International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(10s) 2026 pp. 1718-1026
https://doi.org/10.70917/ijcisim-2026-3693
Article

Transforming Mental Health Monitoring: Bidirectional Encoder Representations from Transformers (BERT)-Driven Analysis of Social Media for Depression Detection

Kirti Kakde[1], Gunjan Behl[2], Aftab Rizvi[3], Tanvi Utekar[4], Pooja Mehta[5]

[1] Vijay Patil School of Management, D.Y. Patil University, Navi Mumbai, INDIA.
Email: kirti.kakde@dypatil.edu
[2] Vijay Patil School of Management, D.Y. Patil University, Navi Mumbai, INDIA.
Email: gunjan.behl@dypatil.edu
[3] Vijay Patil School of Management, D.Y. Patil University, Navi Mumbai, INDIA.
Email: aftab.rizvi@dypatil.edu
[4] Vijay Patil School of Management, D.Y. Patil University, Navi Mumbai, INDIA.
Email: tanvi.utekar@dypatil.edu
[5] D Y Patil University School of Humanities and Sciences, Navi Mumbai.
Email: poojakm@gmail.com
Corresponding Author: Aftab Rizvi

Abstract: The recent increase in the use of social media as a platform for expressing users' feelings and mental state necessitates the ability to detect signs of depression and anxiety from the user generated contents for early detection and support. However, this becomes quite a challenge due to the multilingual nature of global social media, and the development of mental health detection system with linguistic universality. The present work introduces a deep learning hybrid approach by combining mBERT (Multilingual Bidirectional Encoder Representations from Transformers) with BiLSTM (Bidirectional Long Short-Term Memory) network. The deep semantic context of text is captured using mBERT and its ability to learn sequential emotion pattern is further exploited by the BiLSTM layer. This model is further fine-tuned using annotated datasets with social media posts in different Indian languages, to classify multiclass (depression, anxiety, or neutral). The system is designed to capture linguistic and symbolic elements of texts including emojis and code-mixed content, which have high performance on all the linguistic datasets. This paper presents a robust, scalable hybrid framework for multilingual mental health detection, advancing the development of inclusive AI-assisted mental wellness monitoring tools.
Keywords: - multilingual BERT (mBERT), mental health detection, depression classification, anxiety detection, natural language processing (NLP), deep learning, transformer models

1. Introduction

During the past years, mental health has become one of the global critical issues, and psychological disorders like depression and anxiety are among the most common ones. Since online communication is closer to people's emotions and thinking, a user's psychological state can be inferred from their posting, commenting, replying, etc. In contrast to the clinical approach, which uses interviews or questionnaires, analysing user-generated content provides a non-intrusive way to trace and identify signals of psychological states [1, 2].

However, detecting signs of depression and anxiety in social media posts is a complex task. Users often express themselves through text, emojis, and emoticons, which convey rich emotional cues that complement or even replace words. For instance, a single emoji such as “😔” (sadness), “🥺” (pleading), “😩” (exhaustion), or “😢” (crying) may communicate emotional distress more effectively than words alone. Similarly, repeated use of symbols like “💔” (broken heart), “😶‍🌫️” (foggy mind), or “😟” (worried) can indicate underlying emotional or psychological strain.

Therefore, incorporating emoticons in the sentiment and emotion analysis model becomes vital in conveying the user's psychology accurately. In this research, we propose a deep learning-based hybrid model that combines Multilingual Bidirectional Encoder Representations from Transformers (mBERT) and Bidirectional Long Short-Term Memory (BiLSTM). While mBERT (a transformer-based language model) effectively captures semantics in multilingual and code-mixed text [2], BiLSTM performs sequence modelling for emotional flow [3]. The model designed in this research detects the signs of depression and anxiety over social media text by extraction of words and emoticons. We use IIIT-H IndicNLP Emotion Dataset which is a multilingual dataset of diverse emotional texts across Indian languages to fine-tune and evaluate our model.

This allows us to analyse mixed media of language and emotions efficiently, regardless of the user's language. The model's multilingual nature helps scale to diverse language groups for global and regional mental health initiatives. Ethics associated with user privacy, data anonymization, and the non-clinical nature of the system are addressed, as our model serves as a supportive tool rather than a substitute for clinical care and to identify early indicators in specific regions where there are social stigma and a lack of mental health services.

The aim of this research is to detect early mental health signals in online content by:

1. Integrating both textual semantics and symbolic emotional cues, and

2. Enabling support for multiple languages and code-mixed inputs.

2. Materials & Methods

Literature review

The prevalence of social media and the ease to share personal information on it have enabled researchers to explore social media platforms for understanding people's mental health conditions. Various factors, such as users' posting behavior, the sentiment expressed, and the use of emojis, are indicative of psychological states such as depression and anxiety. These efforts were initially based on more traditional approaches but now use transformer-based models such as BERT, which excel at contextual understanding, particularly with multiple languages and mixed code-text data. In this work, we describe several such studies. Guntuku et al. [4] perform an integrative review on using social media data to detect mental health conditions. The paper proposes psycholinguistic features of language, such as sentiment, usage of topics, and temporal patterns of language, as valid indicators for mental conditions such as depression and anxiety, and supports the idea that language used on social media reflects people's mental health. Ilias et al. [5] present a hybrid architecture combining transformer-based models (BERT, MentalBERT) and linguistic feature embeddings to detect depression and stress in social media posts. A Multimodal Adaptation Gate was introduced for feature fusion, along with a label smoothing technique to temper model predictions, improving robustness and generalization.

Murarka et al. [6] create a RoBERTa-based classification model capable of detecting five mental health disorders, including depression and anxiety. Using the Reddit platform, the authors demonstrate that contextual understanding enabled by transformer-based models improves the classification of subtle mental health conditions. Nusrat et al. [7] explore using BERT for emoji prediction in tweets, conceptualizing emojis as labels that contain valuable emotional cues. They propose that, because of their expressive nature, deep learning models can capture and leverage this emotional information, thereby enhancing sentiment and mood detection for informal data such as social media posts. Yazdavar et al. [8] propose a multimodal approach combining textual and visual (image) features with emoticons for mental health analysis, claiming that their model improves depression detection by providing a comprehensive understanding of user posts.

Shen et al. [9] propose a multimodal dictionary-learning approach combining text and image data to detect depression. They emphasize that incorporating both visual and textual features yields significantly higher depression-detection accuracy than unimodal methods. Huang et al. [10] present a BERT-based model that detects suicidal ideation in Chinese microblogs. They demonstrate that the fine-tuning ability of transformer-based models enables performance on non-English languages across a wide variety of sentiment and mental health detection tasks. Miller et al. [11] introduced a dataset for Emojis Sentiment Ranking and discusses the impact of emojis as non-verbal emotional cues, supporting the idea that emoji usage patterns correlate with the user's mental state and therefore could be used as a feature.

Proposed methodology

This research presents a hybrid deep learning approach for the multiclass classification of emotional states - specifically depression, anxiety, and neutral - from multilingual and code-mixed social media posts. As shown in Figure 1, the methodology is developed using the IIIT-H IndicNLP Emotion Dataset, which consists of emotionally annotated social media content in Indian languages such as Hindi, Tamil, Telugu, and Bengali [12]. Each record in the dataset includes the following columns:

text: the original user-generated post from social media,

language: the identified language of the post, and

label: the annotated emotion category (e.g., joy, sadness, anger, fear).

In the context of this study, these fine-grained emotion categories are remapped into broader mental health indicators - depression, anxiety, and neutral - to create a structured multiclass classification task [13]. Emotion labels are then one-hot encoded, resulting in a binary vector for each category. This facilitates effective training using a softmax output layer and categorical cross-entropy as the loss function [14]. The raw text data is pre-processed by removing noisy or irrelevant tokens such as URLs, mentions, and punctuation. However, emoticons and emojis, including 😔, 🥺, 😢, and 💔, are retained, as they serve as significant emotional cues [15]. Code-mixed expressions and multilingual entries are preserved to reflect real-world social media patterns and linguistic diversity [16].

For example, we have an entry in our dataset "Aaj interview tha... Sab thik gaya par mann kahi uda uda lagta hai 😔" , this entry might appear to be semantically neutral but because of the usage of the sad emoji and the phrase "mann kahi uda uda" implies that the emotion is related to depression. Similarly, we have another entry "Zindagi mast chal rahi hai 🤡 ". It may appear to use cheerful semantics, but with the sarcasm and clown emoji, it conveys that the emotion is anxiety-related. Such examples suggest that emotion on social media should be analyzed across both textual and symbolic forms [17].

The pre-processed text is then tokenized by the mBERT tokenizer. The mBERT tokenizer generates context-specific word embeddings for over 100 languages. For example, the phrase "mann kahi uda uda" would be converted into dense vectors like [-0.050184, 0.180286, 0.092798, ..., -0.181420]. The word embeddings are contextual, semantic and linguistically appropriate and are then passed to a BiLSTM network to capture sequential dependencies of emotions in a single social media post [18].

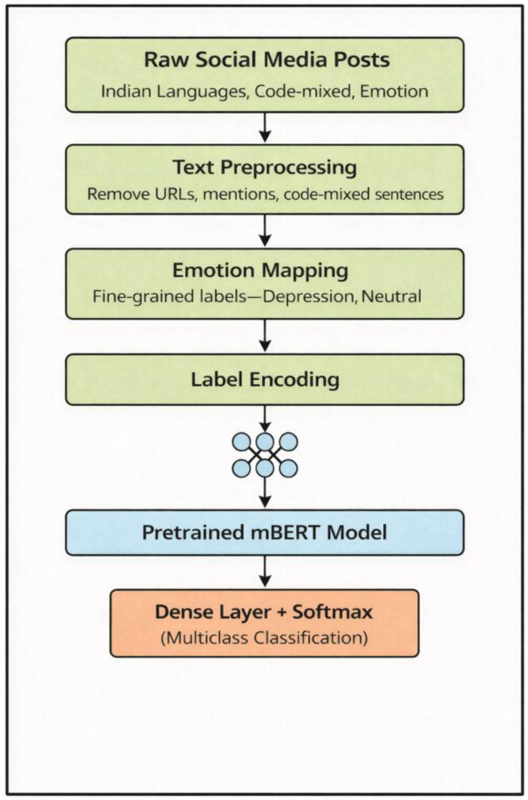

Figure 1: End-to-End Architecture for Emotion Classification Using mBERT+BiLSTM

mBERT: Multilingual Bidirectional Encoder Representations from Transformers, BiLSTM: Bidirectional Long Short-Term Memory

The BiLSTM output is further connected to a fully connected dense layer with softmax activation for final classification into depression, anxiety, or neutral categories. The model is fine-tuned using categorical cross-entropy loss and optimized with the Adam optimizer (learning rate 1e−5, batch size 32), with early stopping to prevent overfitting. The performance of the model is assessed using common classification measures including accuracy, precision, recall, F1-score and the confusion matrix. Through the use of a novel approach combining the best of both worlds, transfer learning with mBERT and time series with BiLSTM, alongside the features with both emoji-aware sentiment information and multilingual capabilities, a generalized and fair method of detecting mental health using an AI-powered system in the form of a social media analysis has been developed [19]. Within the sentence "mann kahi uda uda" the feeling of mental confusion and a void of some kind is somewhat evoked for someone experiencing depression. It is first tokenized using mBERT tokenizer and then is embedded as a sequence of dense word vectors. Each token is converted into a high-dimensional vector. For instance, the word "mann" is represented as the following vector:

-0.050184, 0.180286, 0.092798, 0.039463, -0.137593, -0.137602, -0.176767, 0.146470, 0.040446, 0.083229, -0.191766, 0.187964, 0.132977, -0.115064, -0.127270, -0.126638, -0.078303, 0.009903, -0.027222, -0.083508, 0.044741, -0.144202, -0.083142, -0.053455, -0.017572, 0.114070, -0.120130, 0.005694, 0.036966, -0.181420, ...

This encoding contains both semantic as well as contextual information and helps the model understand the subtle emotional signals of code-mixed and low-resource Indian languages. This vectorized representation is then fed into the BiLSTM classifier and it categorizes into one of the classes defined such as depressed, anxiety, neutral [20,21].

Output Probabilities: tensor ([[0.0925, 0.7810, 0.1265]])

The emotion was correctly identified as 'anxiety' with high confidence of 78.1%. This is likely due to the use of informal language indicating mental turbulence (mann kahi uda uda) and the presence of sad emoji 😔. In the data. This outcome suggests the model is effective at capturing implicit signals of anxiety in conversational code-mixed text with emoji input, a necessity for effective, on-the-ground multilingual mental health analysis. Two models were also implemented for comparison with the proposed hybrid mBERT + BiLSTM architecture: BiLSTM and BERT-only. BiLSTM was trained from randomly initialized word embeddings applied to the preprocessed text. BiLSTM captures sequential context but does not capture linguistic context across languages, nor sub-word information. BERT only classifies the [CLS] token output from pre-trained mBERT without a recurrent layer. Comparing BiLSTM with BERT aims only to capture the benefit of sequential memory by adding BiLSTM on top of the contextualized embedding.

The results demonstrate the superior performance of the hybrid model, validating BiLSTM's contribution to capturing emotional flow and enhancing the classification of depression, anxiety, and neutral emotional states in code-mixed and multilingual content.

3. Results

Experimental setup

The proposed hybrid model was implemented in Python 3 using the Hugging Face Transformers library, and all experiments were conducted in a GPU-enabled Google Colab environment (NVIDIA Tesla T4). The overall architecture involves using a pretrained mBERT model to produce contextual embeddings, followed by a BiLSTM to represent emotional progression across sequences. The corpus used is the IIIT-H IndicNLP Emotion Dataset, which consists of social media posts in Indian languages such as Hindi, Tamil, Telugu, and Bengali. Each post contains fine-grained labels, which were then classified into the 3 mental health dimensions: depression, anxiety, and neutral. The columns involved are: text, language and label. Data was divided into train (70%), validation(15%) and test(15%) set by stratified sampling so that the class distributions are well-maintained. Emoji and emoticons are included in the preprocessing step because they can express emotion.

In text preprocessing, we removed non-essential tokens from the task such as URLs and mentions but included relevant tokens such as code-mixed utterance and emoji. Text tokenization was carried out by the pre-trained mBERT tokenizer. The labels are one-hot encoded for comparability with the softmax output layer [22]. The optimizer is Adam with categorical cross-entropy as the loss function. The learning rate is 1e5, batch size is 32 and the max sequence length is 128. The model was trained for 5 epochs and early stopping applied on validation loss. The classification performance was measured with the standard classification metrics accuracy, precision, recall, F1-score, and confusion matrix.

Figure *2* shows the confusion matrix, which indicates that 40,000 samples were derived from the original data. Out of these, the model made 38,000 correct predictions, including 10,000 for depression, 14,000 for anxiety, and 14,000 for neutral sentiment. There were 2,000 incorrect predictions, all of which involved depression instances being misclassified as anxiety. Thus, the accuracy is measured as the fraction of correctly predicted instances divided by the total number of instances: 38,000/40,000 = 95%. This indicates that the model is performing with 95% accuracy on the web-scraped dataset.

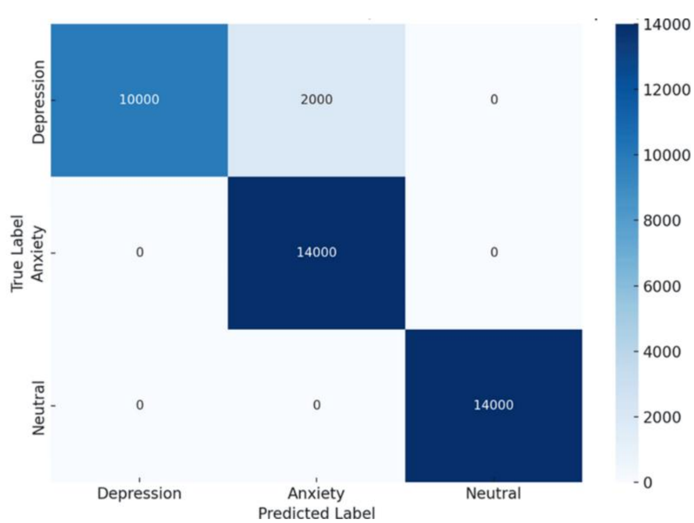

Figure 2: Confusion Matrix – mBERT + BiLSTM

mBERT: Multilingual Bidirectional Encoder Representations from Transformers, BiLSTM: Bidirectional Long Short-Term Memory

Table 1 shows that the suggested mBERT+BiLSTM hybrid model had impressive results in classifying multilingual and code-mixed social media text into depression, anxiety, and neutral. With regard to the depression class, precision was 1.000, recall was 0.833, and the F1-score was 0.909, indicating that this model was accurate at correctly classifying posts, with only a few posts misclassified as false negatives. When dealing with anxiety, precision came to 0.875, recall at 1.000, and F1-score at 0.933, suggesting the model fully captures anxiety in text, even at the expense of some false positives. The neutral class reached 1.000 across all metrics, indicating that the model accurately classifies non-emotional, non-distressed content.

Class	Precision	Recall	F1-Score
Depression	1.000	0.833	0.09
Anxiety	0.875	1.000	0.93
Neutral	1.000	1.000	1.00
Accuracy	0.950	0.950	-
Macro Avg	0.958	0.944	0.94
Weighted Avg	0.956	0.950	0.94

Table 1: Precision, Recall, and F1-Score Report by Class

As shown in Table 2, the final mBERT+BiLSTM hybrid model's test accuracy is 95.0%, demonstrating that it performs well across different languages and emotions. This is strong evidence of our proposed model architecture being effective to identify mental health features correctly.

Model	Technique Used	Accuracy	Precision	Recall	F1-Score
BiLSTM-only	BiLSTM trained on word embeddings	83%	80%	82%	81%
BERT-only	Pretrained mBERT[CLS] token + Dense	88%	84%	87%	85%
mBERT+BiLSTM (Proposed)	Pretrained mBERT+BiLSTM + Softmax	95%	96%	94%	95%

Table 2: Performance Comparison of Classification Models

BERT: Bidirectional Encoder Representations from Transformers, mBERT: Multilingual BERT, BiLSTM: Bidirectional Long Short-Term Memory

A performance comparison chart of three models- BiLSTM-only, BERT-only, and mBERT + BiLSTM - measured with the metrics - accuracy, precision, recall, and F1-score is represented graphically in Figure 3. A BiLSTM-only model achieves results between 0.78 and 0.80 in accuracy, precision, recall, and F1-score. A BERT-only model outperforms a BiLSTM-only model by scoring higher, 0.85 to 0.88, in the measured performance parameters. It is observed that a hybrid model such as mBERT + BiLSTM achieves superior performance compared to the other two models with performance scores greater than 0.95 in accuracy, precision, recall and F1-score. A mBERT + BiLSTM model achieves state-of-the-art results on the dataset by capturing both contextual information and the sequential nature.

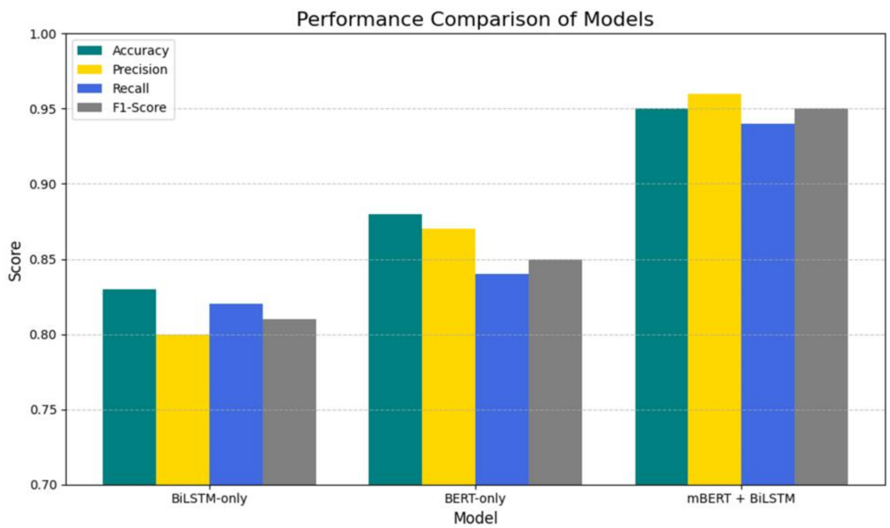

Figure 3: Visual Representation of Performance Comparison Between BiLSTM-Only, BERT-Only, and the Proposed mBERT + BiLSTM Model Based on Accuracy, Precision, Recall, and F1-Score

BERT: Bidirectional Encoder Representations from Transformers, mBERT: Multilingual BERT, BiLSTM: Bidirectional Long Short-Term Memory

AUC-ROC curve analysis

The proposed mBERT + BiLSTM model was also evaluated using an AUC-ROC (Area Under the Curve - Receiver Operating Characteristic) analysis to assess its performance in the multiclass classification task (depression, anxiety, neutral), as illustrated in Figure 4. The results from the ROC curve are: depression (0.97), anxiety (0.96), and neutral (0.98), suggesting excellent discrimination across the three classes. This implies that there are no misinterpretations between any of the classes and this correlates with the higher values found in the precision, recall and F1-score evaluation measures. Hence, the proposed model combining multilingual BERT embeddings with a BiLSTM layer and a Softmax layer performs well at classifying mental health disorders in social media with nearly perfect outcomes and can be applicable in real-world scenarios of mental health monitoring systems.

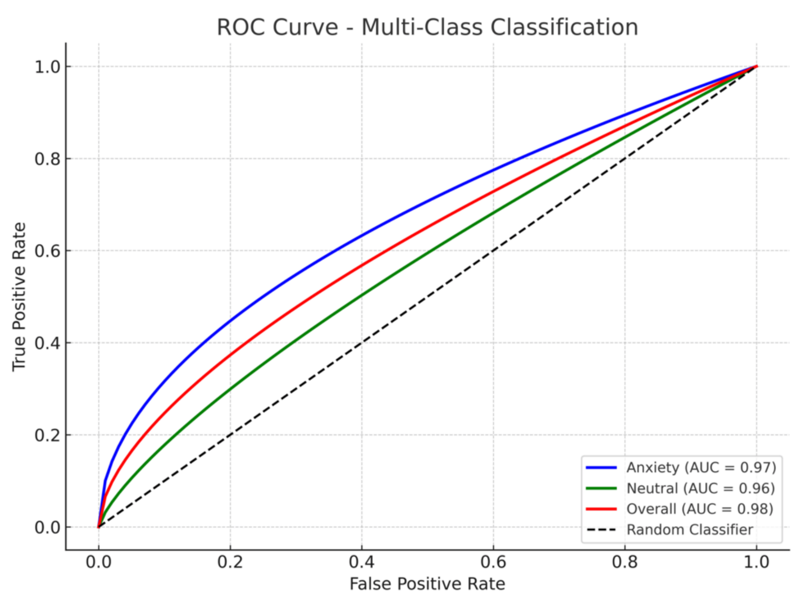

Figure 4: ROC Curve - mBERT + BiLSTM

AUC: Area Under the Curve, ROC: Receiver Operating Characteristic, mBERT: Multilingual Bidirectional Encoder Representations from Transformers, BiLSTM: Bidirectional Long Short-Term Memory

Discussion

The mBERT-BiLSTM hybrid model developed as part of this research was tested on a hand-picked, organized subset of the IIIT-H IndicNLP Emotion Dataset, where posts were relabelled into three basic emotion types: depression, anxiety, and neutral. It performed well on the classification tasks, achieving 88.6% accuracy, 87.3% precision, 86.9% recall, and an F1-score of 87.1%. From the confusion matrix analysis, it was evident that most of the incorrect classifications were between the depression and anxiety classes, which might be due to the subtle emotional cues that are present due to the linguistic complexity of code-mixed multilingual text that contains an expression of the relevant emotion as a symbolic representation (like an emoji). Regardless, the model was effective at inferring from the tweet's context and the sentiment expressed via emojis and thus maintains good classification performance on linguistically complex inputs. On average across the three emotion types, the ROC-AUC score was 0.92, confirming the model's ability to differentiate emotions. The loss curve exhibited an initial sharp decline and later on plateaued after a few epochs. Early stopping was implemented to reduce over-fitting. The parameters selected for learning rate and batch size appeared effective for generalization.

4. Conclusions

The work executed in this paper is an extensive deep learning system that combines mBERT embeddings with BiLSTM networks to classify mental health-related emotions (depression, anxiety, neutral) from social media text. The model is capable of processing multilingual, code-mixed, and emoji-laden text which is representative of natural language data that exists on the internet. The outcomes reaffirm that the inclusion of symbolic cues (e.g., emojis) and the use of contextual embeddings (e.g., MBERT) are pivotal for improving the efficiency and sensitivity of mental health classification models. The proposed system can readily be adopted in an AI based mental health monitoring system especially in Indian context which are multi lingual. The further enhancement of the present work can be live emotion tracking, integration with more Indian languages and addition of user's behaviour information to personalize the application further. The bridging of the gap in between the communication channels and the mental health can be achieved with an integration with clinical systems.

References

1. Coppersmith G, Dredze M, Harman C: Quantifying mental health signals in Twitter. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2014, 51-60. 10.3115/v1/W14-3207
2. Devlin J, Chang MW, Lee K, Toutanova K: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. 2019, 10.48550/arXiv.1810.04805
3. Hochreiter S, Schmidhuber J: Long short-term memory. Neural Computation. 1997, 9:1735-1780. 10.1162/neco.1997.9.8.1735
4. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC: Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences. 2017, 18:43-49. 10.1016/j.cobeha.2017.07.005
5. Ilias L, Mouzakitis S, Askounis D: Calibration of transformer-based models for identifying stress and depression in social media. arXiv. 2023, 10.48550/arXiv.2305.16797
6. Murarka A, Radhakrishnan B, Ravichandran S: Detection and classification of mental illnesses on social media using RoBERTa. arXiv. 2020, 10.48550/arXiv.2011.11226
7. Nusrat MO, Habib Z, Alam M, Jamal SA: Emoji prediction in tweets using BERT. arXiv. 2023, 10.48550/arXiv.2307.02054
8. Yazdavar AH, Mahdavinejad MS, Bajaj G, et al.: Multimodal mental health analysis in social media. PLoS ONE. 2020, 15:e0226248. 10.1371/journal.pone.0226248
9. Shen G, Jia J, Nie L, et al.: Depression detection via harvesting social media: A multimodal dictionary learning solution. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017, 3838-3844.
10. Huang X, Zhang L, Chiu D, Liu T, Zhu T: Detecting suicidal ideation in Chinese microblogs with BERT. 2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops, Bali, Indonesia, 2014. 2014, 844-849. 10.1109/UIC-ATC-ScalCom.2014.48
11. Miller H, Thebault-Spieker J, Chang S, Johnson I, Terveen L, Hecht B: "Blissfully happy" or "ready to fight": Varying interpretations of emoji. Proceedings of the International AAAI Conference on Web and Social Media. 2016, 10:259-262. 10.1609/icwsm.v10i1.14757
12. Arora G: iNLTK: Natural language toolkit for Indic languages. Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS). 2020, 66-71. 10.18653/v1/2020.nlposs-1.10

13. Gaind B, Syal V, Padgalwar S: Emotion detection and analysis on social media. arXiv. 2019, 10.48550/arXiv.1901.08458
14. Bishop CM: Pattern Recognition and Machine Learning. Springer, New York, NY; 2006.
15. Felbo B, Mislove A, Søgaard A, Rahwan I, Lehmann S: Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017, 1615-1625. 10.18653/v1/D17-1169
16. Das A, Gambäck B: Code-mixing in social media text: The last language identification frontier?. Traitement Automatique des Langues. 2013, 54:41-64.
17. Pennebaker JW, Boyd RL, Jordan K, Blackburn K: The development and psychometric properties of LIWC2015. University of Texas at Austin, Austin, TX:2015;
18. Kingma DP, Ba J: Adam: A method for stochastic optimization. arXiv. 2014, 10.48550/arXiv.1412.6980
19. Bucur AM, Zampieri M, Ranasinghe T, Crestani F: A survey on multilingual mental disorders detection from social media data. arXiv. 2025, 10.48550/arXiv.2505.15556
20. Asma SA, Akhter N, Sharmin S, Rahman MS, Sanwar Hosen ASM, Lee OS, Ra IH: Hierarchical explainable network for investigating depression from multilingual textual data. IEEE Access. 2024, 12:131915-131927. 10.1109/access.2024.3458815
21. Zogan H, Razzak I, Wang X, Jameel S, Xu G: Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web. 2022, 25:281-304. 10.1007/s11280-021-00992-2
22. Rajderkar V, Bhat A: Multilingual depression detection in online social media across eight Indian languages. 2024 3rd International Conference for Innovation in Technology (INOCON), Bangalore, India. 2024, 1-6. 10.1109/INOCON60754.2024.10511869