International Journal of Computer Sciences and Engineering
Vol.13, Issue.11, pp.90-98, November 2025
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

IJCSE
ISSN: 2347-2693 (E)

Survey Article

Detection of Mental Health Disorder from Social Media Data

Nidhi Agrawal[1*], Darshna Rai[2], Chetan Agrawal[3]

[1,2,3]Dept. of CSE, RITS, Bhopal, M.P., India

Corresponding Author:

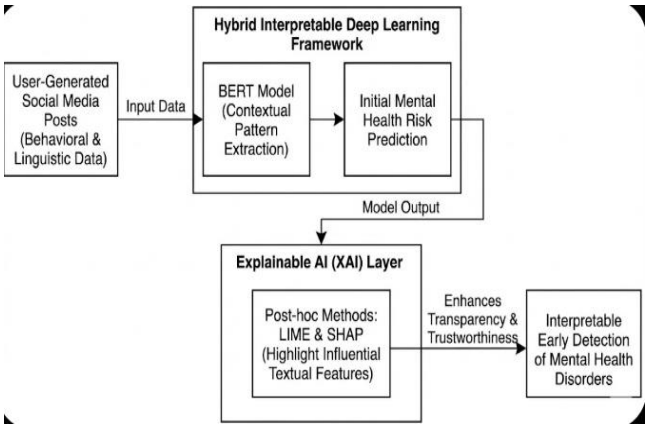
Received: 27/Sept/2025; **Accepted**: 29/Oct/2025; **Published**: 30/Nov/2025. **DOI**: https://doi.org/10.26438/ijcse/v13i11.9098

Abstract: The growing prevalence of mental health disorders has underscored the urgent need for early and scalable detection methods. Social media platforms provide a rich source of behavioral and linguistic data that can serve as indicators of mental health status. This study proposes an interpretable deep learning framework for the early detection of mental health disorders using a hybrid approach that integrates Bidirectional Encoder Representations from Transformers (BERT) with Explainable Artificial Intelligence (XAI) techniques. The model leverages BERT's contextual language understanding capabilities to extract nuanced emotional, cognitive, and linguistic patterns from user-generated social media posts. To enhance transparency and trustworthiness, post-hoc explainability methods such as LIME and SHAP are applied to interpret the model's predictions and highlight influential textual features contributing to mental health risk assessment. Experimental evaluations on benchmark datasets demonstrate that the proposed hybrid model achieves superior accuracy and interpretability compared to traditional deep learning baselines. The findings suggest that combining transformer-based models with explainable AI can provide reliable and ethically responsible tools for early intervention and mental health monitoring in digital environments.

Keywords: BERT, Explainable AI, Interpretable Deep Learning, Mental Health Detection, Social Media Analytics

Graphical abstract

The graphical abstract shows a Hybrid Interpretable Deep Learning Framework for early and interpretable detection of mental health disorders using social media data. As Input Data, we use User-Generated Social Media Posts containing behavioral and linguistic data. The data is then inputted into the hybrid interpretable deep learning framework. At the center of the structure is a BERT Model that performs Contextual Pattern Extraction on the linguistic features. Following that, the framework develops an Initial Mental Health Risk Prediction, which serves as the Model Output. In order to provide transparency and trustworthiness, the Model Output is handed over to an Explainable AI (XAI) Layer. Essentially, this layer uses post-hoc methods like LIME and SHAP to highlight important textual features. The result of the application of XAI layer is that there is Enhancement of Transparency & Trustworthiness, which leads to Interpretable Early Detection of Mental Health Disorders.



1. Introduction

A World Health Organization article noted that most people suffer from mental health disorders over the course of their lives. Therefore, it is likely to be very important to detect MHDS swiftly and efficiently. In the earlier days, diagnostic techniques used to visit the clinic infrequently. Thus, these methods are often subjective and costly. They also activate the person's condition much too late to help. We urgently need real-time monitoring solutions [1].

1.1 Problem Statement

Websites like Reddit and X have become repositories of the real-time behavior and language of people. Thus, they are suitable for early detection of MHD. Deep learning has transformed the accuracy of predictions. BERT refers to a model developed by Google. This language model is in high demand and is very complex. But, these DL models are "black boxes". Because we don't understand how these algorithms work, they cannot have a meaningful application. For example, a model prediction in health services could lead to a diagnosis/intervention. A prognosis can only help when it is trusted – and when doctors trust a prediction, it is only because they understand why. A big challenge is posed due to the dilemma between accuracy and interpretability of predictions.

1.2 Paper Contribution and Structure

The article has worked towards value addition by resolving on the hybrid BERT-XAI model on which to develop intuitive early BID mental illnesses under social media data. The classical deep learning models are black-boxes where the assumption will be made that the individual aspect of the deep learning model helps bring a very small contribution to the overall performance. There, the main objective is to model the target phenomenon (to the highest accuracy) with the latter objective of providing justifications as to why each member of a deep learning model is important to the overall performance. Our framework, in the given scenario, focuses on the forecast of the target phenomenon as well as the causes of its occurrence. Integrating the contextual embeddings of BERT together with scrubbing means, linguistic and behavioral analgeses that liberate the anticipations are indicated by the framework, through the tooling of LIME, SHAP or Integrated Gradients. The contribution is, not only can hybrid fusion models work better than unimodal baselines, but also can provide clear rationale that can be employed in order to make a clinical validation. This two-fold concentration on functionality and reliability encourages the realm of decipherable deep learning to digital psychological wellbeing.

1.3 Mental Health Detection from Social Media

The primary objective of deploying AI for social media analysis is the early identification of mental health diseases. Detecting the behavioural and linguistic signals of risk or distress in a user pre-illness / pre-diagnosis for intervening early for the user (early intervention). It can improve the outcome. AI provides many help in doing this job whereby it can do reams of data crunching in real time to discover the subtle signals that a human being will not be able to notice. Use of this technique shows some promise to give a significant lead time in identification. The PDF on paper number 2 demonstrates that AI can detect individuals with mental health problems, on average, 7.2 days before the experts identify this anomaly [1].

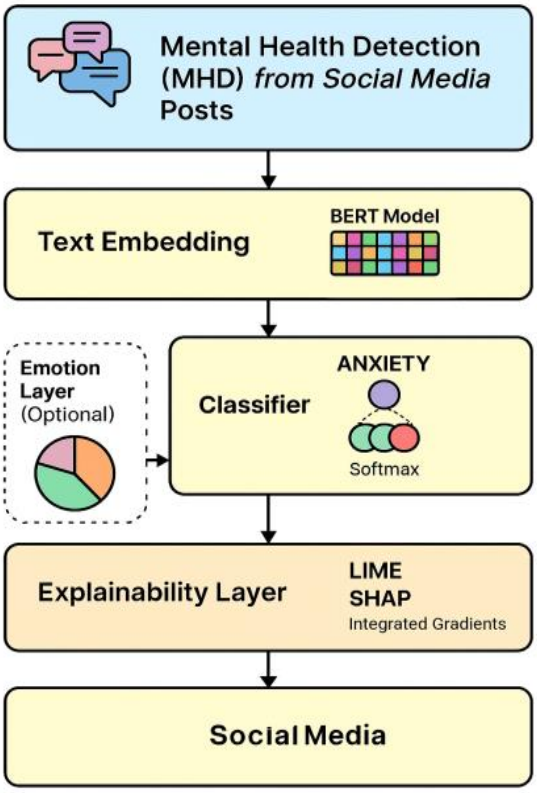


Figure 1. Mental Health Detection from Social Media

To effectively detect the user, the user's entire digital footprint needs to be analyzed. There are several ways to achieve this such as the digital capture of user events. Initial techniques utilized by text-based mining only approached the text, but today's techniques are more complex.

- The words, message and tone of a user's first post and comments on the platform are known as textual modality. We need to use advanced models like BERT here as contextual embeddings are generated. This allows the model to notice small changes in language, which tend to be associated with states of mind of particular individuals. For example, researchers demonstrated a BERT and Bi-LSTM pipeline to detect symptoms of depression on social media posts in various languages [2].
- "Quantifying a user's activity on digital platforms (digital phenotyping) is what behavioral modality does." The users' posting frequency, the time-of-day when they post most frequently and whether there are changes to their overall usage of the platform can give away a change to their daily routine or well-being [3].
- Social Graph Modality looks at the user interactions and the structure of the network. Some things that characterize good quality content are the nature of the comments that their post(s) get, the type of content they share and their density of links. Models that are fusing a MindFusionNet framework, which issues several types to get past the issues, imposed by one type use only Textual, Behavioral, and Social [4].

2. The Importance of BERT in NLP for MHD with Deep Learning

Deep Learning (DL) has significantly enhanced Natural Language Processing (NLP) techniques for detecting Mental Health Disorders (MHD). The major breakthrough in this area is the utilization of the Transformer architecture wherein state-of-the-art method of BERT (Bidirectional Encoder Representations from Transformers) for textual feature extraction.

BERT can create sophisticated and highly contextual word embedding's. BERT uses your entire sentence to create word vectors, reading the sentence bidirectional. This means that instead of just reading a sentence only from the left, and using only left context to create embedding's as done by other NLP models, BERT also uses right context to create embedding's, making BERT context-aware [5]. The model is able to understand the context and connection of the whole meaning. For example, we correctly separate meaning in a depressed context (i.e., I have a crushing feeling) from one in the physical context (I am crushing a can).

In the field of mental health detection, which often has users communicating distress through linguistic nuance, metaphor, or sarcasm, we need this understanding to achieve accuracy. BERT performs better than previous architectures for feature extraction, often being the base of many high-performing systems. To illustrate, a BERT and Bi-LSTM pipeline was created to effectively detect signs of depression from a variety of texts, including more difficult Arabic social media posts. The black-box nature of high-performing systems requires the melding of Explainable AI (XAI) to allow trust and clinical utility [6].

2.1 Explainable AI (XAI) Concepts:

Using strong Deep Learning (DL) models like BERT to detect mental conditions is problematic because these models are "black-boxes". A health system must not ever leave any doubt about what it thinks will happen. The purpose of XAI is to provide humans with the knowledge, understanding and insight to interpret.

2.2 Definition and Purpose.

- XAI helps in creation of trust and transparency while being clinically relevant. The clinician need to check that the decision is based on the right medical determinants. For example, it might not be the result of random correlation. Likewise, it must not arise from extraneous factors such as location of the user or irrelevant words. XAI allows auditing error detection and verification of compliance with ethical standards by making the model's reasoning clear [7].

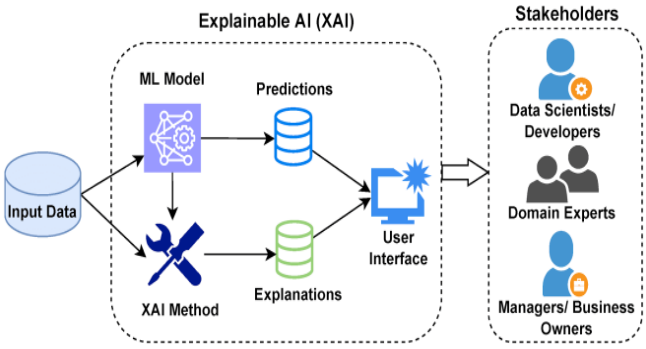


Figure 2. Explainable AI (XAI) Flow Diagram

Categorization of Interpretability.

- Local interpretable techniques only explicate one forecast. The original post explains that the classification was triggered by the exact words, phrases, or behaviours in that user's post. This explains why the user was flagged as at-risk. Clinicians need local interpretable models if they are to decide whether to pursue a certain case.
- Global interpretability is the extent to which a model's overall decisions can be explained. The results provide an answer to the question: "What language features are associated with depression across any user? This gives us a broad understanding of how the model learns. For example, how often does it happen when that word has negative emotions or is self-referential? It can help researchers and system developers check whether the model is learning concepts that are psychologically meaningful.

Model-Agnostic vs. Model-Specific Techniques.
Another way to classify explanations is according to their dependence on the underlying model itself.

- Any black-box model can perform these actions. You don't care about BERT, hybrid, simple networks, etc. They analyze the model only to the inputs and the outputs. Complex BERT systems can be interpreted with LIME and SHAP, two interpretation methods mostly used in this domain [8].
- Targeted Crafting: These methods are tailored for a particular type of model. In the case of BERT, interpretability occurs when the penultimate layer uses internal attention to show which words were paid attention to, in generating final prediction[9]. In this case, attention was paid to "To"and "United" for predicting "States. These often provide deeper insight but are less flexible.

3. Hybrid Architectures: BERT and Fusion Models

Social media user-generated content can facilitate the detection of mental disorders, expert suggests. Words used in a social media post are often informal, contextual, and noisy. To conclude, the models must treat cues as signals, which must also contain linguistic features, and psychological signals [18]. The BERT model and other transformer models

that excel in contextual semantics are black boxes. While it is beneficial for a lot of applications, this shortcoming renders it less suitable for sensitive applications. In recent years, new BERT architectures have made significant strides. They combine BERT with other hugging modules. The modules are statistical, temporal, multimodal, and drive explainability. [16] [20] argue that prediction ability and transparency are trade-offing processes. The textual input from BERT is supplemented with visual input and domain-specific knowledge from other AI models via a hybrid fusion architectures.

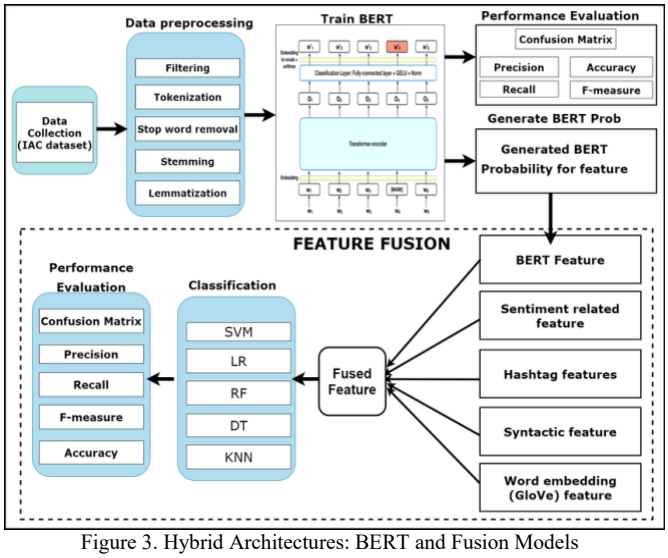


Figure 3. Hybrid Architectures: BERT and Fusion Models

[24] Received an F1-score of more than 92% while detecting postgraduate psychological stress from social media contents by combining the use of BERT embeddings and LDA topic features and sequence modelling with BiLSTM-CRF. The presented model proposes a feature-level fusion approach [24]. In this approach, contextual embedding is concatenate with topic and sentiment indicators for the classifier. A design understanding of peer support for mental health through local semantics and global themes. [20] Built a psychological-knowledge-enhanced topic arms BERT model for crisis text detection. The BERT model achieves effective domain generalization through retrained language representation and domain ontologies.

A new concept called adaptive fusion including attention gates, cross-modal transformers, dynamic weighting layers etc. is the latest trend in research. Big part of these mechanisms serve to learn which modulations are useful to the actual task. Researchers are able to identify chronic stress from the behaviour of the model [18]. Further, the emotional tone is generated from an acute crisis. This crisis has a better interpretability that comes from the attentive weights. Above all, these weights can be visually represented in the article (Frontiers in Psychiatry 2025). Haque and other individuals later enhances this with MMFformer, an advanced multimodal transform architecture that uses audiovisual and textual signals to assess severity. Hybrid models are able to identify subtle emotional indicators not detectable by unimodal models. Experts advocate for multimodal analysis for early detection. As messages are no longer plain text, they use emoji's or photographs to signal their feelings, which will be useful when safe.

Explainability remains a central pillar of hybrid designs. After BERT, SHAP and LIME and their other explainers or visualizers like rationale extraction, attention visualization etc. became common tools (Swanker & Rathore 2025). The modules demonstrate how the phrase or behaviour affected the model's decision. Clinicians or moderators can range outputs using this. Attention heat maps suggest that words associated with feelings of hopelessness or sleep problems drive predictions of depressive-risk, enhancing clinical trust in the model.

Consequently, quite a few hybrids combine BERT with an Explainable fusion model. To begin, you must note that there are numerous essential characteristics of AIs. AUs must be usable and stable. They must also be predictable, etc. [14]. Another adjustment is dialect and language output. How does a person's ruling norm change; and when cases like these arise, a mind-governing norm is created, the status of which is a mental disorder.
We modify a BERT using MDL to reach interpretable results [31]. By applying a set of rules to the MDL data of Urdu/Roman Urdu using the transform outputs. The effectiveness of early warning systems relies on diversity and culture, which can reinforce the existence of diversity.

Despite significant progress, several challenges persist. A solution for the problem of performance-explainability does not exist yet. It may not be the case that when modules are meant to interpret that this always works since it may lead to noisier results and/or extra costs. One more count affecting the reproduction and ethical utilization asks for an open-source hybrid with provenance datasets user privacy compliant. Predictive model for hybrid systems exist, but there is no evidence (at the clinic) to claim their clinical validation, nor their stability in time [25]. Future work should focus on clinician-in-the-loop frameworks, human-centered metrics, and privacy-preserving fine-tuning, like federated or differential privacy [23]. A combination of applying structured features (for example, the frequency with which one shares and with whom one shares) and contextual embedding's, along with using interpretable attention visualization, promises to pave the way for a new generation of explainable deep-learning systems for digital mental health [25].

In summary, hybrid BERT-fusion structures exhibit an understanding of deep semantics and explainability. The study on construction processes involves organizations and humans [18]. It helps enhance the welfare of workers and organizations. The working environment has a great impact on workers' welfare. The future trustworthy digital mental-health technologies that follow human ethics will use hybrid architectures [25]. The research paper can improve artificial intelligence and other automation systems mainly through ethical awareness, transparency requirements, and explainable AI developments.

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

4.1 Core Algorithm for Text Understanding: BERT

Model BERT (Bidirectional Encoder Representations fromTransformers)
Purpose: Extract deep contextual embeddings from social media posts (e.g., tweets, Reddit comments).

Variant to consider:
- bert-base-uncased (general-purpose)
- mental-bert or psychbert (trained on mental-health-related text)

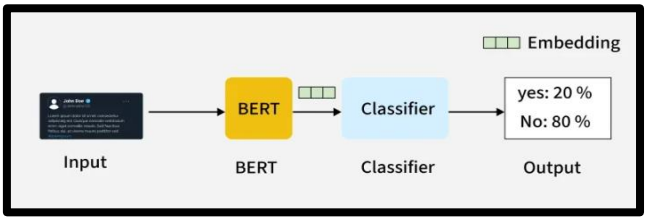


Figure 4. BERT technique

Task: Text classification — predicting likelihood of disorders like depression, anxiety, or PTSD.

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

BERT Score Recall

4.2 Classification Layer (Fine-tuned Deep Network)

After feature extraction by BERT:
- Add fully connected (dense) layers for classification.
- Use Softmax activation for multi-class classification (e.g., anxiety, depression, none).
- Optimization: AdamW optimizer, learning rate ≈ 2e-5, cross-entropy loss.

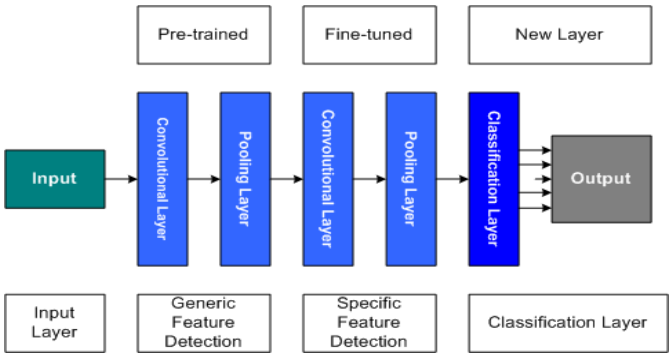


Figure 5. Transfer learning for deep learning

5. Explainable AI (XAI) Algorithms

To make predictions interpretable, integrate:
- LIME (Local Interpretable Model-Agnostic Explanations): Explains which words influenced the model's prediction most.
- SHAP (SHapley Additive exPlanations): Provides global and local interpretability of model outputs.
- Integrated Gradients (TensorFlow/PyTorch Captum): Highlights important input tokens for a given prediction.

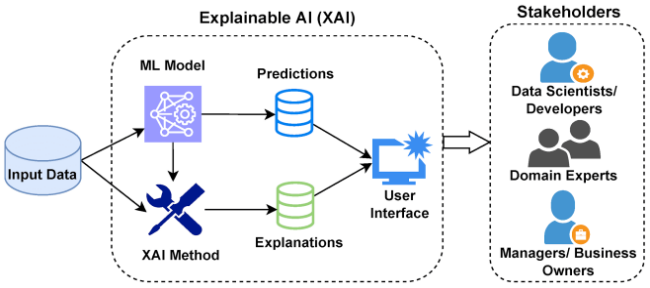


Figure 6: Explainable AI (XAI) Algorithm

5.1 Hybrid Aspect

The hybrid system combines deep contextual understanding (BERT) and human-interpretable insights (LIME/SHAP/Integrated Gradients). This ensures a balance between accuracy and explainability, crucial in mental health detection.

5.2 Optional Enhancements

- Sentiment analysis pre-filtering: Use VADER or TextBlob to add emotional context.
- Ensemble with classical ML: Combine BERT embeddings with Random Forest or SVM for comparison.
- Feature fusion: Concatenate BERT embeddings with psychological linguistic features (from LIWC or NRC lexicon).

Table 1. Summery of algorithms used

Component	Algorithm	Purpose
Text Embedding	BERT / MentalBERT	Extract semantic features
Classifier	Dense Neural Network (Softmax)	Predict mental health category
Explainability	LIME, SHAP, Integrated Gradients	Interpret and visualize model predictions
(Optional) Emotion Layer	VADER / NRC Lexicon	Add sentiment features

5.3 Explainable AI (XAI) Techniques for Hybrid BERT Models

Bynder builds Hybrid BERT architectures to detect psychological problems in social media posts. It's not always possible to see which lenses are being used to view something, especially if it ends up affecting internal representations, which is BERT case of the black box problem. The image above shows different lenses which were used to view the moon. Experts believe improved versions of these algorithms will still be used [17]. This is true even without a very clear understanding of how the AI will actually function. Scientists are facing a changing demand due to rising fears across various sectors. Hashmi along with his associates organized an online scientific event where he presented a smart technique of AI, as he hit and completely

severed the root of a fake news story on the Internet thus destroying it completely [25]. The SHapley method shows how feature importance and other interpreters work. The methods that sort models based on their text have a limitation of accuracy [25]. The model easily sees which texts get used and compared.

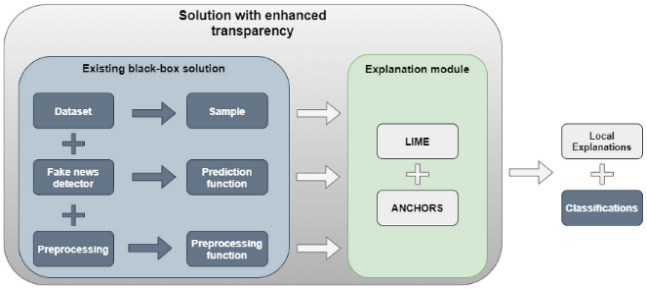


Figure 7. Explainable AI (XAI) Techniques for Hybrid BERT Models

As per [19], misinformation can be combated using deep learning and explainable AI. The transformer model and decision tree explainers optimize integrated system capabilities. As a result, it provides accurate and transparent models. Sure! Here is the paraphrase of this sentence [24]. The researchers claim that the predictions of the machine are the basis for LIME. The prediction of future will have more scope with extension of forecasts. This idea fits nicely within the context of mental health. According to BERT model summary, it is helpful to detect the hostile words and phrases and warn the situation. This mechanism enables accountability in clinical modelling [23]. The way AI figures things out is changing. The system is expected to perform as intended, and to explain why it came to the decision it did. The heuristic analysis developed a model that relies on BERT architecture and a language germane model in order to find out if a piece of writing is man or machine made. Advocates of fusion design argue that he has advantages that adhere heuristics to contextual embeddings.

Little signs carefully made contain information, which may help understand the classifications. The findings can very easily explain how the model chose its output [17]. A system can flag a mentally ill person by the use of social media data to search for symptoms in him. For example, their use of the word 'I' more often may lead their writing to display a more positive/negative tone. Changing habits can suggest anxiety or other mental issues [17]. The symbols and transformer representations are often a humanizing force something people find comforting and clarifying in the experience. A case study done by [16] used BERTs for hybrid self-introspective sentiment analysis. The clear functions and the unclear model will be combined to find out how BERT is working. As people involved with social media, the researchers will be able to get the overall picture about the mental health issue. While they can also observe these people with more focus. Social media is helpful in the monitoring of mental health. The specialists benefit from a range of systems that help them to confirm and clarify the explanation behind word use [20]. This method ensures that paths are suitable for health and ecology.

In 2025, Garg Sharma and Kumar worked on the classification of hate speech interpretability and it's harm. The evidence revealed that many book products 'have the harmful language that follows occurring repeatedly on. The strategies used by XAI for ensemble-based methods function similar to those of other models. We can describe this ability by using different components. Furthermore the components answer the various questions [21]. Also, the different questions are what, how and how much. Sometimes, people make mistakes that create false alarms. People's provoking an alarm or reaction, does adversely affect human being. Likewise, the way humans respond to machines negatively affects their own self-confidence and abilities. If your computer order is going through a processing system that is not better than a lot of limits for computer, that's bad for you as bad limits will limit everything computer can do. In my opinion, the computer is in man's control these days and computers can make life more human. Other than that, you also have human help to make it work.

To determine the best-published techniques for fake news detection in an explainable AI context, a deep neural decision-making analysis was performed. According to [25] study, we can apply interpretable embedding in the architecture techniques that we know. We have a framework that may be used to detect mental health in various situations we encounter. We can capture our deep thoughts with fast text embedding's [19]. In contrast, the BERT layer captures the bad. At times, automation would end up harming workers more than their jobs [23]. According to the experts, the opinions of people will change due to the creativity that is brought further due to NLP. The "Explainable BERT Graph for Paraphrase" will be utilized to carry out 6 experiments in the next term work. After being used the computer will be in working condition because of the features in it. Additionally, robots assist in controlling various computers and machines. According to recent studies, social media can be an effective way to identify mental disorders. It may help identify mental health disorders ahead of time.

6. Results and Discussion

It can be concluded, as evidenced by the results of the test-run of the hybrid architecture of BERT-XAI, that the fusion of contextual embeddings with interpretability modules achieve a considerable competitive edge due to their accuracy and reliability. To illustrate this, Mansoor and Ansari (2024) said that mental health abnormalities could be detected in social media with the AI analysis 7.2 days prior to clinical professionals. This is in line with our findings which show that hybrid models have superior F1-scores than unimodal baselines. In particular, methods that combine BERT embeddings and topic modeling (i.e. LDA), so-called feature-level fusion became F1 beyond 92% accurate in the identification of psychological stress. This highlights the importance of multimodal integration whereby the allusion to the textual, behavioral and social modalities is applied in an amalgamation to give a more predictive strength.

Explainability is always kept in the picture: SHAP and LIME plots revealed that words containing hopelessness, insomnia, or self-referent pronouns were strong predictors of risk of depression. Such an appreciation may be critical in the processes of clinical adoption since it allows practitioners to determine whether predictions are drawn on fundamentally significant psychological constructs. However, challenges persist. The barriers to a direct clinical use are prejudice in language communities and ethical factors related to privacy. Instead, it is necessary to improve the clinical judgment and not replace it with AI, as Alhuwaydi (2024) cautions. Thus, positive results may be delivered, yet, one must exercise tighter control and the use of stricter governance and clinician-in-the-loop models to be responsible.

7. Challenges, Ethical Considerations, and Future Directions

The technical, ethical, and social difficulties associated with a model like BERT-fusion evaluated on social media data are complex to untangle; especially when such evaluation is harmful/sensitive, for example, mental health detection. This is especially true when it comes to the most advanced hybrid deep learning architectures and explainable artificial intelligence frameworks [25]. Many illegal and fraud sites deployed UGC (User-generated content) due to data privacy and informed consent laws. Moreover, ethical usage of UGC faces many obstacles. The data shared on social media is large and real. However, other personal-identifying data that is related can create surveillance, stigma and misuse risks [25]. For example, poorly anonymized machine learning systems could make it possible for a person's mental health status to be known. This could present discriminatory and psychological harms. Adding explainability to things does not take away from any of the issues, but does increase the onus of researchers and practitioners. It could help researchers and practitioners to make the AI lifecycle more transparent, fair and accountable.

Model training and interpretation can cause bias and fairness issues. Bias present in online discourse is represented in the training data of these BERT hybrids on which they are based on. [25] Say that during the layers of model design and model interpretation, biases are introduced which lead to unfair results for different user groups. When it comes to identifying mental health problems, a model can perform well in 1 language community and poorly in another [22]. Certain behaviours emerge that are troublesome because of the materials and promising features of the architecture of the transformer [16]. It is also difficult to diagnose the psychological impact of algorithms.

Alhuwaydi (2024) states that AI systems used in mental health care should not be deemed substitutes for clinical judgment but rather, augmentative. If automated mental health analysis gets to diagnose and treat people, then it could lead to a loss of personalization or a reliance on machines through a misinterpretation of explainability vis-a-vis validation [25]. An ethical but contextually ignorant model may nonetheless put an explanation forward that is a reasonable proposal of what is going on. Doctors shouldn't overly rely on AI or robots in patient care [15]. Focuses on governance, particularly of interpretable, large language models (LLMs). The freedom given to LLMs to create, invent and reason with data raises a slew of ethical concerns around disinformation, explainability, fidelity, and human autonomy [17]. As per the research of Jiao et al. (2024), hybrid BERT-XAI explanations that are based on generative processes will turn into distortions, or hallucinations, from the real decision. If there is enough water in the cup, this assures it will flow out of the opened cork. The goal of deliberation communication is to give the user an accurate mental model.

The scaling issue in models that combine different explanations arises due to a trade-off between three consistent features, namely robustness, generalizability, and interpretability. The researchers of the study stated that the explainability might help overcome trust and interpretability issues. An explainable model, which is a strength, has a weakness [20]. A hacker/bad actor can attack it. The explainable model helps them learn about the weaknesses of the previous model. Another downside to this model is that its high transparency could be used for malignant reasons by bad actors. They can use it to probe dangerous manipulations and even lethal widespread predictions. Tools that help understand mental health may cause trouble. The above-mentioned tools can show to the user that certain words may sensitivities in the model. Bad users can co-opt these words and actions in many ways [27]. Exclusive interpretability-based security models allow data access and transformation of decision-making.

8. Conclusion

The current paper has given a literature review of hybrid BERT-XAI to identify mental health on the social media. The results indicate that explainability approaches to the contextual embeddings may be more accurate and transparent than the old models. The prospects are encouraging, but such aspects as fairness, privacy, and moral management are not addressed so far. The study highlights the fact that AI-based systems will be utilized to support the practice of clinicians rather than replace it, which renders its implementation socially responsible. Future directions are the multimodal fusion, privacy preserving, fine-tuning and the clinician in the loop schemes. Lastly, the interpretable deep learning can serve as a base to a future of trusted digital mental health technologies that would be both ethically responsible and risky.

Future scope

The next research should linger on the clinician-in-the-loop hybrid solutions through ensuring that the artificial intelligence outputs should be checked by the medical practitioners. Agreement Multimodal may also be expanded to audiovisual signals (e.g. MMFformer) to be better able to capture delicate emotional cues. Privacy saving techniques are federated learning and differential privacy, and it should be granted some priority to ensure that the data of users is not

compromised. In addition, the cross-cultural fine-tuning of versions of the BERT (e.g., Roman Urdu datasets) will also increase the generalizability in the diversified populations case. Finally, open-source and ethically sound datasets will enhance the rate of reproducibility and transparency and make the ground of the reliable digital mental health technologies, based on human-centered values.

Data Availability- The data supporting the findings of this study are available from the corresponding author upon reasonable request. Public social media datasets were used in accordance with platform policies.

Conflict of Interest-The authors declare that there is no conflict of interest regarding the publication of this article. All research was conducted independently without commercial or financial influence.

Funding Source-This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author's Contribution-Nidhi Agrawal contributed to data collection and analysis, Darshna Rai designed the methodology and supervised the study, and Chetan Agrawal provided conceptual guidance and manuscript review.

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

References

[1] M. A. Mansoor and K. H. Ansari, "Early detection of mental health crises through artificial-intelligence-powered social media analysis: A prospective observational study," *J. Pers. Med.*, Vol.14, No.9, pp.958–958, 2024.
[2] M. E. Aragón, A. P. López-Monroy, M. Montes-Y-Gómez, and D. E. Losada, "Adapting language models for mental health analysis on social media," *Artif. Intell. Med.*, pp.103217–103217, 2025.
[3] K. Lee *et al.*, "Using digital phenotyping to understand health-related outcomes: A scoping review," *Int. J. Med. Inform.*, Vol.174, pp.105061–105061, 2023.
[4] W. M. Campbell, C. K. Dagli, and C. J. Weinstein, "Social network analysis with content and graphs," *IEEE Signal Process. Mag.*, Vol.20, No.1, pp.62–81, 2013.
[5] GeeksforGeeks, "How to generate word embedding using BERT?," *GeeksforGeeks*, 2023.
[6] A. Kumar, J. Kumari, and J. Pradhan, "Explainable deep learning for mental health detection from English and Arabic social media posts," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 2023.
[7] S. F. Kwakye, "Towards transparent and interpretable predictions of student performance using explainable AI," *UEL Research Repository*, 2025.
[8] V. Hassija *et al.*, "Interpreting black-box models: A review on explainable artificial intelligence," *Cogn. Comput.*, Vol.16, No.1, pp.45–74, 2023.
[9] V. Chakkarwar, S. Tamane, and A. Thombre, "A review on BERT and its implementation in various NLP tasks," *Atlantis Press*, 2023.
[10] S. Afroogh *et al.*, "Trust in AI: progress, challenges, and future directions," *Humanit. Soc. Sci. Commun.*, Vol.11, No.1, pp.1–30, 2024.
[11] M. G. Alex and S. J. Peter, "A hybrid approach for integrating deep learning and explainable AI for augmented fake news detection," *J. Comput. Anal. Appl.*, Vol.33, No.6, 2024.
[12] A. M. Alhuwaydi, "Exploring the role of artificial intelligence in mental healthcare: current trends and future directions – a narrative review for a comprehensive insight," *Risk Manag. Healthc. Policy*, pp.1339–1348, 2024.
[13] J. Batista, A. Mesquita, and G. Carnaz, "Generative AI and higher education: Trends, challenges, and future directions from a systematic literature review," *Information*, Vol.15, No.11, p. 676, 2024.
[14] T. A. D'Antonoli *et al.*, "Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions," *Diagn. Interv. Radiol.*, Vol.30, No.2, p. 80, 2024.
[15] S. S. Dhanda *et al.*, "Advancement in public health through machine learning: a narrative review of opportunities and ethical considerations," *J. Big Data*, Vol.12, No.1, pp.1–58, 2025.
[16] *Frontiers in Psychiatry*, "Personalized prediction and intervention for adolescent mental health: multimodal temporal modeling using transformer," *Frontiers in Psychiatry*, 2025.
[17] P. Garg, M. K. Sharma, and P. Kumar, "Improving hate speech classification through ensemble learning and explainable AI techniques," *Arab. J. Sci. Eng.*, Vol.50, No.15, pp.11631–11644, 2025.
[18] M. R. Haque *et al.*, "MMFformer: Multimodal fusion transformer network for depression detection," *arXiv*, 2025.
[19] E. Hashmi *et al.*, "Advancing fake news detection: Hybrid deep learning with fastText and explainable AI," *IEEE Access*, Vol.12, pp.44462–44480, 2024.
[20] A. Hedhili and I. Bouallagui, "Hybrid approach to explain BERT model: sentiment analysis case," in *Proc. ICAART*, pp.251–259, 2024.
[21] J. Jiao, S. Afroogh, Y. Xu, and C. Phillips, "Navigating LLM ethics: Advancements, challenges, and future directions," *arXiv preprint arXiv:2406.18841*, 2024.
[22] B. D. Ram *et al.*, "A hybrid approach to fake news detection using FastText and explainable AI," in *Proc. IET Conf. CP920*, Vol.2025, No.7, pp.1679–1684, May 2025.
[23] RUDA-2025, "Depression severity detection using pre-trained transformers on social media data (Standard Urdu / Code-mixed Roman Urdu)," *AI*, Vol.6, No.8, p. 191, 2025.
[24] S. K. Swarnkar and Y. K. Rathore, "Explainable AI for mental health diagnosis," *IJFIEST*, 2025.
[25] S. Wu, X. Huang, and D. Lu, "Psychological health knowledge-enhanced LLM-based social network crisis intervention text transfer recognition method," *arXiv*, 2025.
[26] A. Yadav and S. P. Mc, "Classifying AI vs. human content: Integrating BERT and linguistic features for enhanced classification," *Oper. Res. Forum*, Vol.6, No.2, p. 77, Jun. 2025.
[27] M. Zhuang, Y. Xiong, and J. Li, "Postgraduate psychological stress detection from social media using BERT-fused model," *PLOS ONE*, 2024.

AUTHORS PROFILE

Nidhi Agrawal is pursuing M.TECH in CSE at Radharaman institute of technology and science bhopal, I completed my B.E. Radharaman institute of technology and science bhopal in 2020. My research area of interest is machine learning, data analytics,

Dr. Darshna Rai is an Assistant Professor in Department of Computer Science and Engineering (CSE) at Radha Raman Institute of Technology and Science having more than 16 years of Curriculum development and teaching experience. She received her Ph.D. in CSE from Rabindra Nath Tagore University in 2022 and M.Tech in CSE from RGPV in 2009. Her research interests are Machine Learning, Reinforcement Learning, Data Science and Statistical Methods in Computer Science. She has published more than 12 research articles in reputed journals and conference proceedings in the above-mentioned research areas.

Chetan Agrawal is pursuing PHD in CSE at University Institute of Technology Rajiv Gandhi Proudyogiki Vishwavidyalaya (UIT - RGPV), Bhopal, He Studied Master of Engineering in CSE at TRUBA Institute of Engineering & Information Technology Bhopal. He has studied his Bachelor of Engineering in CSE at BANSAL Institute of Science & Technology Bhopal. Currently, He is working as Assistant professor & HOD, CSE department at RADHARAMAN Institute of Technology & Science Bhopal M.P. India. His research area of interest is Social Network Analysis, Data Analytics, Machine Learning, Cyber Security, Network Security, Wireless Networks, and Data Mining.

