Explainable Self-Attentive Transformer Model for Bangla Mental Health Disorder Detection

Abu Saim Hossen Hridoy[1], Nazmus Sakib Shohan[1], Md. Shaharia Alif[1], Nurul Mursalin Ag Mahin[1], and Md. Ayon Mia[1*]

[1]Department of Computer Science and Engineering, Dhaka International University, Dhaka-1212, Bangladesh
{abu.saim.h.h, n.s.shohan.research, sahariaalif, nurulmursalinagmahin, mdayonrahman100*}@gmail.com

Abstract. Detecting mental health expressions in Bangla social media text remains a critical challenge, particularly in a rapidly digitalizing society where users increasingly express emotions and psychological distress online. We used the B-MHD (Bangla Mental Health Disorder Text) dataset, a manually annotated collection of 7,130 Bangla and BanglaEnglish code-mixed social media posts gathered from Facebook, YouTube, Twitter, and Reddit. While previous studies have explored traditional and transformer-based approaches for Bangla sentiment and depression detection, the integration of explainability into transformer architectures for mental health analysis remains underexplored. To address this gap, we conduct a systematic evaluation of classical, recurrent, and transformer-based models for Bangla mental health detection. As part of this evaluation, we employ various machine learning and deep learning modelsincluding SVM, Logistic Regression, BiLSTM, GRU, and LIME-based explanation analysisto investigate token-level contributions and model behaviour. Building upon these insights, this paper introduces one of the first explainable self-attentive transformer-based models designed specifically for Bangla mental-health text recognition, incorporating an attention and mean-pooling mechanism to enhance contextual understanding and interpretability. Experimental findings demonstrate that transformer-based models outperform traditional methods, with BanglaBERT combined with self-attention pooling achieving an F1-score of 97.04% and an accuracy of 96.98%. Through LIME-based explanation analysis, we further interpret token-level contributions, showing that emotionally expressive words strongly influence predictions, while metaphorical or contextually ambiguous phrases remain challenging. This study advances the development of explainable mental health detection systems for Bangla social media contexts.

Keywords: Bangla mental health detection, Transformer model, Self-attention, Social media text analysis, Explainable AI

1 Introduction

The increasing incidence of mental health disorders, including depression, anxiety, and stress, has become a worldwide issue. More than 300 billion people suffered from depression in 2017, an 18% surge over the last decade, as reported by the World Health

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_28

Organization (WHO) [10]. In Bangladesh, *The Daily Star* has reported that 4.6% of all adults and 1% of children have depressive symptoms [15]. With the rapid expansion of social media, individuals increasingly express emotions, frustrations, and life experiences online, creating an opportunity to study mental well-being through language patterns in digital spaces.

Previous studies across several languages have explored sentiment and emotion detection using lexicon-based, statistical, and machine-learning techniques [12]. More recently, deep-learning architectures such as RNNs and LSTMs have improved social media text analysis, outperforming earlier feature-based approaches [16]. Despite this progress, research on Bangla remains limited. Although Bangla ranks among the most widely spoken languages globally, mental health detection in Bangla social-media text is still underdeveloped. The frequent code-mixing between Bangla and English, transliteration, irregular spelling, and informal syntax further complicate computational processing and model generalization.

To address these issues, we present an explainable transformer-based framework that improves contextual understanding and interpretability in detecting mental health disorder expressions from Bangla and Bangla-English code-mixed social media text. Leveraging recent progress in neural architecture, we develop a hybrid model that integrates language representation with interpretation and explanation for emotionally complex and psychologically expressive contents. In addition, we perform extensive comparisons between classical, recurrent, and transformer-based approaches in order to ensure a consistent evaluation of the models' capabilities across architecture families. Our contributions are summarized as follows:

- Development of a self-attentive transformer model integrating attention and mean-pooling for effective representation of Bangla mental health expressions.
- Incorporation of LIME-based explainability to interpret token-level contributions and enhance transparency in model predictions.
- Comprehensive evaluation comparing classical ML, DL, and transformer architectures on the Bangla Mental Health Disorder (B-MHD) dataset.

2 Related Works

Early research in Bangla mental health detection primarily relied on traditional deep-learning architectures. One of the first studies [17], applied an LSTM model to classify depressive and non-depressive social media posts, demonstrating the potential of recurrent neural networks for emotion-based text analysis. Although this work provided an initial foundation, its small dataset and narrow focus limited generalization. The introduction of large-scale pre-trained models such as BERT [6] and RoBERTa [9] marked a major shift, enabling bidirectional contextual representation and improved transfer learning across tasks. These models have since become the basis for mental health detection in multiple languages, including Bangla, where their contextual embeddings outperform traditional feature-based methods.

Several Bangla-specific studies have expanded upon these advancements through clinical annotation and hybrid modeling. [8] developed a clinically validated depression detection dataset and compared classical algorithms (SVM, Random Forest, KNN) with

neural models (LSTM, GRU, CNN-RNN), where GRU achieved the highest performance. Subsequent work,such as TRABSA by [7] combined transformer encoders with BiLSTM and attention layers to retain sequential and contextual dependencies. The use of SHAP and LIME analyses in this study introduced early interpretability into Bangla emotion classification. Ensemble-based frameworks have also been explored; for example, [11] proposed an ensemble of XLM-R, DistilBERT, and BanglaBERT using a probability aggregation method (MaxOfAvgProb), improving consistency across datasets. Similarly, [2] presented a large-scale depression dataset and showed that BanglaBERT outperformed other RNN and transformer variants, emphasizing the benefit of language-specific pretraining.

Recent works have increasingly focused on improving interpretability and contextual understanding in low-resource mental health detection. [1] proposed a CNN-BiLSTM architecture with deep attention visualization, highlighting key emotional tokens contributing to predictions, while [4] investigated CNN-BiLSTM models using various embeddings (TF-IDF, FastText, and BERT), finding that BERT-based representations offered better contextual balance and stability. Collectively, these studies demonstrate a gradual evolution from early recurrent models toward transformer-driven and interpretable architectures, motivating the development of our explainable self-attentive transformer framework for robust and transparent Bangla mental health detection.

3 Dataset

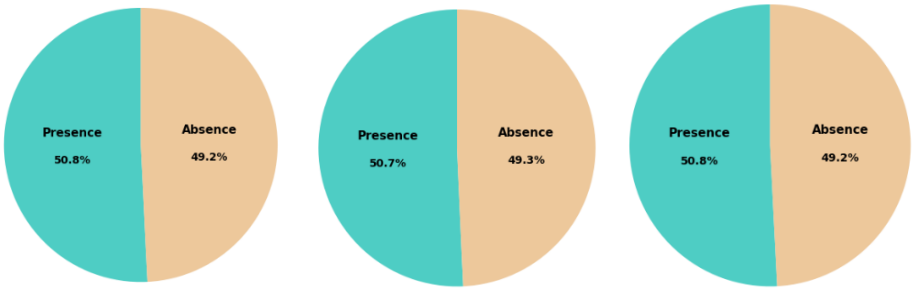
This study employs the B-MHD (Bangla Mental Health Disorder Text) dataset [14], a manually annotated collection of 7,130 Bangla and Bangla-English code-mixed social media posts gathered from Facebook, YouTube, Twitter, and Reddit.

The dataset was curated to capture diverse linguistic expressions of mental health conditions in informal online discourse. Each post is labeled according to the presence or absence of mental health indicators, such as expressions of depression, anxiety, loneliness, or emotional distress. Annotation was performed by native Bangla speakers with linguistic and psychological awareness, following a structured guideline that emphasized both lexical and contextual cues such as self-referential statements, affective adjectives, and expressions of hopelessness-to ensure

Dataset Statistics	
Min Character Length	11
Max Character Length	5784
Mean Character Length	476.43
Min Word Count	2
Max Word Count	1147
Mean Word Count	86.43
Unique Word Count	51362
Dataset split	**#Samples**
-Train	4991
-Val	712
-Test	1427
Total	7130

Table 1: General Statistics of the B-MHD Dataset.

reliable labeling. For data integrity, we pre-processed all records by deleting emojis, hashtags, bleeding URLs, and duplicated posts. The dataset was subsequently stratified into 70% training, 10% validation, and 20% test splits, maintaining the original class proportions to enable consistent evaluation across models, as illustrated in Figure 1. Table 1 shows the main characteristics of the dataset; text lengths (in characters) range from 11 to 5,784 and word counts per message for each post vary between 2 and 1,147 (mean 86). It comprises 51,362 unique tokens due to the diversity of vocabulary and its informal status in social media language, including transliteration and code-mixing.



(a) Train set (b) Validation set (c) Test set

Fig. 1: Class distribution across train, validation, and test splits in the B-MHD dataset, showing balanced representation of samples indicating the presence and absence of mental health indicators.

4 Methodology

Our proposed framework aims to detect mental health disorder-related expressions in Bangla social media text using the B-MHD dataset. Each input text sample x_T is represented and classified through a transformer-based encoder followed by a self-attention and pooling mechanism, as illustrated in Figure 2.

4.1 Text Representation.

Given a Bangla text input x_T, we employ a pretrained transformer based text encoder ϕ_T to obtain contextualized token embeddings:

$$H_T = \{t_{[\mathrm{CLS}]}, t_1, t_2, \ldots, t_n\} = \phi_T(x_T)$$

Here, $t_{[\mathrm{CLS}]}$ denotes the special classification token capturing the overall semantic representation of the text, while $t_1, \ldots, t_n$ represent contextualized token-level embeddings.

4.2 Self-Attention Layer.

To enhance the model's ability to focus on psychologically informative cues within the text, we apply a self-attention mechanism over the hidden representations:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where Q, K, V are learnable projection matrices derived from H_T, and d_k denotes the scaling factor. This operation emphasizes contextually important tokens such as those reflecting emotional distress, anxiety, or self-referential statements.

4.3 Mean Pooling.

The attention-refined representations are aggregated via mean pooling to produce a fixed-length vector h_T:

$$h_T = \text{Mean-Pooling}(A)$$

This pooled embedding captures the global contextual semantics of the input while smoothing over local variations.

4.4 Classification Head.

The final pooled vector h_T is passed through a classification head comprising a linear layer and a softmax activation function to predict the presence or absence of mental health disorder indicators:

$$\hat{y} = \text{Softmax}(W_c \cdot h_T + b)$$

where W_c and b are trainable parameters. The model is trained using the categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

with $C = 2$ representing the binary classes.

4.5 Experimental Setup

All experiments were conducted on the Kaggle platform using an NVIDIA Tesla P100 GPU with 16 GB VRAM, 32 GB RAM, and 8 CPU cores. We used BanglaBERT [3], mBERT [6], and XLM-RoBERTa [5] models for text encoding. Training was performed for up to *10* epochs with early stopping, a batch size of *16*, and a learning rate of *2e-5*, optimized using *AdamW*. The implementation utilized *PyTorch 2.4.0* and *HuggingFace Transformers 4.45.1*, with *NumPy 1.26.4*, *Pandas 2.2.3*, *Matplotlib 3.7.5*, *Seaborn 0.12.2*, and *scikit-learn 1.2.2* used for data preprocessing and analysis.

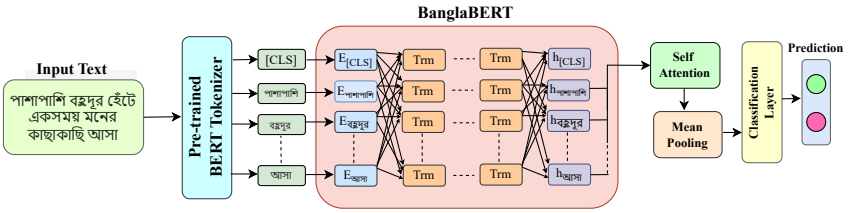


Fig. 2: Overview of the proposed method, where input text is tokenized and encoded by BanglaBERT, followed by a self-attention and mean-pooling layer, and finally a classification layer predicting the presence or absence of mental health indicators.

5 Result Analysis

The comparative performance of all evaluated models is summarized in Table 2. Results show a gradual improvement from traditional machine learning models to deep learning and transformer-based architectures. Classical methods employing Bag-of-Words and TF-IDF representations provided stable but limited performance, as they rely on surface-level lexical patterns and fail to capture deeper contextual meaning in code-mixed text. Deep learning approaches such as LSTM and BiLSTM achieved higher accuracy by modeling sequential dependencies, while the BiLSTM+CNN combination further improved contextual representation through joint learning of local and global features.
Transformer-based models demonstrated additional improvement due to their ability to encode contextual relationships and long-range dependencies. The inclusion of self-attention and mean-pooling layers consistently enhanced performance across all transformer variants, indicating that these mechanisms help the model concentrate on linguistically and psychologically salient tokens. Among them, the BanglaBERT model with the self-attentive pooling delivered the most reliable results, showing clear gains over its [CLS]-only version and the multilingual alternatives. These findings highlight the advantage of language-specific pretraining combined with attention-based feature refinement for detecting the presence or absence of mental health indicators in Bangla social media text.

6 Error Analysis

Quantitative Analysis. We performed a quantitative error analysis of the BanglaBERT + Self-Attention + Mean-Pooling model, which achieved the best overall performance on the B-MHD . The confusion matrix in Fig. 3 becomes evident that the adopted model preserves balanced prediction accuracy for both classes, which can accurately identify most samples, indicating and not-indicating mental health indicators. The misclassifications are few, since only a small fraction of the absence cases were predicted to be presence and vice versa, showing that the two categories have little overlap. This suggests that the model effectively distinguishes psychologically expressive language from general non-indicative content, although occasional errors may arise from short or context-ambiguous posts that lack clear emotional cues.

Models	Pr(%)	Re(%)	F1 (%)	Acc (%)
		ML-based Models		
LR + BoW	93.14	93.15	93.13	93.13
LR + TF-IDF	92.49	92.50	92.50	92.50
SVM + BoW	91.76	91.58	91.51	91.51
SVM + TF-IDF	93.19	93.20	93.20	93.20
XGB +BoW	92.22	92.23	92.22	92.22
XGB + TF-IDF	92.07	92.07	92.07	92.08
		DL-based Models		
LSTM	90.55	89.88	89.92	89.97
BiLSTM	93.94	93.88	93.89	93.90
GRU	91.76	91.58	91.51	91.51
BiGRU	93.27	93.28	93.27	93.27
BiLSTM+CNN	94.40	94.38	94.39	94.39
		Transformer-based Models		
XLM-RoBERTa				
+ [CLS]	97.00	94.88	95.43	95.42
+ Attention + Mean	**97.35**	96.55	96.95	96.91
mBERT				
+ [CLS]	92.10	94.10	93.00	93.12
+ Attention + Mean	93.92	96.00	94.95	94.81
BanglaBERT				
+ [CLS]	96.09	96.68	95.89	95.84
+ Attention + Mean (Ours)	96.84	**97.24**	**97.04**	**96.98**

Table 2: Comparison of ML, DL, and transformer models on the B-MHD dataset for Bangla mental health disorder detection. Results are reported in Precision (Pr), Recall (Re), F1-score (F1), and Accuracy (Acc). Best-performing scores are highlighted in **bold** and color.

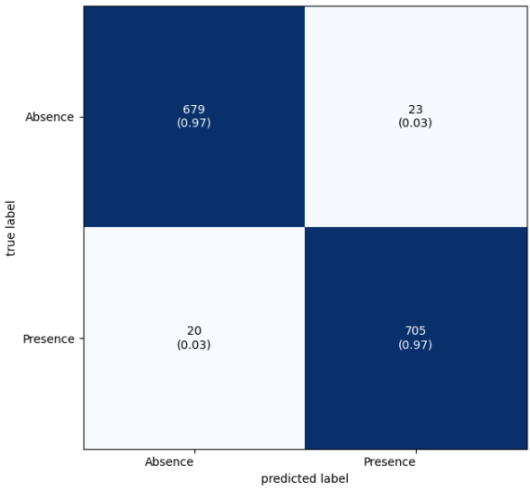


Fig. 3: Confusion matrix of the BanglaBERT model with self-attention and mean-pooling, showing balanced classification between the presence and absence of mental health indicators.

Qualitative Analysis. Examining representative cases in Table 3 reveals several recurring patterns in the proposed model's predictions. Posts explicitly expressing emotional distress or hopelessness are generally recognized correctly, indicating that the model effectively captures strong lexical and affective cues associated with mental health discourse. However, subtle expressions of sadness or indirect emotional cues are sometimes misidentified as absence cases, which indicates there may be challenges in recognizing implicit and metaphorical signs. On the other hand, such supportive or motivational sentences are also wrongly recognized as presence examples by the model, suggesting it may confuse empathetic or supportive language to self-referential mental health expressions. These confusions typically occur when affectionate language that is positive conflict with language related to mental health. Overall, these observations highlight that while the model performs reliably on explicit cues, it remains challenged by figurative, context-dependent, or emotionally neutral expressions that require a deeper pragmatic understanding of Bangla social media text.

Text	Actual	Predicted
আমাদের দেশের যে কোন দোকানে যেয়ে যদি আপনার আবদারের কথা বলেন, তারা সেটা খুশি মনেই মানে। (If you go to any shop in our country and make a request, they gladly agree to it.)	Absence	Absence
আমি একদম একা। সবসময় ডিপ্রেশনে থাকি। আমি মরে গেলেও হয়তো কেউ আফসোস করবে না, খুঁজ নিবে না। (Im completely alone. Im always in depression. Even if I die, maybe no one will regret it or come looking for me.)	Presence	Presence
আমি সহজে কোন কিছু ধরতে পারি না মাথা অনেক বেশি slow কাজ করে একটা বিষয়ের উপর focused থাকতে পারি না এই অবস্থা থেকে বের হতে আমি কি করতে পারি? (I cant grasp things easily; my mind works very slowly. I cant stay focused on one topic. What can I do to get out of this situation?)	Presence	Presence
আপনি নিজে আপনার জীবনের সমস্যাগুলি সমাধান করতে পারেন। (You are capable of solving your own life problems.)	Absence	Presence
অনেক ভালোবাসার পড়েও আমি আজ হেরে গেছি! (Even after giving so much love, I've lost today.)	Presence	Absence

Table 3: Examples of misclassifications produced by the proposed method on Bangla mental health detection, illustrating consistent error patterns across the two label categories.

7 Explainability Analysis

To gain interpretability into model behavior, we applied Local Interpretable Model-agnostic Explanations (LIME) [13] to examine token-level attribution patterns in the BanglaBERT + Self-Attention + Mean-Pooling model. LIME provides insight into how individual tokens contribute to the prediction of mental health indicator presence or absence by approximating the models local decision boundary with an interpretable surrogate. In Figure 4(a), the model correctly predicts the presence of mental health

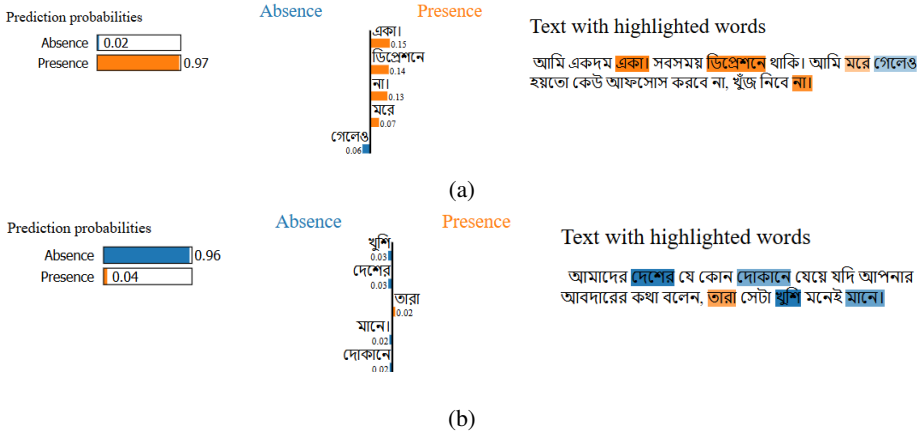


Fig. 4: LIME visualizations for Bangla mental health detection showing (a) correctly predicted presence and (b) correctly predicted absence of mental health indicators, demonstrating token-level contribution analysis.

indicators, with words such as একা (alone) and ডিপ্রেশনে (in depression) receiving the highest positive contributions toward the prediction. Figure 4(b) shows an absence case, where neutral words like দেশের (countrys) and দোকানে (shop) dominate, aligning with non-mental-health-related content. Figure 5 illustrates a misclassified instance where the token সমস্যাগুলি (problems) influenced the model toward predicting presence despite the sentence conveying a motivational tone. These visualizations reveal that while the model captures explicit emotional cues effectively, it sometimes misinterprets contextually neutral or metaphorical expressions, highlighting the complexity of semantic understanding in Bangla social media text.

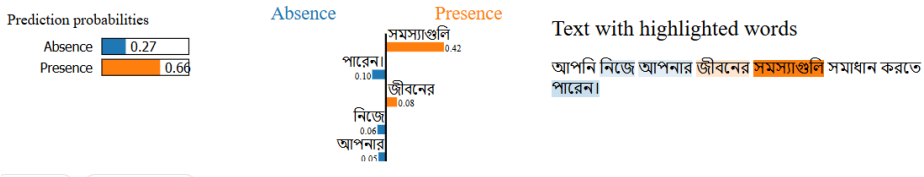


Fig. 5: LIME visualization showing a misclassified sample that led to predicting presence instead of absence.

8 Conclusion

We present a comprehensive comparative study of classical, recurrent, and transformer-based architectures for Bangla mental health detection in social media text. Our evalu-

ation shows that transformer models, particularly BanglaBERT with self-attention and mean-pooling, outperform traditional and recurrent approaches, achieving an F1-score of 97.04% and accuracy of 96.98%. The integration of LIME-based explainability further enhances interpretability, revealing that emotionally expressive and self-referential tokens play a decisive role in classification, whereas figurative or contextually ambiguous phrases often lead to misclassification. These insights highlight the effectiveness of combining language-specific pretraining with attention-driven interpretability for analyzing psychological expressions in Bangla social media discourse. In future work, we plan to explore large language models (LLMs) and advanced methodological frameworks, such as instruction-tuned and prompt-guided architectures, to further improve contextual understanding and interpretability.

Bibliography

[1] Absar, N., Islam, M.M., Somaya, Z.N.: Explainable depression detection from low-resource languages using cnn-bilstm with deep-attention mechanism. Machine Learning for Computational Science and Engineering **1**(2), 29 (2025)

[2] Ahmed, T., Hossain, M., Karim, M.: Bangla depression detection dataset and comparative analysis using transformer models. Data in Brief **55**, 110845 (2024)

[3] Bhattacharjee, A., Hasan, T., Ahmad, W.U., Samin, K., Islam, M.S., Iqbal, A., Rahman, M.S., Shahriyar, R.: Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. arXiv preprint arXiv:2101.00204 (2021)

[4] Chowdhury, S., Rahman, M., Ferdous, S.: Embedding-based cnn-bilstm model for bangla depressive text classification. Computers and Electrical Engineering **119**, 109113 (2025)

[5] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)

[6] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186 (2019)

[7] Jahin, M.M., Rahman, M.M.: Trabsa: Transformer-based robust sentiment analysis with bilstm and attention mechanism for explainable text classification. IEEE Access **12**, 45231–45243 (2024)

[8] Kabir, M.K., Islam, M., Kabir, A.N.B., Haque, A., Rhaman, M.K.: Detection of depression severity using bengali social-media posts on mental health: study using natural-language-processing techniques. JMIR Formative Research **6**(9), e36118 (2022)

[9] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[10] Organization, W.H.: Mental health: Depression (2017), http://www.who.int/mental_health/management/depression/en/, wHO Report

[11] Rahman, S., Haque, T., Akter, R.: Bangla depression intensity classification using transformer ensembles. arXiv preprint arXiv:2403.18111 (2024)

[12] Rani, S., Kumar, P., Sharma, R.: A review on emotion detection by using deep learning techniques. Artificial Intelligence Review (2024). https://doi.org/10.1007/s10462-024-10831-1, https://link.springer.com/article/10.1007/s10462-024-10831-1

[13] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

[14] Tasnim, F.: B-mhd: Bengali mental health disorder text dataset. Science Data Bank (2024), https://doi.org/10.57760/sciencedb.15744

[15] The Daily Star: Depression: Lets talk (2017), https://www.thedailystar.net/health/depressionlets-talk-1384978, published March 2017

[16] Tran, N., Ta, P., Nguyen, H., Nguyen, H.D.: Hybrid contextual and sentiment-based machine learning model for identifying depression risk in social media. Expert Systems with Applications **291**, 128505 (2025). https://doi.org/10.1016/j.eswa.2025.128505, https://www.sciencedirect.com/science/article/pii/S0957417425021244, online: 13 June 2025

[17] Uddin, A.H., Bapery, D., Arif, A.S.M.: Depression analysis from social-media data in bangla language using long short-term memory (lstm) recurrent neural-network technique. In: Proceedings of the International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). pp. 1–4. IEEE (2019)