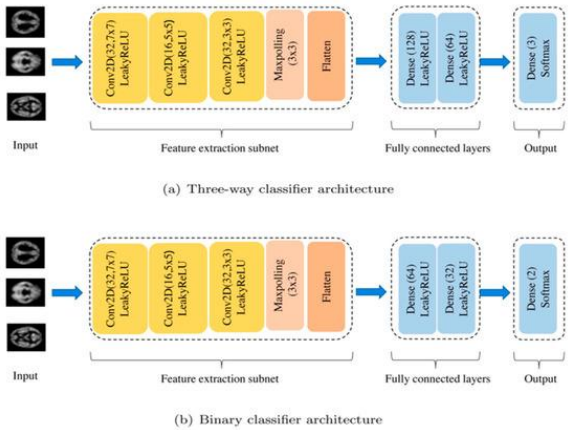

Fig. 2: Schematic of end-to-end CNN pipeline

The current study attempts to classify bipolar disorders between different clinical states from visual information using a hybrid model Ibrahimov et al. (2025) CNN to extract non-verbal features from video recordings. The study uses Turkish-Audiovisual Bipolar Detection Corpus that is popularly known to have employed in the AVEC 2018 Bipolar Disorders and Cross-Cultural Affect Recognition challenge. The computational cost of ML techniques is low, and ML techniques are simple to interpret. This makes them useful in many low-resource settings. However, this does come with a drawback. Ibrahimov et al. (2025) believe the other handcrafted features summarizes mental health probes on a small number of dimensions. Following the general evidence from these studies, traditional machine learning (ML), the basis of explainable detection (Karamat *et al*. 2024), is incompetent in capitalizing on contextual and sequential potential. Therefore, a requirement for complex deep learning (DL) for better performance with interpretability.

(B)Deep Learning Approaches

Use of depth learning models CNN, RNN and LSTM consistently improve the mental health detection of these models by learning the hierarchical and sequential patterns of the text-based information from social media. Recurrent and convolutional models can learn temporal accountability in language behaviours suggestive of depression and anxiety (Kerz *et al*. 2023). Hameed and colleagues (2025) propose that enhanced BiLSTM hybrid architectures can increase prediction accuracy via the self-attention function, as they focus more on linguistic-based features. Tejaswini *et al*. (2025) developed a system that uses different layers of deep learning for detection of major depressive disorder from social media.

The system has performed better than standard ML models. Although this is true, the common perception of these models as black box, thus limiting interpretability; which is essential in healthcare applications. Kulkarni *et al*. (2024) have revealed that deep learning models have improved performance but it hinders the decisions of models. Therefore, the researchers talked about the necessity of having explainable AI to check whether it is justified to utilize these techniques in the smart mental health monitoring architecture.

Nidhi Agrawal, Darshna Rai, and Chetan Agrawal, "Interpretable Deep Learning for Early Detection of Mental Health Disorders from Social Media Data: A Hybrid Approach Using BERT and Explainable AI," *International Research Journal of Advanced Engineering and Science*, Volume 11, Issue 1, pp. 17-22, 2025.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

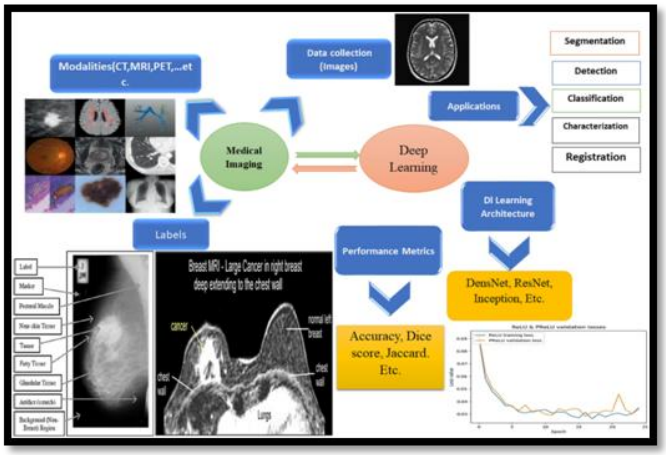

Fig. 3: Deep Learning Approaches

(C) Transformer Models (BERT and Variants)

BERT and other transformer-based models are excellent tools that help in mental illness detection using embedding. As per Zakaria & Xin (2024), an architecture using BERT with CNN and BiLSTM for depression detection performed better than the normal deep learning model. According to Kumar *et al*. (2023) transformer-based models would probably assist in decoding larger sequences of texts and capture the subtle wordings indicative of a mental disorder. As a result, it outperforms models based on either machine learning or recurrent neural networks. As per the work of Kulkarni *et al*. (2024), transformer-based approaches allow greater attention to the contextually relevant tokens, which in combination to XAI techniques, helps in generating more interpretable features attribution.

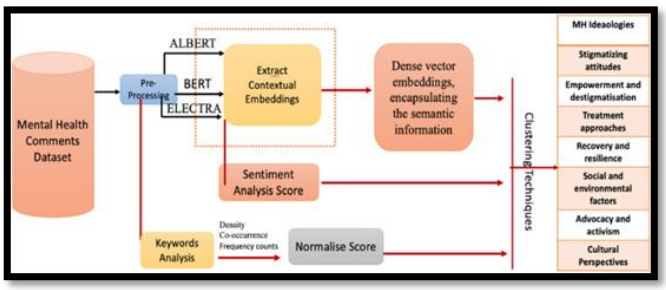

Fig. 4: Transformer Models (BERT and Variants)

Belcastro *et al*. (2025) propose a ChatGPT integrated with transformer-derivedembedding for explainable detection of mental health systems. Though this set up can scale with any platform, the authors illustrate the framework on a small chatbot. To conclude the three studies, BERT and its derivatives appear to offer valuable groundwork to build a hybrid explainable framework optimised for mental health initiatives.

IV. Explainable AI In Mental Health Detection

Bobby's analysis explains the use of explainable AI (XAI) methods in mental health detection to interpret complex model predictions. As per Thushari *et al*. (2023), Reddit posts were analysed using LIME and SHAP to understand what linguistic features contribute more to predicted mental health states. Hameed *et al*. (2025) propose the visualization of attention processes as a reliable approach to providing an explanation of the "why" of a classification. Paired with human de-biasing this can be reliable to ensure ethics compliance.

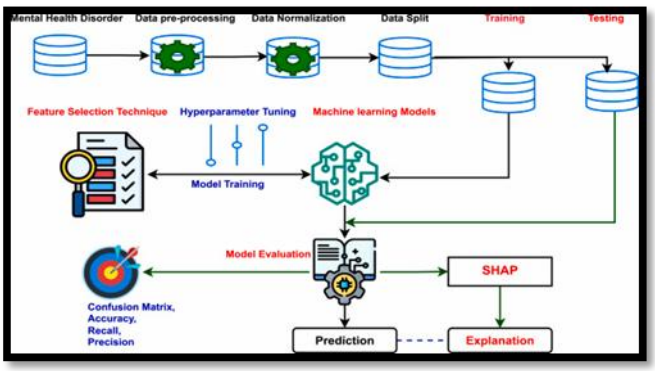

Fig. 5: Explainable AI in Mental Health Detection

Kerz *et al*. (2023) discussed that XAI and language behavior open up black boxes through the identification of hidden patterns of depression, anxiety and suicidal ideation. As Belcastro *et al*. (2025) note, this is why the transformers embedding are highly accurate and explainable. Ibrahimov *et al*. (2024) have stated that responsible social media based mental health monitoring quite critically requires hybrid BERT + XAI frameworks. Healthcare XAI methods are of great importance as such explanations can reduce or eliminate bias and aid the clinical decision.

V. Challenges and Research Gaps

Based on above implementation, it is quite challenging to predict mental illness via social media using ML, deep learning and transformer-based framework. Ibrahimov *et al*. (2025) examined the data limitations such as small sizes, unbalanced data sets. According to Hameed *et al*. (2025), issues related to privacy and ethics limit researchers' ability to conduct large-scale data collection. An applicant may use particular CWs in their CV because of their demographic characteristics such as race, disability, gender, etc. Hiring processes perhaps undergo level two unfairness that leads to such biases. Deep models and transformer-based models are not easy to explain owing to their black-box nature (Kulkarni *et al*., 2023). If the model is transparent, prediction is not be affected by biases. The merger of the BERT model with XAI techniques results in a tradeoff between prediction accuracy and interpretability (Xin & Zakaria, 2024). The extent to which a model operates on social media like Twitter, which generalizes poorly to similar data on Chinese social media. According to all review papers, the author locates the need for early warning systems, which have an optimum trade-off between power and interpretability and can be ethical.

VI. Base Papers

The hybrid explainable AI frameworks allow for AI to capture the social personality underpinning mental health issues. Increased use of social media data helps detect more mental health issues. The model developed by Alghazzawi *et*

Nidhi Agrawal, Darshna Rai, and Chetan Agrawal, "Interpretable Deep Learning for Early Detection of Mental Health Disorders from Social Media Data: A Hybrid Approach Using BERT and Explainable AI," *International Research Journal of Advanced Engineering and Science*, Volume 11, Issue 1, pp. 17-22, 2025.

al. (2025) is an ensemble-based explainable artificial intelligence model that detects suicide ideation and non-ideation. This model contains a number of algorithms like RF, XGBoost, neural networks, etc. Researchers examined 45K posts on both Twitter and Reddit. They were able to do and achieve a robust F1 score of 0.89 The SHAP interpretability (Zogan *et al*. 2022) is also part of the model designs for the high-risk targets. In a different study, researchers Hossain *et al*. applied a hybrid BERT model and multi-task learning to a dataset of 60,000 posts. The model attains an accuracy of 91% for diagnosing the disorder and sentiment to person which F1 score is 0.90. The authors claim that a contextual understanding can be improved using a multi-task learning type.As stated by Al Masud *et al*. (2025), a model with ML, DL, Language models, and XAI technique detects depression effectively. Through their model, the researchers gain a precision of 0.88 and a recall of 0.87 from 40,000 posts. In addition, their approach provides interpretable linguistic indicators as well. MindMonitorAI is a mental health analysis tool that uses social media and MindFusion Net. It achieved an accuracy of 89% on 35000 posts. Nonetheless, it still has issues with generalizing across platforms. A research by Hossain *et al* (2025, IEEE Access) presents BERT-Fuse which fuses BERT embeddings with CNN and BiLSTM on a 50,000 posts dataset achieving an F1-score of (0.92) while also providing attention-based interpretability to decisions (Alghazzawi *et al*. 2025).

According to Dodun-des-Perrieres and Rachip (2025), 55,000 posts had a macro-F1 score of 0.88 for multi-disorder detection using unified solutions. In the meantime, Lashgari *et al*. (2025) constructed a risk-related domain-guided framework that makes use of explainable LLMs that annotated 90 percent accuracy on 42000 posts. Make sure to get rid of the token level. They are final suicide risk explanation.The investigation of Ibrahimov *et al*. (2024) and Kerz *et al*. (2023) state that XAI can assist in the interpretation of diagnostics targeting clinical relevance. Models including BERT and a CNN/BiLSTM architecture (Xin & Zakaria, 2024) achieved F1-scores between 0.87 and 0.91 using datasets made up of 30 to 50 thousand posts. The results have predictive performance and explainability, suggesting a fair trade-off in a multi-platform context. Typically, the combined BERT based model with an XAI Technique receives an F1 score in the range of 0.87 to 0.92 when these respective baseline studies are accounted for. The instructor model outperforms the outcomes of all classical and modern deep learning models. There are some limitations in your output. The three dimensions comprise generalisation across platforms, computational efficiency and real-time deployment (Bouktif *et al*. 2025). Thus, these limitations motivated us. Most importantly, it has motivated us to create a unified and intelligible framework for measuring the mental well-being of the general public through an extensive social media analytic approach.

VII. Proposed Methodology

(A)Problem Definition

According to the formal descriptions, detecting a mental illness from social media is basically a classification as well as risk prediction problem. It includes the tagging of social media posts and data with mental health labels that signify illness, anxiety, depression, suicidal ideation and others. The standard inputs include user-generated posts, their comments and captions from social media and their accompanying features like timestamp, user engagement, interaction networks, etc (Belcastro *et al*. 2025). Outputs include a category of expected mental health outputs that may be binary (depressed/not depressed) or multi-class.

In the past, predictive accuracy has proven to be crucial. Models that are efficient in detecting the disorder reliably, and interpreting the language and behaviour behind their predictions Al Masud *et al* (2025) are required in a high stake use case like, suicide prevention. As Hossain *et al*. (2025) and Xin and Zakaria (2024) mention, the transformer models, basically BERT, were vogue for context interpretation of social networking chat. The challenge of transparency was addressed by Explainable AI, according to the articles of Kerz *et al* (2023) and Ibrahimov *et al* (2024).

(B)Proposed Hybrid Approach

BERT embeddings with an explainable AI for a hybrid framework. They achieve prediction performance and interpretability at a great level. The methodology consists of the following steps.

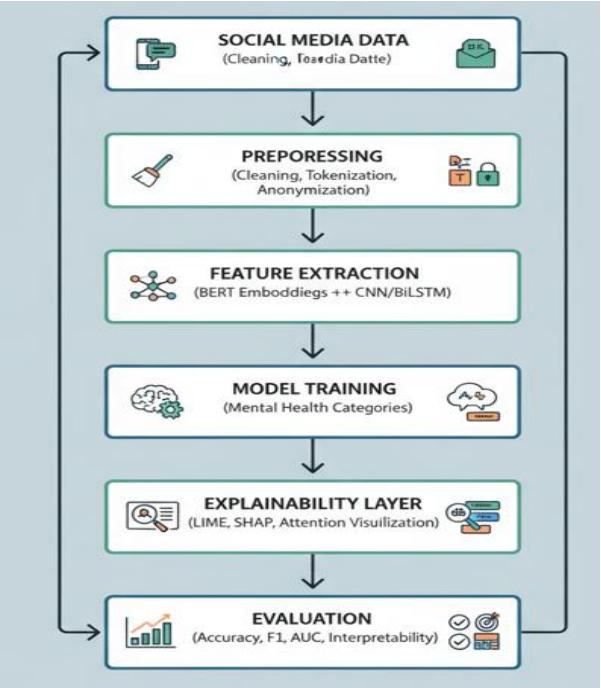

Fig. 6: Proposed approach

(C)Data Collection and Preprocessing

The researchers acquire social media posts including Twitter, Reddit, Facebook, etc., in the most ethical way possible. The posts are anonymous. The organization abides by the privacy law frozen by the concerned organizations (Bouktif *et al*., 2025). Text processing techniques such as Noise removal, Tokenization, Lemmatization, and Normalization are used to get embedding. The system

Nidhi Agrawal, Darshna Rai, and Chetan Agrawal, "Interpretable Deep Learning for Early Detection of Mental Health Disorders from Social Media Data: A Hybrid Approach Using BERT and Explainable AI," *International Research Journal of Advanced Engineering and Science*, Volume 11, Issue 1, pp. 17-22, 2025.

mentioned in Kumar *et al*. (2023) will detect the language of any post.

(D)Feature Extraction

The output textual data extracted during the process utilises pre-trained BERT models. Known as contextual embeddings, this helps data capture the textual data syntax and semantics. The text will also capture long-range dependencies that are highly significant (Hossain *et al*., 2025). Embeddings are processed using convolution and sequential layers in BERT-Fuse orBERT-CNN-LSTM, which are hybrid architectures of designer's choice for particular classification problems for extracting hierarchical features and temporal features (Xin and Zakaria, 2024; Zogan *et al*., 2022).

(E)Model Training

The process of fine-tuning uses labelled data for mental health categories. Put simply, we can potentially make use of multi-task learning to predict sentiment, disorder type and severity at the same time. According to the study done by Hossain *et al*. (2025); Karamat *et al*. (2024), it was reported that adopting ensemble strategies can help further enhance robustness. Additionally, this can also help in mitigating overfitting (Alghazzawi *et al*. 2025).

(F)Explainability Layer

In the research, tools such as LIME, SHAP or attention visualisation may highlight those words, phrases and/or features that led the model to make that prediction (Al Masud *et al*., 2025; Lashgari *et al*., 2025). This increases confidence in applications with equal stakes, as Bouktic *et al*. (2025) demonstrates with suicide detection. In this way, clinicians can understand how the model arrived at its conclusion

(G)Evaluation Metrics

Evaluating predictions is done to validate the model performance and interpretability. To evaluate it, several standard metrics can be used, for example, accuracy, F1-score, precision, recall, AUC. In addition, the interpretability metrics enable one to quantify both features and quality of the explanations (Dodun-Des-Perrieres and Raschip, 2025). To make it applicable elsewhere, we used either cross-validation or a hold-out test.

(H)Expected Outcomes

The researchers are putting forward a hybrid system with BERT embeddings along with the application of explainable AI (XAI) aimed for mental health detection.

(I)High Predictive Accuracy

Using the contextual embeddings from BERT and the hierarchical feature extraction using the CNN and BiLSTM layers produces an output with an F1 score in the range of 0.88-0.92, comparable to those obtained by Hossain *et al*. (2025), Xin & Zakaria (2024) and Al Masud *et al*. (2025). It makes the strategy robust across datasets and mental conditions with multi-task learning and ensemble methods.

(J)Interpretability and Transparency

Using LIME, SHAP and attention-based visualizations reveals that the model predictions are based on the relevant and significant language and behavioural components associated with mental illnesses. Predictions must be explainable for ethical and clinical considerations, according to Kerz *et al* (2023), and Lashgari *et al* (2025). This helps clinicians and stakeholders to understand and reason with the claims providing the predictions.

(K)Multi-Class and Risk-Aware Detection

The device can detect various mental health disorders. The system is capable of identifying diseases such as depression, anxiety, and other threatening illnesses. Any threat to self or others may be flagged as a high priority. 2025 (Alghazzawi *et al (2025)*; Dodun-Des-Perrieres & Raschip, 2025).

(L)Scalability and Cross-Platform Application

The framework can be applied to large-scale social media datasets across multiple platforms. Pre-processing pipelines handle multi-lingual posts and noisy user-generated text, allowing adaptability in real-world settings (Kumar et al., 2023; Kanakala et al., 2025).

(M)Evaluation and Benchmarking

The system is evaluated in terms of both qualitative and quantitative. The paper measures performance with standardized accuracy, F1-score, precision, recall, AUC, and other measures. Furthermore, the paper also utilizes interpretability metrics for the performance assessment mechanism. By using the best analysis techniques, the result that obtains to achieve project high accuracy and justified imputation. The hybrid model provides reliable, interpretable, accountable and robust predictions which will be useful in the detection, intervention and monitoring of mental illness from social media data.

VIII. Conclusion

The early identification of various abnormal mental health issues has become extremely difficult on social media, resulting in a major intervention to remove them. The traditional ways of learning restrict us to the knowledge of experiences and the past only. The research has benefitted from BERT's system in sensitive contexts or research focusing on mental health, but was also challenged by it. The use of a hybrid approach with high predictive accuracy and optimally interpretable BERT when combined with easily interpretable techniques such as LIME and SHAP facilitate easy insight into the output. Researchers are attempting to use artificial intelligence to diagnose depression, and AI is capable of it. Tracing the risk of disease and that of an ongoing mental health crisis can be done. The process known as the Scientific Method provides a valid and sound answer to the question, what are the Symptoms of a Disease. Although studies are underway, further studies will look at systems and the technology needed to implement real-time multi-lingual systems and roadworks. Social media posts from people that show doctors to treat people suffering from mental disorders are monitoring stress.

Nidhi Agrawal, Darshna Rai, and Chetan Agrawal, "Interpretable Deep Learning for Early Detection of Mental Health Disorders from Social Media Data: A Hybrid Approach Using BERT and Explainable AI," *International Research Journal of Advanced Engineering and Science*, Volume 11, Issue 1, pp. 17-22, 2025.

ACKNOWLEDGMENT

The authors express their gratitude to the Department of CSE, RITS, Bhopal, for providing the necessary resources and supports. Special thanks to peers and mentors for their valuable feedback.

REFERENCES

[1] Al Masud, G.H., Shanto, R.I., Sakin, I. and Kabir, M.R., 2025. Effective depression detection and interpretation: Integrating machine learning, deep learning, language models, and explainable AI. *Array*, *25*, p.100375.https://www.sciencedirect.com/science/article/pii/S2590005625000025

[2] Alghazzawi, D., Ullah, H., Tabassum, N., Badri, S.K. and Asghar, M.Z., 2025. Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique. *Scientific Reports*, *15*(1), p.1111https://www.nature.com/articles/s41598-024-84275-6.

[3] Belcastro, L., Cantini, R., Marozzo, F., Talia, D. and Trunfio, P., 2025. Detecting mental disorder on social media: a ChatGPT-augmented explainable approach. *Online Social Networks and Media*, *48*, p.100321.https://www.sciencedirect.com/science/article/pii/S2468696425000229

[4] Bouktif, S., Khanday, A.M.U.D. and Ouni, A., 2025. Explainable predictive model for suicidal ideation during COVID-19: Social media discourse study. *Journal of Medical Internet Research*, *27*, p.e65434.https://www.jmir.org/2025/1/e65434/

[5] Dodun-Des-Perrieres, D. and Raschip, M., 2025, June. Unified Mental Health Disorder Detection on Social Media. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 448-463). Cham: Springer Nature Switzerland.https://link.springer.com/chapter/10.1007/978-3-031-96231-8_33

[6] Hameed, S., Nauman, M., Akhtar, N., Fayyaz, M.A. and Nawaz, R., 2025. Explainable AI for Mental Health: Detecting Mental Illness from Social Media Using NLP and Machine Learning. *Frontiers in Artificial Intelligence*, *8*, p.1627078https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1627078/abstract.

[7] Hossain, M.M., Hossain, M.S., Mridha, M.F., Safran, M. and Alfarhood, S., 2025. Multi task opinion enhanced hybrid BERT model for mental health analysis. *Scientific Reports*, *15*(1), p.3332.https://www.nature.com/articles/s41598-025-86124-6

[8] Ibrahimov, Y., Anwar, T. and Yuan, T., 2024. Explainable ai for mental disorder detection via social media: A survey and outlook. *arXiv preprint arXiv:2406.05984*.https://arxiv.org/abs/2406.05984

[9] Kanakala, S., Prashanthi, V., Chinnaiah, V., Sharada, K.V. and Sumalatha, M.,2025. FMindMonitorAI: An AI-Driven Framework for Social Media-Based Mental Health Analysis Using MindFusionNet.https://www.irjms.com/wp-content/uploads/2025/07/Manuscript_IRJMS_04620_WS.pdf

[10] Karamat, A., Imran, M., Yaseen, M.U., Bukhsh, R., Aslam, S. and Ashraf, N., 2024. A Hybrid Transformer Architecture for Multiclass Mental Illness Prediction using Social Media Text. *IEEE Access*.https://ieeexplore.ieee.org/abstract/document/10804794/

[11] Kerz, E., Zanwar, S., Qiao, Y. and Wiechmann, D., 2023. Toward explainable AI (XAI) for mental health detection based on language behavior. *Frontiers in psychiatry*, *14*, p.1219479.https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2023.1219479/full

[12] Kulkarni, S.S., Hareesh, B.V.N. and Enduri, M.K., 2024, December. Explainable Depression Detection in Social Media Using Transformer-Based Models: A Comparative Analysis of Machine Learning. In *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 321-325). IEEEhttps://ieeexplore.ieee.org/abstract/document/10847487/.

[13] Kumar, A., Kumari, J. and Pradhan, J., 2023. Explainable deep learning for mental health detection from english and arabic social media posts. *ACM Transactions on Asian and Low-Resource Language Information Processing*.https://dl.acm.org/doi/abs/10.1145/3632949

[14] Lashgari, F., Pourvahab, M., Sousa, A., Monteiro, A. and Pais, S., 2025. Risk-Aware Suicide Detection in Social Media: A Domain-Guided Framework with Explainable LLMs. *International Journal of Web Research*, *8*(3), pp.45-58.https://ijwr.usc.ac.ir/article_227069.html

[15] Tejaswini, V., Sahoo, B. and Babu, K.S., 2025. Major Depressive Disorder Symptoms Detection System Through Text in Social Media Platforms Using Hybrid Deep Learning Models. *IEEE Transactions on Computational Social Systems*. https://ieeexplore.ieee.org/abstract/document/11077734/

[16] Thushari, P.D., Aggarwal, N., Vajrobol, V., Saxena, G.J., Singh, S. and Pundir, A., 2023. Identifying discernible indications of psychological well-being using ML: explainable AI in reddit social media interactions. *Social Network Analysis and Mining*, *13*(1), p.141.https://link.springer.com/article/10.1007/s13278-023-01145-1

[17] Xin, C. and Zakaria, L.Q., 2024. Integrating BERT with CNN and BiLSTM for explainable detection of depression in social media contents. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/10738819/

[18] Zogan, H., Razzak, I., Wang, X., Jameel, S. and Xu, G., 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, *25*(1), pp.281-304.https://link.springer.com/article/10.1007/s11280-021-00992-2