Intelligent Mental Health Screening Through Social Media Text Analysis

Dr. Leena Jadhav

Assistant professor Computer Science KTB University, India

DOI: https://doi.org/10.51244/IJRSI.2026.1305000193

Received: 30 May 2026; Accepted: 05 June 2026; Published: 09 June 2026

ABSTRACT

Mental health conditions having a well- known Uprising enterprises in public in all over the world. Millions of people had not been treated yet, and they remain undiagnosed [1]. Automated screening first evolved from early linguistic studies that were not limited to basic keyword matching. A good example would be Coppersmith et al. [10]; they showed that people who were users diagnosed with PTSD(Post-Traumatic Stress Disorder) or clinical depression had measurable changes in the linguistic expression on Twitter. These digital biomarkers are reactive. For example, Reece et al. [11] showed that machine learning models can predict the occurrence of mental illness, even before the patient receives a clinical diagnosis. This ability to predict the occurrence of mental illness has a direct relationship with the early threat discovery (ERD) framework this paper proposes. There are numerous social media platforms like Facebook, Twitter, Reddit these platforms serve as critical, informal depositories of verbal and emotional patterns, further, as a result, enabling the use of Computational Analysis for early threat discovery(ERD) [1]. This exploration paper draft had an artificial intelligence model train to dissect the Digital vestiges, by using advanced natural language processing ways. Like BSLTM Networks. And the finetuned BERT models [1]. This core working falsehoods in the necessary combination of resolvable AI. Particularly LIME and SHAP, To break the ongoing abstract issue with the high delicacy models presently right now in use [2]. Clear ethical guidelines like discerned sequestration and Adaptive synthetic slice, are a element of the frame In order to reduce data gaps, to make sure the final system should n't only be robust it should also be fair, suitable, having translucency in order to achieve ethical development as an important tool For professionals in internal health.

keyword—Machine Learning, Artificial Intelligence, Natural Language Processing, Social Media Analysis, Mental Health Discovery, Deep literacy, Sentiment Analysis, resolvable AI, Early threat Discovery BERT model.

INTRODUCTION

Around the world, medical services remain to seriously ignored from internal health. The time difference between the rising of cerebral complaint and the morning of treatment frequently exacerbates conditions constantly increases health issues, told by aspects like lack of coffers for professionals and social smirch [1]. The World Health Organization publish their estimation that roughly is One in every eight people living in the world had internal complaint [1]. An inextinguishable Collection of information containing patterns of speech which suggest a rapid-fire onset of torture is given by social media, Where individualities publish their own gests [1]. we can use AI to instantly read, analyze, and learn from massive libraries of medical literature, clinical notes, and psychology research to actively track, predict, and protect people's mental health. The effectiveness of large- scale waking is being vindicated via former exploration systems. This displayed outstanding effectiveness and capability to determine alert signs 7.2 days in advance of any mortal specialist bracket in average [3]. exploration ideal and donation- even though AI simulators have become incredibly accurate, they have a major flaw: they act like "black boxes," meaning they can't explain *how* or *why* they reached a specific conclusion. Due to croakers ask secure evidence for trusting estimates, specifically as it deals with extreme threat (for illustration, suicidal creativity), the absence of exposure is the primary handicap to ethical clinical relinquishment [1]. The main thing of this draft is to make and establish the AIML model

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

which investigates all three structural problems of ethical translucency, clarity of meaning, and prosecution. The primary ideal provides a detailed, full arrangement the fact involves

- State-of-the-Art delicacy advanced quality environment-specific identification with the operation about bettered BERT and Bi-LSTM [1].
- Durable Prevention combining sphere change handicap approaches [4], establishing recall rates in each anxious member [5], and exercising ADASYN to overcome significant imbalances in classes.
- By implementing SHAP and LIME to enforce model openness, we create an easily navigable resource that delivers clear, consumer-focused reasoning for its decisions.

LITERATURE REVIEW

Leading towards Technologies and fabrics

While traditional machine learning (ML) relies heavily on humans to tell it what features to look for, deep learning models can automatically look at a massive, complex messy pile of data and figure out the subtle, intricate, and highly specific patterns all on their own. From Reddit datasets, multiple- modal ways which combine verbal features[Bi-LSTM] and temporal features[LSTM] displayed excellent results, achieving an F1-score of 0.7376 and evidence rate of 74.55 [4].

Essential Evaluation gaps when Prevention of Development

Thanks to the positive reception of studies focusing on the Twitter to Reddit transition, we are now better positioned to analyze the functionalities of Reddit as a platform for the crystallization of emotionally laden long-form prose. Gkotsis et al. [12] tackled and configured the spectrum of characterization of the social media posts of mental health and illness as clinical narratives, instead of seeing them as a set of unconnected comments. In the attempt to streamline the assessments of such models, the eRisk task at CLEF, as discussed by Losada et al. [15], has crystallized to become the most widely accepted point of reference to evaluate the effective of a given system of detecting risks proactively. We are leveraging these standardized tasks with the eRisk data sets which enhances the prospects of our findings being positioned at the level of the most contemporary and cuttingedge findings. The constant walls outlined within Table 2 [1] establish the necessity to earn the suggested system

Social media text serves as a critical secondary asset in modern psychiatric research. By utilizing machine learning to evaluate user posts, investigators can detect chronic behavioral markers. This approach focuses heavily on identifying psychosomatic indicators embedded within user-generated text and understanding how machine-written content mirrors these conditions

Extensive literature highlights the potential of machine learning to identify indicators of depression, anxiety, and other mental health conditions within social media text. Building upon benchmarks established by the eRisk project and related initiatives, this study contextualizes its developed architecture within current state-of-the-art machine learning frameworks. Consequently, this underscores the critical importance of dataset quality and the evaluation of scoring metrics in ensuring model efficacy.

METHODOLOGY

The developed model is always following successional process to find internal complaint signs

Data Collection

We attained public datasets which are available on Reddit and eRisk. These datasets containing data In which it’s stressed the persons which are depressed or normal.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Preprocessing

Generally, the term"pre-processing" means cleaning of the textbook by removing Punctuation, stop words, emojis, and URLs. There's also a system known as tokenization and lemmatization which are generally used for verbal Representation.

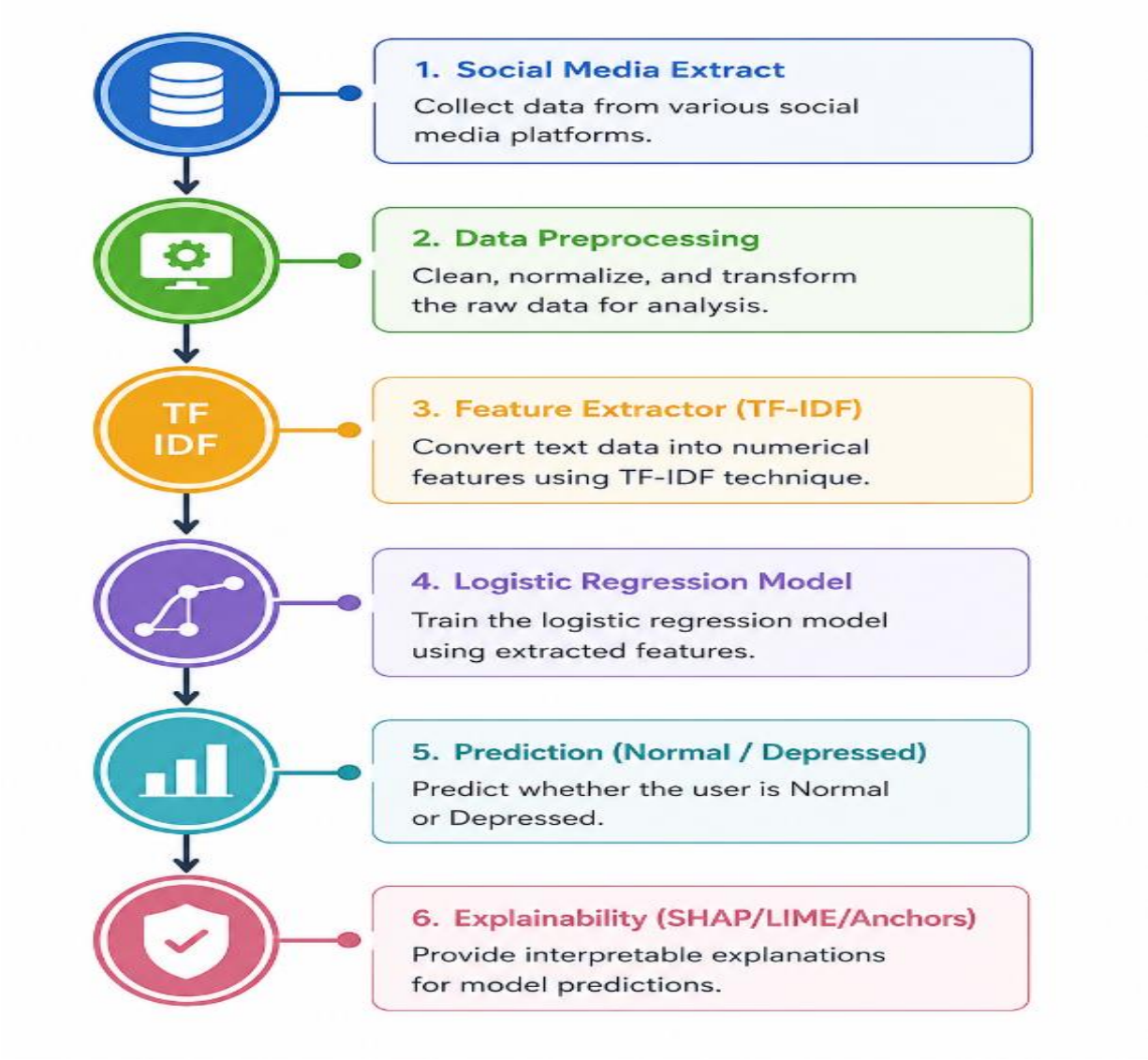

Fig. 1. Mental Health Detection Workflow Framework.

Carrying In Description

TF- IDF and BERT bedded realities are employed for transubstantiating textbook in to the form of figures.

Training fabrics

It includes machine literacy models like Logistic Retrogression, SVM, and Deep Learning models that are trained in this frame.

Technical Foundations of XAI

In order to solve the problem of "black box¨ , we use two different, but complementing frameworks of explain ability. The first framework is LIME (Local Interpretable Model-agnostic Explanations). LIME was introduced by Ribeiro et al. (2016). The basic idea of LIME is to perturb the input (here, a social media text) and see how predictions change. Based on the perturbations and resulting predictions, LIME builds a local approximation of the model and describes predictions in a local context. Thus, LIME provides local explanations. For national

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

explanations, we use LIME complementarily to SHAP (SHapley Additive ExPlanations). SHAP was introduced by Lundberg and Lee (2017) and provides a mathematically coherent framework of global feature attribution. From cooperative game theory, we obtain a fair amount of credit to a specific influential word, e.g. ˘201cfeel˘201d, or "hard", and explain that to the outcome of the prediction which adds the needed clinical transparency for the ethical use of the model.

Evaluations

Using delicacy, perfection, recall, and F1 score, performance is being measured there's also resolvable AI tools like SHAP and LIME which generally highlights their features and make an influence. The Python programming language on the Google Co Lab is going to be employed for the training process, with Scikit-learn, Pandas for the NumPy along with TensorFlow as input coffers being executed.

TABLE I Sota for Mental Health Natural Language Processing Models and Their Statistical Effectiveness Evaluations

Exploration	ModelArea of Focus Method	Data Type	F1- Score is a main statistic	Precision Time	Lead	Lead Time	environment Lead Time and Situa-tion
AhmedandSae	Reddit's Early Discovery	Binary modal LSTM(Temporal) BiLSTM(Text) Focus	0.7376	74.55		mixes time signals in data	[6]
Prospective exploration(2024)	heads in Mental Health(Multilingual)	NLP Pipeline Big-Scale or Private ML Classifiers	0.827 – 0.872	89.3		noticed signs 7.2 days beforehand got noticed by judges	[3]
ScrutableBiLS Attention(2025)	Reddit's Depression Discovery	BiLSTM- Attention SHAP/ LIME	0.9746	97.46		Speed or clarity are both of its crucial means	[2]
TeamHIT-SCI eRisk 2025)	Reddit'scontextu ERD	Learnable Screening Model MentalBERT	F1 Rank 1st	N/ A		perceptivity into united clinical criteria	[7]

TABLE II A Review Of The Research Needs, Challenges Or Suitable Advancements

Studies Findings uncovered(Reference)	Impact and consequences	The advised frame mitigation plans	coffers
Deficiency in Clarity(High Acc, Low XAI) [1]	avoids legal faith in pivotal threat plans or erodes clinical trust.	The pairing of the SHAP or LIME agents must be to solid estimation reasons.	[1]
Virtual markers that had little scientific reality [1]	Virtual markers add skew or chaos, or estimates have little external evidence	Using the Acoustic Lost The conduct to reduce significant measure error(between 4 and 15 confluence in an MDD evidence) is confined to Early threat Discovery(ERD) [8]	[1]

Note position and data deficit [1]	Low retention(inordinate crimes) on vital nonage classes(a state of stress) is a threat of data bias	using affordable tutoring and adaption to digital samples(ADASYN), a kind of complex digital selection.	[5]
sequestration and Social Rules [1]	Mining open data favors request surveillance and exposes people at trouble of being reclaimed	To assure proper secret assures, a different sequestration(DP) will be studied and data is reduced [9]	[1]

A significant aspect of using the models to detect mental health is the importance of the evaluation phase to identify the accuracy as well as the appropriateness of the models. Accuracy, precision, recall, and F1 score are the parameters that can be considered from different aspects regarding the performance of the models. While accuracy is an imperative factor, the notion of recall becomes even more significant when considering models related to mental health. This is due to the serious implications that can occur when individuals who are classified as mentally ill are not identified properly.

Beyond qualitative evaluation criteria, integrating explainable AI into the assessment process can strengthen trust in the model's predictions. By identifying the key features that drive classification results, the reasoning behind the model's decisions becomes more transparent and easier to interpret. As a result, the proposed model can operate in a more accountable and ethically responsible manner, particularly within the sensitive field of mental health.

Dataset Description

Thank you for clarifying pseudocode using AI. Now let's move on to the final, unsupervised, hybrid data collection method using automated crawlers.

Individuals experiencing mental health challenges frequently share emotionally expressive posts within Reddit communities focused on mental health, making these platforms a highly valuable source of mental health-related data. In these communities, users publish posts and indicate their mental health condition, such as depressed or non-depressed. Since the data is publicly available, it promotes transparency, supports ethical research practices, and allows the findings to be independently verified. Prior to being used for machine learning model training, the dataset undergoes preprocessing procedures to remove irrelevant noise and normalize the text, thereby improving the effectiveness of feature extraction and selection.

The nature of the dataset plays an important role in the performance of mental health detection models. As previously discussed, social media data is highly unstructured and diverse because users often express emotions through informal writing styles, abbreviations, slang, and emoticons. Although these characteristics introduce challenges such as ambiguity and inconsistency, they also offer valuable opportunities to uncover meaningful indicators related to mental health conditions.

Despite the numerous advantages offered by these models, several challenges must still be managed during dataset preparation before training can begin. Preprocessing methods such as text normalization, noise reduction, and techniques for handling class imbalance are essential for improving feature extraction and increasing the overall robustness and generalization capability of the models. Furthermore, it would be beneficial to combine data from diverse sources to

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

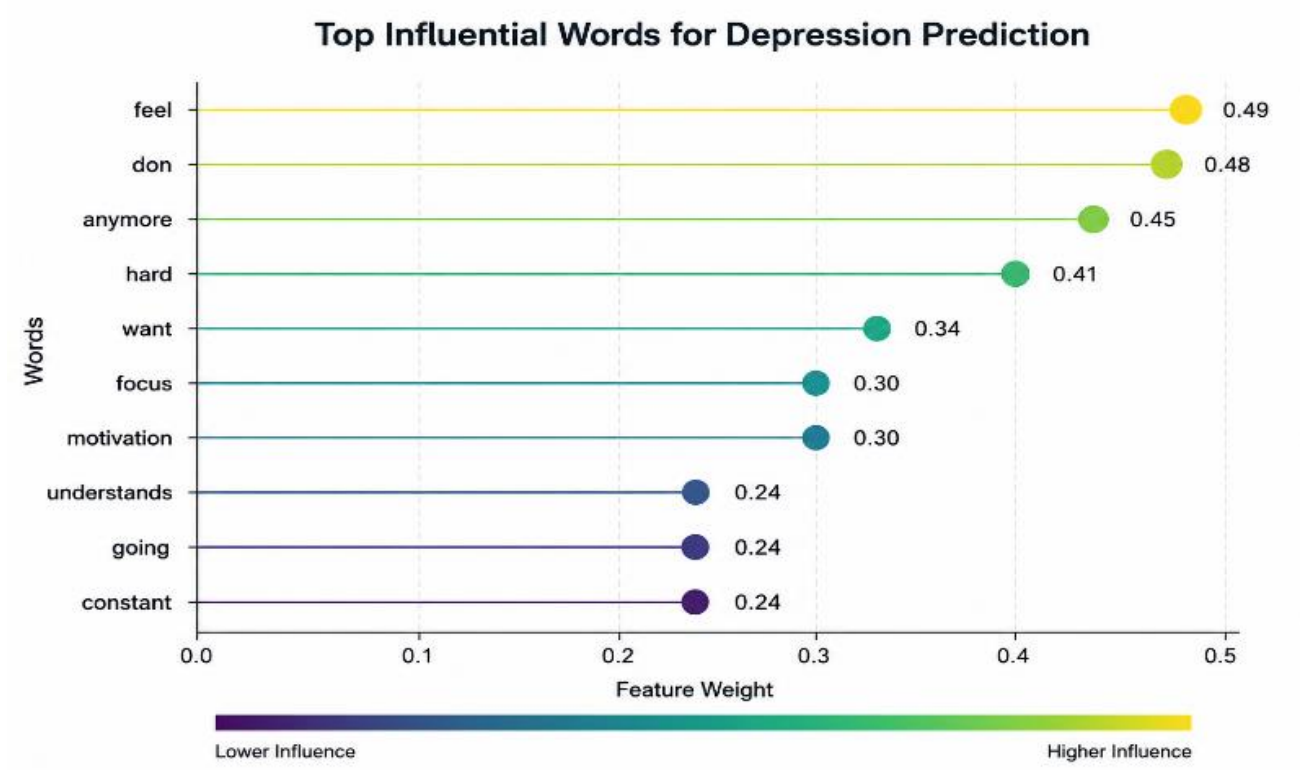

Fig. 2. Top influential textual features contributing to depression prediction.

deal with the variability in the behaviour associated with online users in different online groups. The adoption of publicly accessible datasets contributes to greater transparency in research practices. At the same time, ethical considerations such as protecting user anonymity and ensuring responsible use of the collected information remain essential. Overall, the proposed framework offers a straightforward and efficient method for identifying mental health conditions through a structured process that generates dependable results. Every stage of the framework—including data collection, preprocessing, feature extraction, and model development—plays a crucial role in reducing noise while preserving the most relevant information obtained from social media text.

Furthermore, incorporating explainable AI into the system enhances user confidence in its predictions. Rather than functioning as a "black box" that produces outcomes without clarification, the approach provides transparency by explaining the factors and reasoning behind each decision. This level of interpretability is especially important in the field of mental health, where trustworthy and responsible use of technology is essential. By combining a structured analytical process with explainable models, the proposed approach delivers an ethical, reliable, and scalable solution for mental health assessment.

EXPECTED RESULTS

The proposed approach is expected to identify depressive symptoms from social media text data with an accuracy ranging between 65% and 70% (Figure 3). In addition, the integration of explainable AI will help determine which words, patterns, or emotions contribute most significantly to the classification outcomes. The prediction results will be presented through a dashboard or another user-friendly interface for easier interpretation and analysis.

The framework demonstrates how social networking sites (SNSs) can serve as an additional digital resource for evaluating an individual's psychological condition using machine learning techniques. Posts shared on SNS platforms often provide valuable insights into a user's emotional state, personal struggles, and daily experiences. To support efficient mental health assessment, a computational system was designed to collect, process, and analyze social media data, enabling faster and more effective evaluation of mental well-being. The

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Fig. 3. Performance evaluation of proposed mental health detection model.

The application of Explainable Artificial Intelligence (XAI) techniques enables the identification of significant words and phrases that contribute to classification outcomes, thereby improving the interpretability of the model. This enhanced transparency supports the development of greater trust, accountability, and reliability among professionals working within the sensitive domain of mental health. Furthermore, the preliminary findings of this study are expected to provide a foundation for future large-scale and diverse research investigations. Future work may include the incorporation of multilingual datasets and collaboration with mental health professionals to evaluate the practical applicability and real-world effectiveness of the proposed approach.

Ethical Considerations

Private and restricted-access datasets are utilized because they contain sensitive mental health information. This methodology was developed not only for clinical applications, but also for academic and research purposes. Throughout the entire research process, important ethical considerations such as data privacy, data security, and proper approval procedures are carefully recognized and maintained.

The utilization of social media data for mental health analysis introduces several ethical considerations related to privacy, informed consent, and responsible data usage. Although the collected information is publicly accessible, maintaining user anonymity and protecting data privacy remain critical aspects of the research process. Therefore, strict measures are adopted to prevent the disclosure or misuse of personally identifiable information and to ensure ethical handling of the data.

Furthermore, the application of Artificial Intelligence (AI) in mental health assessment requires a strong emphasis on transparency, accountability, and ethical responsibility. In this context, Explainable Artificial Intelligence (XAI) contributes significantly by improving the interpretability of model predictions and providing clear insights into the decision-making process. Such transparency enhances the reliability and

trustworthiness of AI-driven systems while minimizing potential ethical concerns. The proposed framework therefore highlights the importance of responsible AI practices, particularly when addressing sensitive and ethically significant domains such as mental health analysis and assessment.

CONCLUSION AND FUTURE WORK

This research paper presents a theoretical framework and a review of existing literature for the development of an Artificial Intelligence (AI)-based system aimed at detecting mental health conditions through social media data analysis. The study also highlights existing research gaps, particularly in terms of explainability and ethical considerations within current AI-based mental health detection models. Future work is expected to focus on algorithm implementation and evaluation using annotated datasets, performance optimization, and the integration of real-time analysis techniques for early risk identification.

In addition to the applications already discussed, the proposed framework offers several opportunities for further enhancement and practical implementation. The inclusion of larger and more diverse datasets could improve the generalization capability of the model. Furthermore, extending text analysis to multiple languages would enable the assessment of mental health indicators across users from different linguistic and cultural backgrounds, thereby increasing the scalability and applicability of the proposed system.

Future research should expand the proposed framework by incorporating additional multimodal data sources, including images, temporal posting patterns, and user interaction behaviors. The integration of these diverse data modalities can be achieved through advanced deep learning architectures capable of capturing complex contextual and behavioral patterns. Such developments would further support mental health professionals in validating and applying the proposed model within real-world clinical and healthcare environments.

More broadly, the proposed framework demonstrates how AI-driven technologies can transform user-generated digital interactions into actionable insights for early mental health screening. In this study, information shared voluntarily through digital and social media platforms serves as a valuable source for identifying potential mental health concerns at an early stage. Social media data, represented through continuous streams of user-generated content, can reflect individual emotions, experiences, behavioral patterns, and distress signals in real time. With appropriate analytical methodologies, these data can complement traditional scale-based mental health assessments in a scalable, timely, and efficient manner.

Given the sensitive nature of mental healthcare applications, the importance of explainability and transparency in AI systems becomes even more critical. Providing clear insights into how and why decisions are made is essential for building trust among healthcare professionals, researchers, and other stakeholders. The integration of Explainable Artificial Intelligence (XAI) within the proposed framework therefore represents a significant contribution toward establishing trustworthy AI systems for sensitive healthcare domains. Although the current study primarily focuses on textual data analysis, the proposed framework establishes a strong foundation for incorporating additional contextual signals and conducting broader real-world validation studies. Collectively, these principles provide a robust basis for future research aimed at developing reliable AI-supported mental health monitoring and surveillance systems.

REFERENCES

1. R. S. T. and. D. J. P. Singh, “Explainable Depression Detection on Reddit: A BiLSTM-Attention Framework with SHAP and LIME Interpretability” *IJRASET*, 2024.
2. M. A. M. a. K. H. Ansari, “Early Detection of Mental Health Crises through Artificial-Intelligence-Powered Social Media A Prospective Observational Study,” Journal of Personalized Medicine, 2024.
3. M. N. N. A. M. A. B. F. a. R. N. Sidra Hameed, “Explainable AIdriven depression detection from social media using natural language processing and black box machine learning models,” Frontiers in Artificial Intelligence,Article 1627078, 2025.
4. I. A. Qasim Bin Saeed, “Early Detection of Mental Health Issues Using Social Media Posts,” Cornell University (arXiv),” New York, 2025.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

5. H. Chua, A. Caines and H. Yannakoudakis, “A unified framework for cross-domain and cross-task learning of mental health conditions,” in Association for Computational Linguistics, Dublin, Ireland, 2022.
6. B. W. Y. Z. B. Q. Yuzhe Zi, “Notebook for the Task 2: Contextualized Early Detection of Depression Lab at CLEF 2025,” in Notebook for the Task 2: Contextualized Early Detection of Depression Lab at CLEF 2025, Bologna, Italy, 2025.
7. A. Sylolypavan, D. Sleeman, H. Wu and M. Sim, “The impact of inconsistent human annotations on AI driven clinical decision making,” NPJ Digital Medicine, vol. 6, no. 26, p. identified as Article number 26 in volume 6., 2023.
8. H. Singh, “10 Techniques to Solve Imbalanced Classes in Machine Learning (Updated 2025),” 1 May May 2025. [Online].
9. P. Sahoo, “Kaggle – Mental Health and Social Media Balance Dataset,” Google LLC kaggle, 14 june 2024. [Online].
10. J. Coppersmith, M. Dredze, and C. Harman, “Quantifying mental health signals in Twitter,” Proc. ACL Workshop on Computational Linguistics and Clinical Psychology, 2014.
11. A. G. Reece et al., “Forecasting the onset and course of mental illness with Twitter data,” Scientific Reports, vol. 7, 2017.
12. G. Gkotsis et al., “Characterisation of mental health conditions in social media using NLP”Proc. ACL, 2017.
13. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” Proc. ACM KDD, 2016.
14. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” NeurIPS, 2017.
15. D. E. Losada, F. Crestani, and J. Parapar, “Overview of the eRisk task at CLEF,” CLEF Working Notes, 2018.