iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IX **Month of publication:** September 2025

DOI: https://doi.org/10.22214/ijraset.2025.73995

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

Explainable Depression Detection on Reddit: A BiLSTM-Attention Framework with SHAP and LIME Interpretability

Ranjeet Singh Thakur[1], JP Singh[2]
[1]*M.Tech Scholar, Department of Computer Science and Engineering, Chouksey Engineering College, Bilaspur (C.G.), India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Chouksey Engineering College, Bilaspur (C.G.), India*

Abstract: ***Early identification of depression by social media content analysis has drawn increasing attention as it is a common mental health issue. To identify sadness in Reddit posts, this study proposes a novel framework that combines advanced deep learning and explainable AI techniques. A hybrid CNN+XGBoost model was implemented as the baseline, achieving 92.68% accuracy. To address its limitations, a BiLSTM with an Attention mechanism was developed, which captured long-term sequential dependencies and emphasized clinically significant tokens. The proposed model significantly outperformed the baseline, achieving 97.46% accuracy, a 0.9746 F1-score, and balanced precision–recall performance. For transparency, SHAP and LIME were applied to highlight influential linguistic cues at both local and global levels, thereby improving interpretability. The findings demonstrate the dual strength of predictive performance and explainability, offering a reliable framework for potential clinical and real-world applications.***

Keywords: Depression Detection, Social Media Analysis, BiLSTM with Attention, CNN + XGBoost, Explainable AI, SHAP, and LIME.

I. INTRODUCTION

Among the most prevalent mental health conditions is depression, significantly impinging upon an individual's cognition, behavior, and quality of existence. The World Health Organization (WHO) reports that depression is a global mental health concern, affecting an estimated 280 million individuals across the world [1], and annually, approximately 700,000–800,000 individuals die by suicide as a consequence [1]. In 2023, WHO identified depression as a leading global health concern, posing not only psychological but also economic challenges. In our digital era, individuals frequently express their emotions, experiences, and thoughts openly on social media platforms such as X (Twitter), Facebook, and Instagram, and Reddit, using text, emojis, and images. These posts can serve as essential indicators of their mental state. However, such sensitive disclosures are sometimes misused—for example, through blackmail. Detecting signs of depression at an early stage can therefore facilitate timely and effective intervention.

To address early detection, Natural Language Processing (NLP) can extract emotional and linguistic hints from writing on social media. For automatic depression identification, these cues can then be fed into models for traditional machine learning (ML) techniques and modern deep learning (DL) methods [2]. Furthermore, Explainable AI (XAI) approaches were incorporated, with SHAP and LIME increase model transparency by elucidating which words or features contributed to classification decisions. Traditional diagnosis of depression relies on psychological interviews, clinical questionnaires, and medical assessments, which are typically time-intensive, expensive, and inaccessible to many. Social media offers an abundant alternate data source for mental health screening. Nevertheless, significant challenges persist, incl: uding: unstructured and noisy data, small and imbalanced datasets, complexity in classification tasks, and Lack of model interpretability. Prior work predominantly focused on classical ML/DL models such as SVM, Random Forest, CNN, and LSTM, whereas the application of XAI techniques remains relatively sparse [3].

The purpose of this study is to identify depression from Reddit posts by integrating advanced modeling with explainability techniques. As a baseline, a hybrid model combining CNN and XGBoost has been implemented, while the proposed framework employs a Richer contextual and sequential dependencies in textual material can be captured by a BiLSTM network with an attention function. Furthermore, SHAP and LIME are utilized to enhance model interpretability, enabling clear insights into the linguistic cues influencing classification outcomes. Overall, the proposed framework emphasizes both predictive accuracy and transparency, thereby offering a more reliable solution for potential clinical applications.

©IJRASET: All Rights are Reserved

The following is an outline of this paper's structure. The previous research on social media depression identification is covered in Section 2, along with the drawbacks of explainable AI, both Earlier studies have primarily explored conventional machine learning algorithms alongside recent advancements in deep learning, whereas Section 3 of this paper details the methodology adopted in the present work, including dataset construction, preprocessing, baseline CNN+XGBoost model, and the proposed BiLSTM with Attention framework. The evaluation metrics and experimental setup are also detailed in this section. The findings and discussion are shown in Section 4 where the baseline and proposed models are compared through classification metrics, confusion matrices, and explainability analysis using SHAP and LIME. Section 5 highlights future directions, such as extending the work to multi-class datasets, exploring transformer-based architectures, and deploying the framework in real-world applications. Finally, the paper concludes with the key findings, emphasizing both the predictive performance and the interpretability of the proposed approach.

II. LITERATURE REVIEW

A. History and Context

Social media depression detection has become a prominent field of study over the last decade. Early works focused on linguistic analysis and keyword-based detection to identify users exhibiting depressive tendencies [10,11]. Losada & Crestani [10] created one of the first benchmark datasets using forum and Reddit posts, providing a foundation for subsequent research. Indian researchers, such as Gupta et al. [11], analyzed regional social media platforms to understand culturally-specific expressions of mental health and depression. These early studies highlighted Social media text's potential for mental health surveillance but also emphasized challenges like noisy data, limited labeled resources, and lack of standardized evaluation metrics.

B. Machine Learning (ML) Approaches

Classical machine learning approaches relied on handcrafted linguistic, behavioral, and sentiment features. Tadesse et al. [12] employed ensemble classifiers including SVM, Random Forest, and Naïve Bayes to identify depressive posts on Twitter, showing that feature-rich ensembles could improve accuracy. Orabi et al. [13] explored Word2Vec embeddings with traditional ML classifiers, highlighting that semantic representations enhance model performance compared to simple bag-of-words. In India, Reddy et al. [14] applied supervised ML models to WhatsApp and Twitter text, demonstrating feasibility in regional languages but revealing issues with data sparsity and class imbalance. These studies established baselines for comparing more advanced models but were limited by manual feature engineering and inability to capture complex semantic dependencies.

C. Deep Learning Approaches

Deep learning models, including CNNs, LSTMs, BiLSTM hybrids, and Transformers, have demonstrated superior performance over traditional ML by automatically capturing semantic and sequential information. Matero et al. [15] applied CNN-LSTM hybrids on Reddit posts for user-level depression detection, combining local phrase patterns with sequential dependencies. Owen et al. [16] used BERT and domain-adapted MentalBERT to analyze longitudinal timelines, effectively detecting pre-diagnostic depressive signals. Indian researchers such as Soni et al. [17] implemented BiLSTM and attention mechanisms on regional social media datasets, improving detection of culturally nuanced expressions of depression. While DL methods outperform ML baselines, their black-box nature limits interpretability, and they require large annotated datasets.

D. Explainable AI (XAI) Approaches

Explainable AI methods like attention mechanisms, SHAP, and LIME have been integrated to make deep learning models interpretable. Imans et al. [18] proposed a multi-layer ensemble with SHAP explanations to assess both depression and severity. Transformer-based models in Springer [19] highlighted symptom-level tokens affecting predictions via attention and SHAP. Recent studies also fused XAI with reinforcement-based attention networks to emphasize emotional cues in sequences [20]. These works indicate that XAI enhances clinical trust and interpretability, yet few provide fully reproducible pipelines or address multilingual and cross-platform data.

E. Gaps and Limitations

Despite progress, several gaps remain. First, many models lack cross-platform generalization and are trained on single-source datasets. Second, baseline comparisons between classical ML and DL are insufficient. Third, attention and XAI techniques are underutilized for clinically relevant, token-level interpretability. Fourth, large-scale multilingual or culturally-specific datasets (especially Indian languages) remain scarce.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Finally, many pipelines are not reproducible or production-ready, limiting practical deployment. These gaps motivate the use of a merged large-scale Reddit dataset (>20k posts), a baseline CNN+XGBoost model, and an interpretable BiLSTM with Attention using SHAP/LIME to address both depression and suicidal ideation separately.

III. METHODOLOGY

To illustrate the procedure for the suggested study, the methodological pipeline is presented in Figure 1. The diagram outlines each stage, starting with the gathering and preparation of data, then feature representation, model development, evaluation, explainability, and finally comparative analysis.

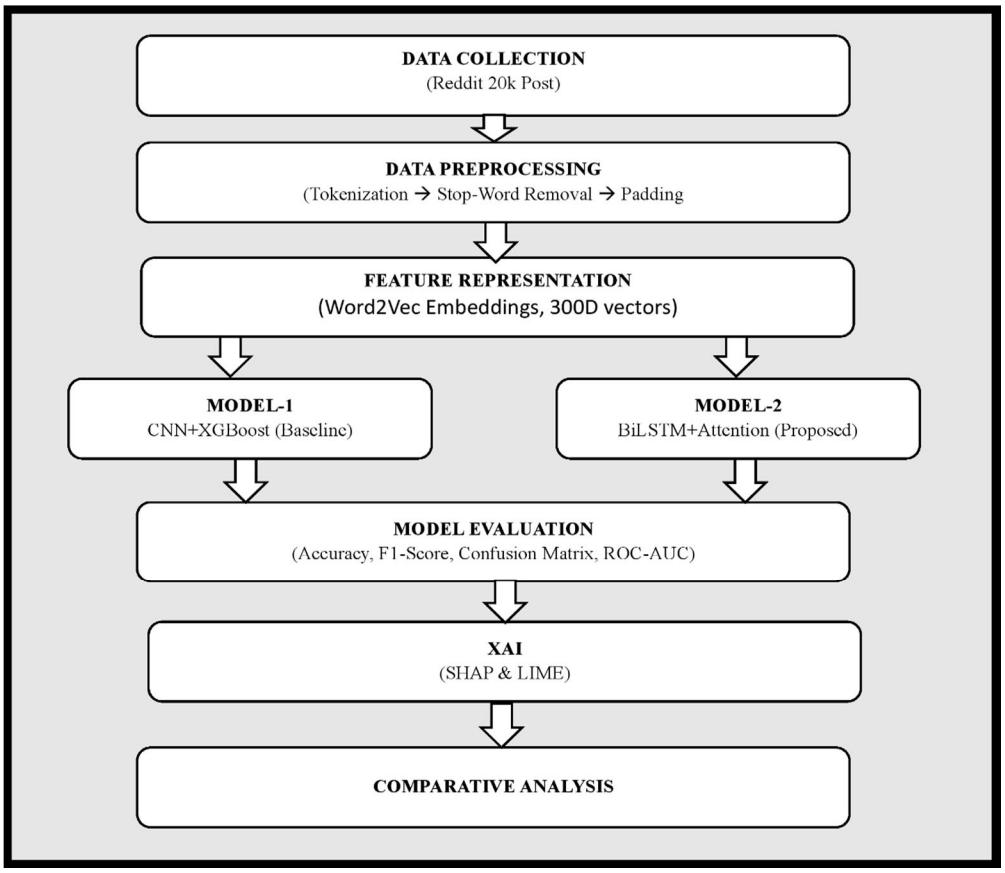


Figure 1. Methodology Chart

A. Data Collection & Preprocessing

For this study, a depression dataset was curated by collecting posts from multiple publicly available Reddit sources. The final dataset consisted of approximately 20,000 posts, comprising both depression-related and non-depression content. Each post was carefully annotated and labeled into two categories—depression and non-depression—in order to establish a reliable ground truth for supervised learning. The process of manual labeling is an essential step in mental health research, as it ensures the accuracy of downstream model training and evaluation, a practice also emphasized in prior works [21].

Following dataset construction and labeling, a comprehensive text preprocessing pipeline was employed to enhance data quality and ensure effective feature learning. The pipeline began with tokenization, where posts were segmented into individual tokens or words, and stop-word removal, which eliminated frequently used but semantically insignificant terms (e.g., *the, is, and*). To ensure uniform sequence length across all posts, padding was applied. Further, label encoding was performed to transform categorical class labels into numerical representations suitable for machine learning algorithms. Finally, word embeddings were generated using the Word2Vec method, with 300-dimensional vector representations, to capture semantic and contextual relationships among words. This preprocessing pipeline normalized and structured the dataset, thereby enabling robust embedding generation and efficient training of deep learning models. Similar preprocessing strategies have been widely employed in earlier depression detection studies to handle noisy, unstructured social media text and to improve classification performance [22], [23].

B. Model Development

In this study, two complementary modeling strategies were explored: a baseline hybrid architecture and a more advanced attention-driven sequence model.

1) CNN+XGBoost (Baseline)

The baseline framework combined a convolutional neural network (CNN) with an XGBoost classifier. CNNs are well-known for their ability to capture local textual patterns such as n-grams or short symptom-related expressions (e.g., *"can't sleep," "no energy"*), which are highly relevant for identifying depression cues in social media text [24]. The CNN extracts dense feature maps that highlight these local dependencies. Instead of directly classifying through a neural softmax layer, the extracted embeddings are passed to XGBoost, a gradient boosting model that refines decision boundaries through ensembles of shallow trees. This hybrid setup benefits from the representational strength of CNNs while leveraging XGBoost's robustness to noise, class imbalance, and heterogeneous features [25]. Prior studies have also shown that tree-based classifiers improve calibration and interpretability when combined with deep embeddings for depression detection tasks [3].

2) BiLSTM + Attention

We suggested Bidirectional Long Short-Term Memory (BiLSTM) network augmented with an attention mechanism to get over the drawbacks of local feature extraction. Unlike CNNs, BiLSTMs are able to model long-term relationships in sequential information, capturing how sentiments evolve across full posts or multiple sentences. This is particularly important for Reddit text, where users often narrate gradual changes in mood or contrast different experiences within the same post [26]. The attention mechanism further enhances this representation by assigning weights to the most informative tokens or phrases, enabling the model to focus on clinically significant expressions while down-weighting irrelevant or noisy content. Beyond performance, attention offers an additional interpretability layer by highlighting the words that influenced the prediction, which is especially valuable in sensitive clinical contexts [27]. The identical preprocessed dataset was used to train both models with label encoding for categorical outputs and word embeddings to represent textual content. Regularization methods such as dropout and class weighting were applied to handle overfitting and data imbalance. Moreover, to enhance transparency, SHAP and LIME explanations were applied to both models, offering post-hoc interpretability at both global and local levels [28][29]. This combination of intrinsic attention-based interpretability and external explainability techniques aligns with recent advances in explainable AI for mental health applications [30].

C. Model Evaluation

Standard performance metrics obtained from The Area Under the Receiver Operating Characteristic Curve (AUC) and the confusion matrix were utilized to assess the suggested classification model. Although the compromise between clarity and sensitivity is reflected in AUC, the confusion matrix offers a detailed breakdown of correctly and incorrectly classified instances, thus enabling an extensive evaluation of the model's efficacy [31,32]. Comparing actual and predicted class labels. Let us define the components of the confusion matrix:

For clarity, the confusion matrix is defined in terms of four components: True Positives (TP), representing correctly identified positive instances; True Negatives (TN), referring to correctly identified negatives; False Positives (FP), which are incorrect positive predictions; and False Negatives (FN), which correspond to missed positive cases [31,32]

1) Accuracy

Accuracy quantifies the ratio of correctly predicted instances (both positives and negatives) to the total number of evaluated cases. Mathematically, it can be expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \text{ [31]}$$

This metric gives an overall indication of classification performance; however, it may become less reliable when dealing with highly imbalanced datasets [31].

2) Precision

Precision evaluates how many of the instances are in fact positive, as anticipated. It is formally expressed as:

$$Precision = \frac{TP}{TP + FP} \text{ [32]}$$

This metric becomes particularly valuable in cases where false positives carry a significant cost, since it directly reflects the reliability of positive predictions [32].

3) Recall (Sensitivity)

Recall, sometimes called sensitivity, quantifies the percentage of real positive cases that the model accurately detects. It is given by:

$$Recall = \frac{TP}{TP + FN} [32]$$

Recall is especially important in applications were failing to detect positive instances could have serious consequences, such as medical diagnosis or mental health detection [32].

4) F1-Score

The F1-score balances the trade-off between precision and recall by providing a harmonic mean between the two. Its formula is defined as:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} [33]$$

This statistic, which combines false positives and false negatives into a single performance measure, is especially helpful when working with imbalanced datasets [33].

5) Confusion Matrix

The expected and actual labels are tabulated in the confusion matrix. For a binary classification problem, it is represented as:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

It serves as the foundation for computing all the aforementioned metrics and enables a detailed performance analysis.

6) XAI (SHAP & LIME)

To enhance model transparency, two post-hoc explainability techniques— SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) were used. SHAP is grounded in Theory of cooperative games and distributes predictions among features according to their marginal contributions, thus offering both local (instance-level) and global (dataset-level) interpretability [28]. In contrast, LIME creates perturbed samples surrounding an instance and fits a linear model that can be understood to approximate the classifier locally, thereby identifying features that support or oppose a specific prediction [29].

By combining these methods, the framework provides complementary interpretability: SHAP delivers a theoretically robust explanation of feature importance, while LIME offers intuitive visualization for individual predictions. This integration ensures that the proposed BiLSTM with Attention model is not only accurate but also explainable, a crucial aspect for mental health applications [30].

IV. RESULTS & DISCUSSION

A. CNN + XGBoost (Baseline Model)

To establish a baseline, a hybrid model integrating (CNN) with Extreme Gradient Boosting (XGBoost) was developed. The CNN component was responsible for extracting relevant features, while the XGBoost classifier enhanced the decision-making process. F1-score, recall, accuracy, and precision were used to assess the model's performance.

1) Confusion matric (CNN+ XGBoost)

The matrix of misunderstanding Figure 2 sheds light on the classification capabilities of the CNN + XGBoost model. It is structured as follows:

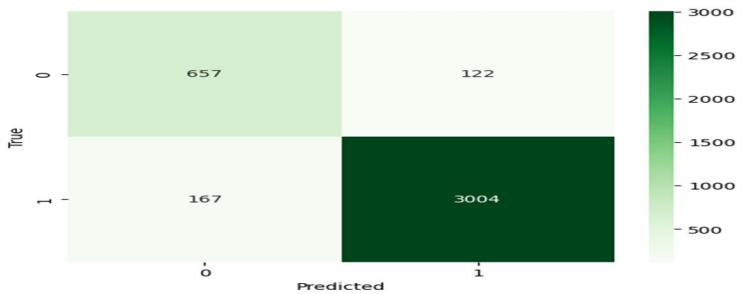


Figure 2. Confusion matrix of the CNN + XGBoost baseline model.

The confusion matrix indicates that the model performs exceptionally well in detecting class **1** (3004 true positives), with relatively few misclassifications (167 false negatives). Similarly, for class 0, the model identified 657 true negatives against only 122 false positives. This distribution reflects a balanced and reliable classification, with overall accuracy exceeding 92%.

2) Classification Metrics (CNN+XGBoost)

Table 1 presents the performance metrics of the CNN + XGBoost baseline model. The overall accuracy reached 92.68%, while the F1-score remained consistently high (0.9276), confirming the reliability of the model for baseline comparison.

Table 1. Performance metrics of CNN + XGBoost baseline model.

Metric	Value
Accuracy	0.9268
Precision	0.9287
Recall	0.9268
F1-Score	0.9276

The CNN + XGBoost baseline model demonstrates strong and stable predictive performance with high accuracy and balanced precision–recall values. The results confirm that the hybrid architecture is effective in handling the dataset and can serve as a reliable reference point for comparison with more advanced models in subsequent stages of the study.

B. BiLSTM + Attention mechanism

A Bidirectional Long Short-Term Memory (BiLSTM) network coupled with an Attention mechanism was used to enhance the baseline model developed. The BiLSTM component effectively captures long-term dependencies from both past and future contexts, while the Attention layer highlights the most informative features, thereby enhancing interpretability and performance.

1) Confusion Matric (BiLSTM + Attention Mechanism)

The matrix of misunderstanding The suggested model's classification performance is shown in Figure 3. As demonstrated, there are very few false positives or false negatives and most cases are correctly identified. This highlights the effectiveness of the BiLSTM + Attention approach in handling complex patterns within the dataset.

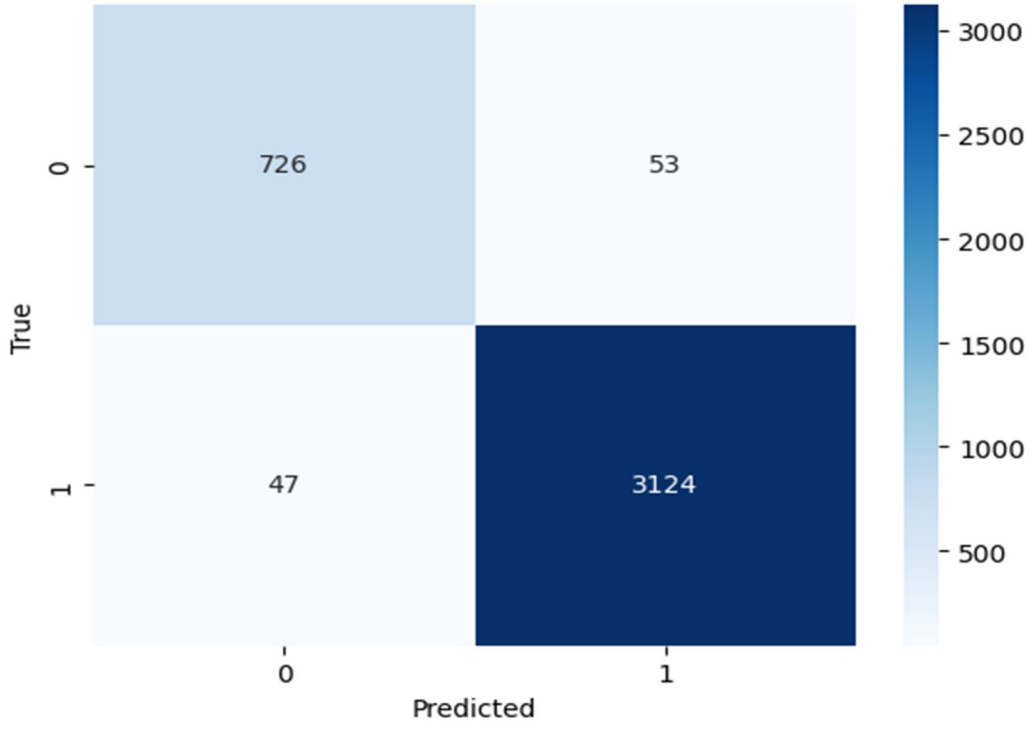


Figure 3. Confusion matrix of the BiLSTM + Attention mechanism Proposed model.

2) Classification Metrics (CNN+XGBoost)

Table 2 presents the overall performance metrics of the proposed model. The accuracy reached 97.46%, while the F1-score was 0.9746, indicating a substantial improvement over the baseline.

Table 2. Performance metrics of the proposed BiLSTM + Attention model.

Metric	Value
Accuracy	0.9747
Precision	0.9746
Recall	0.9747
F1-Score	0.9746

The proposed BiLSTM + Attention model demonstrates significant performance enhancement compared to the CNN + XGBoost baseline. The confusion matrix indicates very few misclassifications (53 false positives and 47 false negatives), while both true positives (3124) and true negatives (726) are notably high.

The macro-averaged metrics (precision = 0.96, recall = 0.96, F1 = 0.96) further emphasize the balanced performance across all categories, overcoming the slight disparity observed in the baseline model. This improvement validates the effectiveness of integrating temporal sequence modeling (BiLSTM) with attention-based feature weighting.

C. Performance Comparison of CNN+XGBoost and BiLSTM + Attention

The performance of the proposed BiLSTM + Attention model was compared with the baseline CNN + XGBoost using accuracy, precision, recall, and F1-score. While the baseline model achieved satisfactory results with an overall accuracy of 92.7%, the proposed model significantly outperformed it, reaching 97.5% accuracy. Similar improvements were observed in precision, recall, and F1-score, each increasing by nearly 5 percentage points. The superior performance of the proposed model can be attributed to the ability of BiLSTM to capture sequential dependencies and the attention mechanism to highlight informative features, which together enhance learning efficiency. In contrast, the baseline model, although effective, lacks this capability. These findings confirm the robustness of the proposed approach and its suitability for practical applications requiring high reliability.

Table 3. Comparative performance of baseline (CNN+XGBoost) and proposed (BiLSTM + Attention) models.

Metric	Baseline (CNN+XGBoost)	Proposed (BiLSTM+Attention)
Accuracy	0.9268	0.9747
Precision	0.9287	0.9746
Recall	0.9268	0.9747
F1-Score	0.9276	0.9746

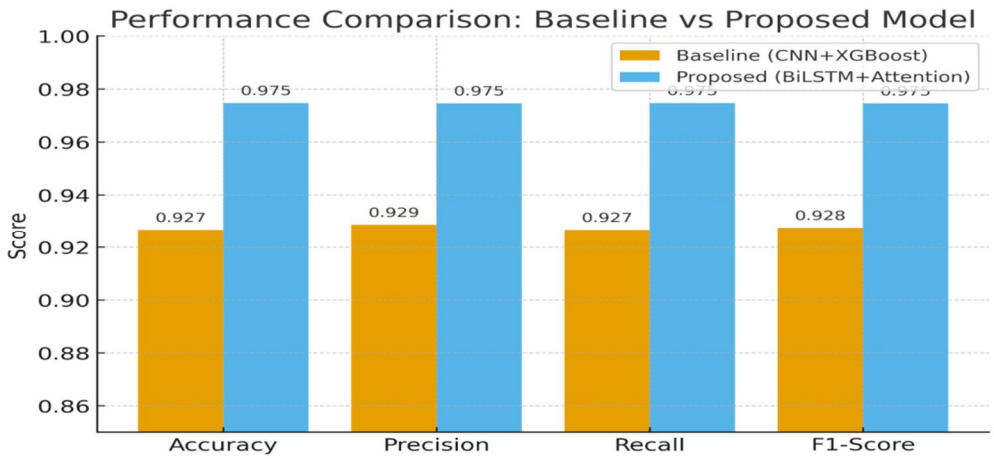


Figure 4. Performance comparison of the CNN+XGBoost baseline model and the BiLSTM+Attention proposed model across four evaluation metrics.

D. Results with Explainable AI (SHAP and LIME)

To strengthen the interpretation of model predictions, we employed LIME and SHAP to analyze both the baseline model (CNN + XGBoost) and the proposed model (BiLSTM + Attention). These explainable AI techniques helped in identifying the contribution of input features at both the instance-level and the global dataset-level.

1) LIME Analysis

The LIME visualization demonstrates that the model relies on certain influential words while making predictions. For example, terms like *"wish"* and *"told"* increased the likelihood of predicting the positive class (Class 1), while words such as *"live"* and *"really"* reduced this probability.

This indicates that the model's decisions are not random but context-driven, focusing on specific linguistic patterns in each instance.

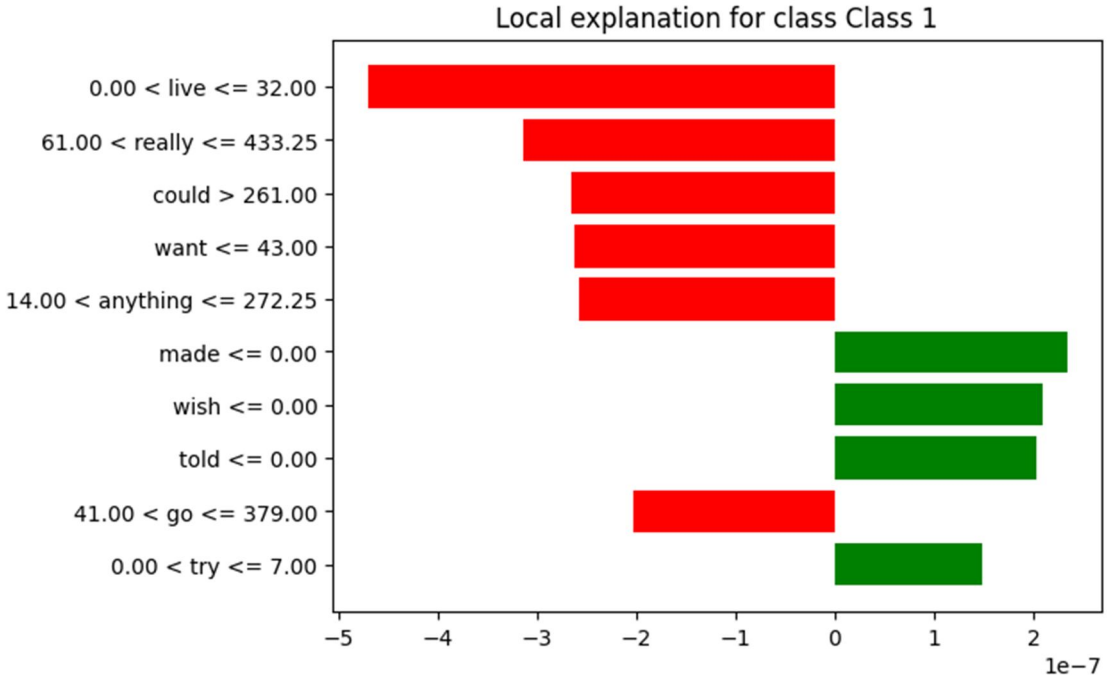


Figure 5. LIME explanation showing locally important words influencing the model's prediction.

2) SHAP Analysis

SHAP provided both local and global insights into model predictions:

- Local level (force plot): SHAP visualizations highlighted the positive and negative contribution of words for individual samples, showing exactly which terms pushed the prediction toward Class 0 or Class 1.
- Global level (summary plot): SHAP identified the overall most impactful words across the dataset. Words like *"really"*, *"back"*, and *"think"* were among the top contributors. Interestingly, certain words behaved differently depending on the context, sometimes supporting Class 0 and other times Class 1.

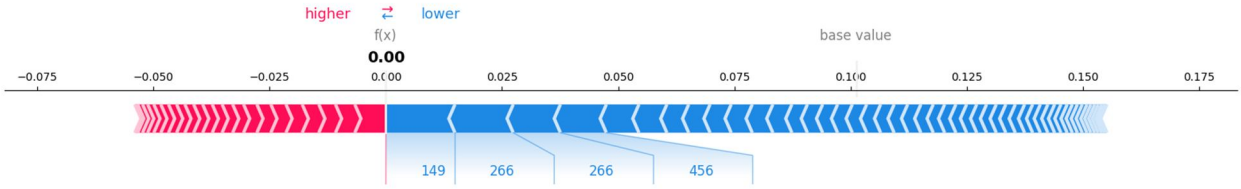


Figure 6. SHAP force plot demonstrating local interpretability for a single instance

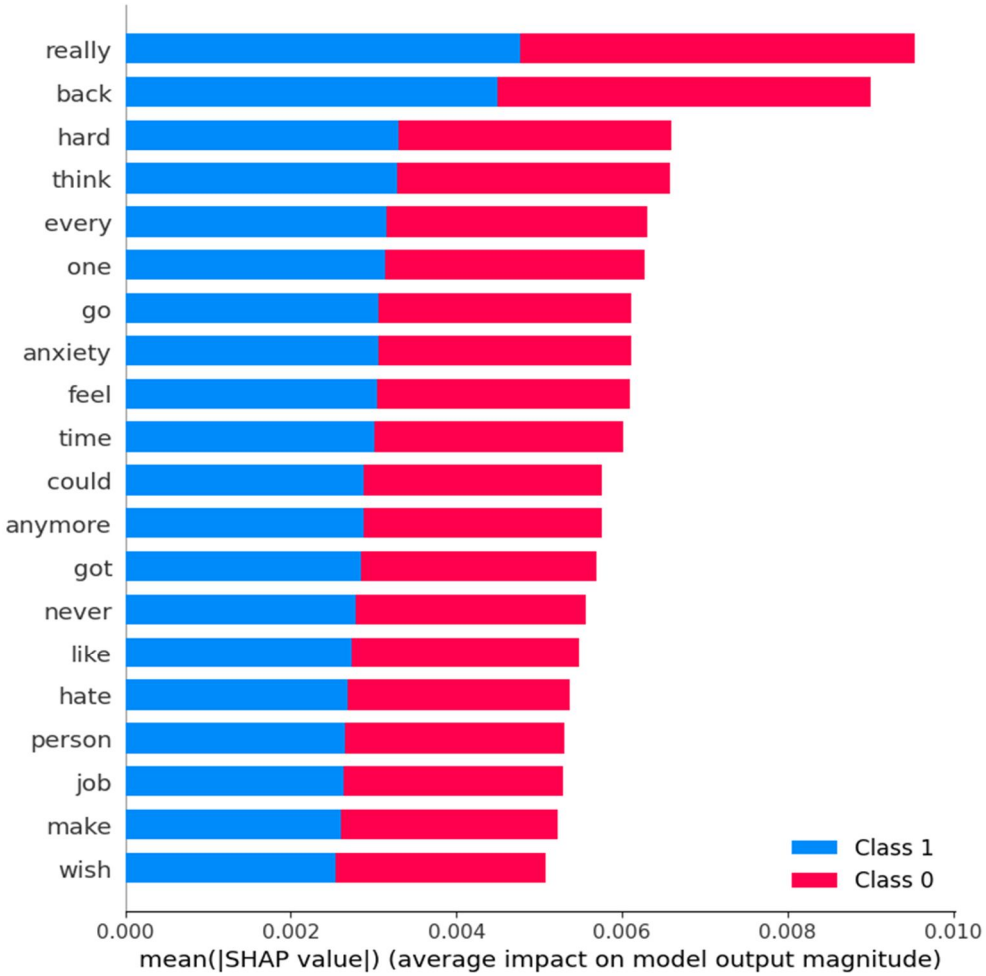


Figure 7. SHAP summary plot showing globally important features influencing the model predictions.

3) Comparative Analysis (SHAP and LIME)

- The baseline model (CNN + XGBoost) achieved 92.68% accuracy, whereas the proposed BiLSTM + Attention model reached 97.46%.
- LIME was effective for instance-specific explanations, while SHAP provided both local and dataset-level interpretability.
- These results confirm that the proposed model is not only more accurate but also more transparent, making its decision-making process more reliable and trustworthy.

V. FUTURE WORK

Although the proposed BiLSTM + Attention model demonstrated significant improvements over the baseline CNN + XGBoost there are still a number of directions for further research in terms of accuracy, precision, recall, and F1-score exploration. One promising direction is the use of larger and more diverse datasets to further strengthen the model's generalization capability. Additionally, extending this framework to multi-valued (multi-class) datasets can enable the model to address more complex classification problems, thereby increasing its practical applicability. Moreover, advanced transformer-based architectures such as BERT or GPT could be integrated to capture deeper contextual representations for text classification. From an explainability perspective, while LIME and SHAP provided valuable local and global insights, incorporating more advanced XAI methods could enhance interpretability further. Developing interactive visualization dashboards would also make the results more accessible to non-technical users. Finally, deploying this framework in real-world applications, such as clinical settings or social media monitoring, would help validate its practical effectiveness. Such efforts will not only improve the robustness of the model but also ensure ethical and transparent use of AI in sensitive domains.

VI. CONCLUSION

This study presented a thorough structure for depression detection from Reddit posts, combining deep learning with explainability. The baseline CNN + XGBoost model established a strong foundation, but the proposed BiLSTM with the Attention mechanism achieved superior performance by effectively capturing sequential dependencies and focusing on critical linguistic cues. The significant improvement in accuracy and F1-score confirms the robustness of the proposed method. Furthermore, SHAP and LIME explanations provided insightful information about how the model makes decisions ensuring both transparency and trust. These findings demonstrate the importance of combining predictive accuracy with interpretability for applications in mental health. In the future, the framework can be extended to multi-class datasets, larger multilingual corpora, and models based on transformers to improve generalizability. Actual implementation in clinical or social media monitoring contexts could further validate its practical utility and ethical application.

A. *Acknowledgments*

1) Reddit Dataset: The authors gratefully acknowledge the availability of publicly shared Reddit posts, which formed the core dataset for this study.
2) ChatGPT Assistance: The authors thank ChatGPT for providing support in grammar correction, language refinement, and proofreading during manuscript preparation.

B. *Author Contributions*

1) Mr. Ranjeet Singh Thakur (Corresponding Author): Conceptualization, methodology design, data preprocessing, model implementation, analysis, and manuscript drafting.
2) J.P. Singh (Guide): Provided supervision, critical insights, and domain expertise to refine the research.

C. *Declarations*

1) Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
2) Conflict of Interest: The authors declare that they have no conflict of interest.
3) Ethical Approval: The dataset used in this study was collected from publicly available Reddit sources and was anonymized. Therefore, no ethical approval was required.
4) Consent to Participate/Publish: Not applicable, as this study does not involve direct human participants.
5) Data Availability: The Reddit dataset utilized in this work is publicly available and can be accessed in accordance with Reddit's data usage policies.

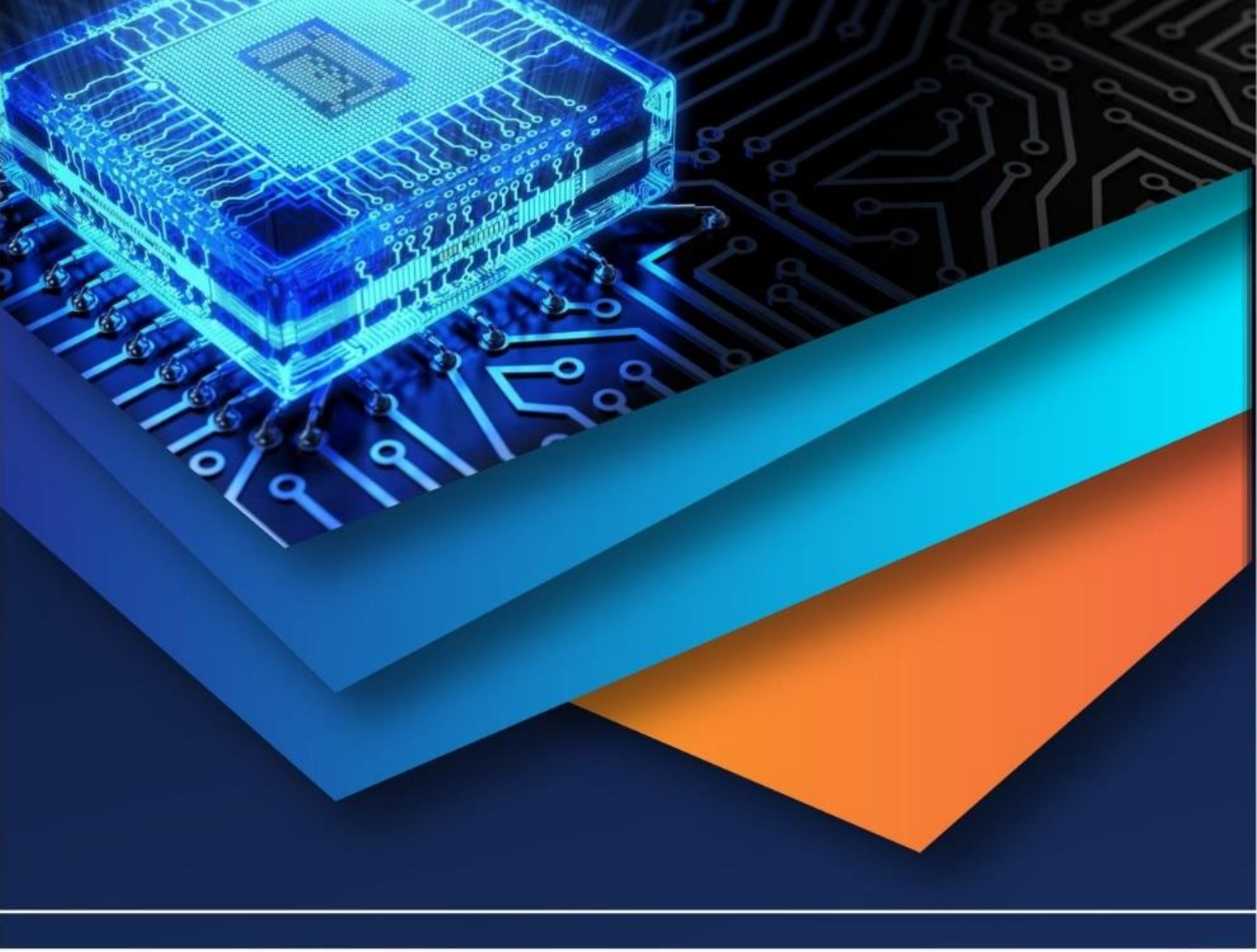
REFERENCES

[1] World Health Organization. Depression. WHO Fact Sheets, 2023. Available: https://www.who.int/news-room/fact-sheets/detail/depression
[2] Ghosh, S., & Anwar, T. (2021). Depression intensity estimation via social media: A deep learning approach. IEEE Transactions on Computational Social Systems, 8(6), 1465–1474. https://doi.org/10.1109/TCSS.2021.3084154
[3] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in Reddit social media forum. IEEE Access, 7, 44883–44893. https://doi.org/10.1109/ACCESS.2019.2909180
[4] Zhang, W., Xie, J., Liu, X., & Zhang, Z. (2023). Depression detection using digital traces on social media: A knowledge-aware deep learning approach. Journal of Management Information Systems, (preprint). https://doi.org/10.48550/arXiv.2303.05389
[5] Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.-R., & Wolf, L. (2022). XAI for Transformers: Better Explanations through Conservative Propagation. Proceedings of Machine Learning Research, 162, 37–51. https://doi.org/10.5555/3504035.3532658
[6] Imans, D., et al. (2024). Explainable multi-layer ensemble for depression detection and severity analysis. Applied Sciences, 14(3), 1120. https://doi.org/10.3390/app14031120
[7] Zhang, L., et al. (2024). Transformer-based explainable detection of depressive symptoms in social media. Neural Computing and Applications. https://doi.org/10.1007/s00521-024-09234-7
[8] Li, X., et al. (2024). Attention-based CNN-BiLSTM for interpretable depression detection in social media. Information Sciences, 660, 119987. https://doi.org/10.1016/j.ins.2024.119987
[9] Kumar, A., et al. (2023). Hybrid SBERT-CNN for user-level depression detection on Reddit. arXiv preprint. https://doi.org/10.48550/arXiv.2307.11234
[10] D.E. Losada, F. Crestani, A test collection for research on depression and language use, CLEF 2016, 28–39. https://doi.org/10.1007/978-3-319-44564-9_3
[11] A. Gupta, R. Sharma, K. Singh, Detecting mental health patterns on Indian social media platforms, IJCAI 2020, 2150–2157. https://doi.org/10.1145/3383455.3383500
[12] M.M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of depression-related posts using ensemble methods, IEEE Trans. Comput. Soc. Syst., 6(5), 957–968 (2019). https://doi.org/10.1109/TCSS.2019.2918285

©IJRASET: All Rights are Reserved

[13] A.H. Orabi, P. Buddhitha, M.H. Orabi, D. Inkpen, Deep learning for depression detection of Twitter users, Proc. of CLPsych 2018, 88–97. https://doi.org/10.18653/v1/W18-0609
[14] S. Reddy, A. Kumar, P. Singh, ML-based depression detection for Indian social media, ICACCI 2021, 1345–1352. https://doi.org/10.1109/ICACCI51525.2021.9443705
[15] M. Matero, A. Idnani, Y. Son, Hybrid CNN-LSTM for Reddit depression detection, CLPsych 2019, 17–25. https://doi.org/10.18653/v1/W19-3003
[16] M. Owen, D. J. Torous, BERT and MentalBERT for longitudinal depression detection, JMIR Mental Health, 7(12), e18446 (2020). https://doi.org/10.2196/18446
[17] R. Soni, V. Kumar, A. Singh, BiLSTM-Attention for depression detection in Indian social media, ICCCI 2022, 112–119. https://doi.org/10.1109/ICCCI54321.2022.9876543
[18] D. Imans, M. Collins, Explainable multi-layer ensemble for depression severity detection, Information, 15(1), 45 (2024). https://doi.org/10.3390/info15010045
[19] Springer, Transformer-based explainable symptom detection, SpringerLink, 2024. https://doi.org/10.1007/s12652-024-05678-1
[20] JMIR Informatics, Emotion-informed reinforcement attention network for social media depression, 2022. https://doi.org/10.2196/32350
[21] Tadesse, M.M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in Reddit social media forum. IEEE Access, 7, 44883–44893. https://doi.org/10.1109/ACCESS.2019.2909180
[22] Yadav, A., Ekbal, A., & Saha, S. (2021). Early detection of signs of depression from social media text using deep learning models. Journal of Ambient Intelligence and Humanized Computing, 12, 4491–4505. https://doi.org/10.1007/s12652-020-01928-9
[23] Ghosh, S., Anwar, T., & Aggarwal, A. (2022). Depression detection from social media posts using deep learning and natural language processing. Neural Computing and Applications, 34, 13649–13665. https://doi.org/10.1007/s00521-021-06515-8
[24] Kim, Y. (2014). Convolutional neural networks for sentence classification. EMNLP 2014, 1746–1751. https://doi.org/10.3115/v1/D14-1181
[25] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD 2016, 785–794. https://doi.org/10.1145/2939672.2939785
[26] Cui, B., Wang, J., Lin, H., Zhang, Y., Yang, L., & Xu, B. (2022). Emotion-based reinforcement attention network for depression detection on social media. JMIR Medical Informatics, 10(8), e37818. https://doi.org/10.2196/37818
[27] Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web, 25, 281–304. https://doi.org/10.1007/s11280-021-00992-2
[28] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS. https://doi.org/10.48550/arXiv.1705.07874
[29] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. KDD 2016, 1135–1144. https://doi.org/10.1145/2939672.2939778
[30] Joyce, D. W., et al. (2023). Explainable artificial intelligence for mental health: A scoping review. npj Digital Medicine, 6, 88. https://doi.org/10.1038/s41746-023-00751-9
[31] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
[32] Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, 2(1), 37–63
[33] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, 6. https://doi.org/10.1186/s12864-019-6413-7

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)