

ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
<http://dx.doi.org/10.70102/afts.2026.1836.963>

MULTILINGUAL XLM-R FOR MENTAL HEALTH SIGNAL DETECTION IN LOW-RESOURCE TAMIL LANGUAGE THROUGH CROSS-LINGUAL AND BILINGUAL LEARNING

Deivanai Gurusamy^{1*}, Midhunchakkaravarthy Janarthanan²

^{1*}Research Scholar, School of AI Computing and Multimedia, Lincoln University College, Selangor, Malaysia. e-mail: deivanai@lincoln.edu.my,
orcid: <https://orcid.org/0000-0002-5028-2128>

²Associate Professor, School of AI Computing and Multimedia, Lincoln University College, Selangor, Malaysia. e-mail: midhun@lincoln.edu.my,
orcid: <https://orcid.org/0000-0002-0107-885X>

Received: June 15, 2026; Revised: July 27, 2026; Accepted: August 31, 2026; Published: September 29, 2026

SUMMARY

Mental-health NLP (Natural Language Processing) remains strongly English-centric, limiting language-inclusive digital health technologies for low-resource languages such as Tamil. This study investigates whether multilingual representation learning can transfer a specialised mental-health classification task to Tamil and whether bilingual English–Tamil training improves performance over Tamil-specific learning. To experiment this, two publicly available mental-health datasets were combined into four categories: Normal, Anxiety, Depression, and Suicidal Ideation. English samples were translated into Tamil, with 1,000 translations manually validated. Eight experimental setups compared TF-IDF-based classifiers and XLM-R models across English, Tamil, cross-lingual transfer, and bilingual training configurations. Accuracy, macro-F1, class-wise performance, confusion patterns, paired predictions, and explainability were analysed. McNemar's test compared Tamil-specific and multilingual XLM-R predictions. Tamil-specific XLM-R achieved 73.20% [mean $\pm$ SD] accuracy and 0.7166 [mean $\pm$ SD] macro-F1, whereas Multilingual XLM-R achieved 73.67% [mean $\pm$ SD] accuracy and 0.7231, respectively. English-only cross-lingual transfer produced lower performance (64.78% accuracy; 0.6110 macro-F1), indicating that multilingual pretraining alone did not ensure effective transfer of the specialised task to Tamil. Bilingual training showed a modest numerical improvement over Tamil-specific training, including a 4.38-percentage-point increase in Suicidal Ideation-class recall, but the overall difference was not statistically significant ($\chi^2 = 2.113$, $p = 0.1461$). Explainability analysis showed that some highly attributed tokens were linguistically ambiguous or difficult to interpret. Multilingual language capability does not necessarily ensure effective specialised mental-health NLP for low-resource Tamil. Tamil task-specific learning substantially improved cross-lingual performance, while bilingual training provided only a modest additional benefit. The findings highlight the importance of language-specific adaptation and rigorous cross-lingual evaluation for inclusive mental-health information-processing systems.

Key words: *mental health, tamil language, XLM-R, explainable artificial intelligence, early detection, social media.*

INTRODUCTION

Mental health conditions remain a major global public-health concern, while access to appropriate care is constrained by stigma, discrimination, workforce shortages, and uneven availability of services [1] [2]. Consequently, conventional healthcare pathways may fail to capture distress that is expressed indirectly or does not reach formal services. Digital communication, including social media and other user-generated content, provides a substantial source of mental-health-related language that can be analysed using NLP (natural language processing) and machine-learning methods for screening, monitoring, and prevention-oriented research [3] [4]. These approaches can process large volumes of textual information consistently, creating opportunities for population-level analysis and earlier identification of potential mental-health-related signals. Despite this potential, mental-health NLP remains strongly concentrated on English, while multilingual resources are comparatively limited [4] [5]. A recent survey identified 108 datasets spanning 25 languages, but found that low- and mid-resource languages remain underrepresented, and that depression-focused datasets dominate multilingual mental-health research [5] [16]. Linguistic and cultural differences may also influence how individuals' express emotions, symptoms, distress, and help-seeking behaviours online, potentially affecting the generalisability and fairness of NLP systems [3] [6] [15]. These limitations make Tamil a relevant low-resource setting for examining whether multilingual language models can support specialised mental-health text analysis beyond high-resource languages. This study does not aim to replace clinical diagnosis. Rather, it evaluates computational methods for identifying mental-health-related textual signals that could support future screening, referral, or prevention-oriented systems when deployed with appropriate human oversight [2] [4]. Multilingual transformer models provide a basis for investigating cross-language learning in this context. XLM-R (XLM-RoBERTa) was pretrained on 100 languages and demonstrated strong performance across cross-lingual benchmarks, including for several relatively low-resource languages [7]. However, multilingual pretraining does not necessarily guarantee reliable transfer for specialised downstream tasks, and transfer performance can vary according to language, task, and the availability of language-specific information [8]. These considerations are particularly important for mental-health NLP, where linguistic and cultural differences can influence the expression and interpretation of psychological distress and self-disclosure in online communication [3], [5]. Against this background, the present study evaluates eight experimental configurations comprising conventional machine-learning baselines, XLM-R models trained on English and Tamil, English-to-Tamil cross-lingual transfer, and combined English–Tamil bilingual training. The experiments are designed to distinguish the effects of language-specific learning from cross-lingual transfer and bilingual training under a common Tamil test setting. The study further applies Layer Integrated Gradients to examine token- and subword-level contributions to selected model predictions. Specifically, it investigates whether language-specific and bilingual XLM-R approaches can improve the detection of four mental-health-related textual signal categories—Normal, Anxiety, Depression, and Suicidal Ideation—in Tamil. The findings are intended to provide empirical evidence on multilingual information processing for low-resource Tamil and to inform the future development of language-accessible, explainable, and human-supervised digital mental-health applications.

Research Contributions

Theoretical Framework: To create a robust theoretical framework for cross-lingual transfer evaluation of specialized high-risk domains' texts in an agglutinative low resource language (Tamil). We theoretically explain and empirically show that it is not guaranteed that general multilingual pre-training leads to semantic transfer to a specific domain.

Empirical & Statistical Validity: By showing through paired prediction and McNemar's significant test for statistics ($p = 0.1461$) that even though there are some benefits of task adaptation on language-specific adaptation (+8.42%), there are only minimal gains for bilingual joint training over language-specific adaptation.

Subword Attributions Analysis: With the use of Layer Integrated Gradients (LIG), demonstrate that successful prediction does not equate to human-readable subword attributions in low-resource multilingual models, showing the important shortcomings of current approaches to XAI in cross-lingual

digital health applications.

Research Significance. The importance of this research is that it deals with the problem of insufficient effectiveness of mental health signal detection techniques for low-resource languages, especially Tamil, due to the relatively poor availability of linguistic resources, annotated data, and language models in those languages. The study explores multilingual XLM-R in terms of Tamil-specific, cross-lingual, and bilingual learning paradigms and offers insights into the application of cross-lingual knowledge transfer for improved mental health signal detection in low-resource contexts. This study increases the validity of its results with the help of comparative testing, re-running of experiments, and statistical analysis of the model's performance.

The rest of this paper is structured as follows. Section II describes the related works concerning multilingual and cross-lingual approaches for mental health signal detection with a focus on low resource languages. Section III provides details regarding the materials and methods which include dataset preparation, preprocessing, experimental setup, multilingual XLM-R architecture, training process, evaluation metrics, and statistical testing. Section IV explains the experimental results and comparative evaluation of the suggested models including Tamil specific, cross-lingual, and bilingual XLM-R architectures. Section V analyzes the results and model performance compared to earlier work. Finally, Section VI concludes the paper.

RELATED WORK

Mental-Health NLP and Social-Media Analysis

Natural language processing and machine learning have increasingly been applied to social-media and user-generated text to identify mental-health-related signals, including depression, anxiety, and suicidal ideation [2] [4] [6]. These methods enable large-scale analysis of linguistic patterns and have supported research on mental-health prediction and early-risk detection [3] [4] [14]. However, differences in datasets, annotation schemes, populations, and task definitions make comparisons across studies difficult. Social-media text is also noisy and context-dependent, while privacy, representativeness, and the distinction between computational signal detection and clinical diagnosis remain important concerns [3] [5] [6].

Multilingual and Low-Resource Mental-Health NLP

Previous multilingual mental-health NLP research has shown that performance is not uniform across languages and that available resources differ substantially in size, annotation quality, and task coverage [5] [13]. Research on multilingual encoders has similarly shown that cross-lingual representations do not necessarily produce equivalent downstream performance across languages, with language-specific information remaining important for some tasks [7] [8] [12]. Although XLM-R provides strong multilingual representations through large-scale cross-lingual pretraining, these findings suggest that multilingual transfer should be evaluated for the specific target language and task rather than assumed to be effective. For specialised mental-health classification, language-specific expressions and contextual patterns may influence how textual signals are represented across languages. However, evidence directly comparing English-to-Tamil transfer with Tamil-specific and English-Tamil bilingual training remains limited. This gap motivates a controlled evaluation of these alternative learning strategies for Tamil mental-health text classification.

Explainable AI for Mental-Health Text Analysis

Explainability is particularly relevant when NLP models are applied to sensitive mental-health information. Recent research has used methods such as SHAP and LIME to examine feature contributions in mental-health prediction from social-media behaviour [9]. SHAP provides feature-attribution estimates based on cooperative game-theoretic principles, while LIME explains individual predictions through locally interpretable approximations [9] [10] [11]. However, attribution methods should not automatically be interpreted as evidence that highly weighted tokens are clinically meaningful or causally responsible for predictions. This is especially relevant to multilingual and

low-resource text, where tokenisation, spelling variation, and linguistic ambiguity may affect model attribution. Against this background, the present study compares English-only cross-lingual transfer, Tamil-specific learning, and combined English–Tamil bilingual training using a common Tamil test set. It further examines selected model predictions using Layer Integrated Gradients to assess token- and subword-level attribution patterns. Table 1 presents related studies and explains how the present study differs from them.

MATERIALS AND METHODS

Study Design

Table 1. Representative related studies and distinction of the present study

Existing Work	Objective	Data/Language Scope	Contribution	Difference from Present Study
[5]	Multilingual mental-disorder detection	Surveys multilingual resources and identifies research gaps	108 datasets across 25 languages	Maps the multilingual mental-health dataset landscape and resource gaps
[7]	Multilingual representation learning	Develops a multilingual pretrained language model	100 languages	Introduces XLM-R for cross-lingual representation learning
[9]	Explainable mental-health prediction	Predicts mental-health outcomes using explainable ML	Social-media behavioural data	Applies SHAP and LIME for model interpretation
[12]	Fine-grained depression signs	Identifies diverse depression-related expressions	Social-media posts from Brazilian undergraduate students	Evaluates 21 fine-grained depression signs
[13]	Depression-related language	Models for linguistic changes before and after depression diagnosis	Social-media posts from users before and after diagnosis	Introduces a dataset and evaluates multiple predictive models
[14]	Comorbid mental-health classification	Evaluates models for depression–anxiety comorbidity	Social-media posts; English	Introduces the multi-label benchmark and evaluates models from Mental-BERT to GPT-4
[15]	Suicidal-ideation evidence extraction	Extracts and summarizes evidence of suicidal ideation	Social-media content; English	Evaluates LLM-based evidence extraction and summarization
[16]	Suicide-related information extraction	Reviews AI/NLP approaches for identifying social determinants related to suicide	Suicide-related literature and computational studies	Highlights limitations in social-determinant information and representation of minority groups
Present study	Mental-health-related textual signal detection	Compares language-specific, cross-lingual, and bilingual learning strategies	English and Tamil	Describes a methodology for assessing the transfer deficit of pretrained encoder models (XLM-R) for use in specialized fields with limited resources (mental health in Tamil language), by conducting empirical paired significance tests (McNemar’s test) along with Layer Integrated Gradients (LIG).

This study used a comparative experimental design to evaluate conventional machine-learning and multilingual transformer approaches for four-category mental-health-related text classification in English and Tamil. The workflow of the study is given in figure 1. Eight experimental configurations

(E1–E8) were organised to compare conventional classifiers, XLM-R-based language-specific training, English-to-Tamil cross-lingual transfer, and combined English–Tamil bilingual training. The experimental configurations were designed to distinguish the contribution of language-specific task adaptation from cross-lingual and bilingual learning. The English and Tamil datasets were derived from the same underlying samples, and corresponding English–Tamil pairs were assigned jointly to the training, validation, or test partition. This ensured that a translated counterpart of a sample used for testing was not included in the training partition. The Tamil test set was identical across E6, E7, and E8, enabling direct paired comparison of Tamil-specific, cross-lingual, and bilingual XLM-R configuration.

Data Sources and Corpus Preparation

Two publicly available mental-health datasets were used: Sentiment Analysis for Mental Health [17] and Mental Health Risk & Emotion Detection Data (NLP) [18], both obtained from Kaggle. The datasets were standardised using common fields for textual content and class labels before being combined into a unified corpus. Only samples corresponding to four target categories - Normal, Anxiety, Depression, and Suicidal—were retained. Records belonging to other source categories were excluded. These categories were treated as computational classification targets rather than clinically confirmed diagnoses. In particular, the Suicidal category represents textual patterns associated with the corresponding dataset label and does not indicate an individual's clinical status or immediate suicide risk. Duplicate records and samples with missing text or labels were removed. Text pre-processing included lowercasing, removal of URLs, usernames, redundant spaces, and dataset-specific tokens. The resulting corpus contained 50,463 English-language samples. The original class distribution was retained without resampling, and class imbalance was quantified using the imbalance ratio, defined as the majority-class size divided by the minority-class size [19] [20]. The corpus was stratified and divided into 70% training, 15% validation, and 15% test sets, resulting in 35,324, 7,569, and 7,570 samples, respectively. The source-data field was removed after corpus preparation, and the target labels were encoded as Normal = 0, Anxiety = 1, Depression = 2, and Suicidal Ideation = 3. The class distribution across the resulting partitions is reported in table 2.

English-to-Tamil Translation and Validation

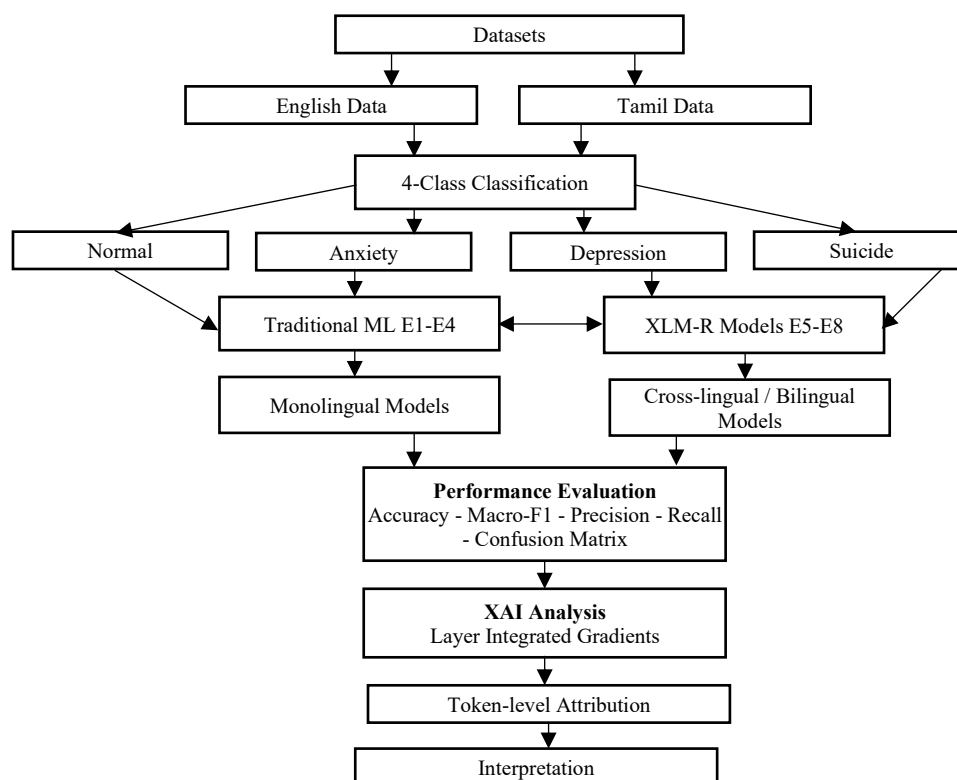


Figure 1. The study workflow

The English corpus was translated into Tamil using Google Translate to construct the Tamil-language corpus. Translation was performed before the train, validation, and test partitions were established, and each Tamil translation was retained with its corresponding English source sample. Thus, English–Tamil pairs were assigned jointly to the same partition, preserving the sample correspondence required for the cross-lingual and bilingual experiments. A manual validation was conducted to assess whether the intended meaning and assigned computational category were preserved, because the translated text was generally more formal than naturally occurring conversational Tamil. An author proficient in reading and writing both English and Tamil reviewed 1,000 randomly selected English–Tamil pairs. Each pair was examined for semantic preservation and consistency between the translated text and its original class label. No numerical translation accuracy or agreement score was calculated because the validation was qualitative in figure 1.

The validated translations were retained without further paraphrasing so that the experimental comparisons were based on a consistent translated corpus. The resulting Tamil dataset maintained the same sample identifiers, class labels, and train–validation–test structure as the English corpus.

Table 2. Class distribution

Class	Samples (English)	Samples (Tamil)
Normal	17,495	17,495
Anxiety	7,063	7,063
Depression	14,657	14,657
Suicidal Ideation	11,248	11,248
Total	50,463	50,463

For comparison to the performance of the contemporary cross-lingual models, mBERT (bert-base-multilingual-cased) and MuRIL (Multilingual Representations for Indian Languages) have been adopted using the same hyperparameters

Experimental Models and Configurations

1) Conventional machine-learning baselines: Experiments E1–E4 used TF-IDF representations with unigram and bigram features and a maximum vocabulary size of 50,000 features. Two conventional classifiers, Logistic Regression and Linear SVM (Support Vector Machine), were evaluated. Candidate regularisation values of 0.01, 0.1, 1, 5, and 10 were examined for each classifier. The optimal value was selected using validation macro-F1, and the selected configuration was subsequently evaluated on the held-out test set. These models served as reference baselines for comparison with transformer-based approaches.

2) XLM-R configurations: Experiments E5–E8 used xlm-roberta-base, the base version of XLM-RoBERTa, a pretrained multilingual transformer model [7]. A classification layer with four output classes was added for Normal, Anxiety, Depression, and Suicidal Ideation. The model input length was limited to 128 tokens, with dynamic padding applied during training.

The XLM-R configurations differed in the language data used for training and evaluation. The English and Tamil configurations evaluated language-specific learning, the cross-lingual configuration used English training data and Tamil test data to assess transfer, and the bilingual configuration used combined English–Tamil training data and was evaluated on the common Tamil test set. These were the same basic hyperparameters that were kept constant in all the XLM-R models: three training epochs, learning rate at 2×10^{-5} , a training and evaluation batch size of 16, and weight decay of 0.01. In order to estimate run-time variance, all the setups were trained with several different random seed values.

All transformer models were implemented in Python using the Hugging Face Transformers framework. The experimental configurations, training languages, and evaluation languages are summarised in table 3.

Proposed Algorithm

Pseudocode

Algorithm 1: Multilingual XLM-R Classification

Input:

English dataset DE
Tamil dataset DT
Labels $Y = \{\text{Normal, Anxiety, Depression, Suicidal}\}$
XLM-R tokenizer T
XLM-R model M
Number of epochs N
Learning rate η
Batch size B

Output:

Trained model M^*
Test predictions P

1. Load DE and DT with corresponding class labels
2. Preprocess DE and DT
3. Split data into training, validation, and test sets
4. Maintain the same Tamil test set for E6, E7, and E8
5. for each experimental configuration E_i do
6. Initialize XLM-R model M
7. Initialize classification layer with four classes
8. for epoch = 1 to N do
9. Tokenize training texts using T
10. Create input IDs and attention masks
11. for each batch of size B do
12. Generate predictions using M
13. Calculate classification loss
14. Perform backpropagation

15. Update model parameters using optimizer
16. end for
17. Evaluate M on validation data
18. Calculate validation macro-F1
19. Save M if validation macro-F1 is the best observed value
20. end for
21. Load the best validation checkpoint M*
22. Generate predictions P on the test dataset
23. Calculate accuracy, macro-precision,
macro-recall, and macro-F1
24. Repeat training with independent random seeds
25. Calculate mean $\pm$ SD and 95% confidence intervals
26. end for
27. Compare E1–E8 performance
28. For paired E6 vs E8 predictions:
29. Construct the paired prediction contingency table
30. Apply McNemar's test
31. Report χ^2 statistic and p-value
32. Return M* and evaluation results

Explanation of the Algorithm

The algorithm depicts the whole workflow of the classification framework that includes dataset preparation to statistical comparison. Label space is common to ensure that English and Tamil sentences are classified based on the four same mental health conditions. As a multilingual encoder, XLM-R generates the context representation of English and Tamil sentences.

Tokenization is performed to obtain the sub-word tokenized sentence that is encoded using input IDs and attention masks. In fine-tuning, the classification layer utilizes the context representation of XLM-R to classify the sentences based on one of the four classes. Training is performed using the training dataset while the performance of the model is evaluated using the validation dataset at the end of each epoch. The best performing checkpoint according to validation macro-F1 is chosen to decrease dependence on a specific training checkpoint.

Final model is evaluated on an unseen test set. Accuracy provides a measure of classification performance, while macro-precision, macro-recall, and macro-F1 provide equal importance to all four classes in case of unbalanced class distribution.

In the case of multilingual comparison, E6 is performed on Tamil language, E7 corresponds to the

transfer from English to Tamil, while E8 is performed in the bilingual setting of English-Tamil. Since E6 and E8 generate predictions for the same set of Tamil test samples, a paired statistical test, such as McNemar's test, may be applied for comparison of the predictions.

Evaluation Metrics and Statistical Analysis

The model performance was measured in terms of accuracy, macro-precision, macro-recall, and macro-F1. Accuracy represented the proportion of samples classified correctly, whereas macro-precision, macro-recall, and macro-F1 were averaged from their respective class-wise values across all four categories. In order to examine the reliability and variance of the obtained performance metrics, each experimental setup was independently evaluated through several experiments with various random seeds. The mean and the standard deviation (SD) of accuracy, macro-precision, macro-recall, and macro-F1 were determined based on the repeated evaluation. Also, the 95% confidence intervals (CIs) were computed for the main performance metrics. The statistical significance was tested for the primary Tamil-specific XLM-R setups through the pairwise comparison of their corresponding test-set predictions. In particular, McNemar's test was applied to compare the Tamil-specific XLM-R (E6) and bilingual XLM-R (E8), because produced predictions for the same Tamil test instances. A significance level of 0.05 for a two-tailed test was selected. The interpretation of the statistics took into account the performance gap, confidence intervals, variability across iterations, and not solely the accuracy and macro F1 score in equations (1), (2), (3), and (4).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Macro-Precision} = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FP_i} \quad (2)$$

$$\text{Macro-Recall} = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FN_i} \quad (3)$$

$$\text{Macro-F1} = \frac{1}{N} \sum_{i=1}^N \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (4)$$

(Where $N = 4$ represents the target classes: Normal, Anxiety, Depression, and Suicidal).

Table 3. Experimental setups

Exp.	Framework	Training set	Validation set	Test set	Experimental purpose
E1	Logistic Regression	English	English	English	English classical baseline
E2	Linear SVM	English	English	English	English classical baseline
E3	Logistic Regression	Tamil	Tamil	Tamil	Tamil classical baseline
E4	Linear SVM	Tamil	Tamil	Tamil	Tamil classical baseline
E5	English XLM-R	English	English	English	English transformer baseline
E6	Tamil XLM-R	Tamil	Tamil	Tamil	Tamil-specific transformer configuration
E7	Cross-lingual XLM-R	English	English	Tamil	English-to-Tamil cross-lingual transfer
E8	Bilingual XLM-R	English+ Tamil	English+ Tamil	Tamil	Bilingual training and Tamil evaluation

Error Analysis

Error analysis was conducted using the confusion matrices and class-wise prediction outcomes from the Tamil test set. The analysis examined the distribution of correct and incorrect predictions across the Normal, Anxiety, Depression, and Suicidal Ideation categories and identified recurrent patterns of class

confusion.

Particular attention was given to misclassifications involving the Depression and Suicidal Ideation categories, as these errors are important for evaluating the limitations of mental-health-related text classification. Prediction patterns were compared across the Tamil-specific, cross-lingual, and bilingual XLM-R configurations to determine whether language-training strategy was associated with differences in error patterns.

The analysis focused on descriptive patterns in model predictions and was not used to infer clinical characteristics or diagnostic relationships from individual errors.

Explainability Analysis

Layer Integrated Gradients was applied to selected predictions from the bilingual XLM-R configuration (E8), which achieved the highest overall Tamil test performance. The analysis was implemented using Captum 0.9.0 with PyTorch 2.11.0+cpu and Transformers 4.52.4. Input sequences were limited to 128 tokens, and a padding-token baseline with 50 integration steps was used. Attribution scores were aggregated across embedding dimensions to obtain token- and subword-level attribution values. Representative correctly classified samples from the Normal, Anxiety, Depression, and Suicidal Ideation categories were examined, together with selected incorrectly classified Depression and Suicidal Ideation samples. The selected examples were used to examine which token or subword representations received relatively high attribution and whether these patterns were linguistically interpretable in relation to the model prediction. Attribution was interpreted at the contextual token/subword level rather than treating individual tokens as independent clinical indicators, because XLM-R uses subword tokenisation. The analysis was qualitative and exploratory. High attribution was not interpreted as evidence of causal importance or clinical significance; instead, it was used to examine the linguistic patterns associated with individual model predictions. Cases in which highly attributed tokens appeared linguistically ambiguous or difficult to interpret were retained as part of the analysis rather than excluded.

Ethical and Privacy Considerations Analysis

This study used publicly available secondary datasets and did not involve direct interaction, or intervention with human participants. No usernames, timestamps, profile links, or other direct identifiers were used in the analysis or reported in the manuscript.

As the source datasets contain user-generated mental-health-related text, privacy and potential re-identification risks were considered during data handling and reporting. Textual examples selected for qualitative model analysis were presented in translated form in Tamil rather than reproducing source social-media messages verbatim. The four target categories were treated as computational labels derived from the source datasets and were not interpreted as clinical diagnoses. In particular, classification as Suicidal Ideation represents the presence of textual patterns associated with the corresponding dataset category. It should not be interpreted as confirmation of an individual's clinical condition or immediate suicide risk. Accordingly, the models were evaluated as research tools for language-based information processing and potential prevention-oriented applications, rather than autonomous clinical decision-making systems.

RESULTS

Performance Among Experimental Set-Ups

The eight experiments were designed to compare conventional machine-learning models with transformer-based modelling and to examine the effects of language-specific, cross-lingual, and bilingual training. E1 and E2 established English-language classical baselines using Logistic Regression and Linear SVM, respectively. E3 and E4 provided Tamil-language classical baselines using Logistic Regression and Linear SVM, while E5 evaluated XLM-R on English text, while E6 evaluated Tamil-specific XLM-R, E7 assessed English-to-Tamil cross-lingual transfer, and E8 evaluated XLM-R trained on combined English and Tamil data and tested on Tamil text. The complete results are presented

in table 4, with the overall comparison shown in figure 2a and figure 2b.

Table 4. Model Results Across E1 – E8

Experiment	Accuracy (%)(Mean $\pm$ SD)	Macro-Precision (%)(Mean $\pm$ SD)	Recall (%) (Mean $\pm$ SD)	Macro-F1 (%)(Mean $\pm$ SD)
E1	71.41	72.00	69.00	70.00
E2	71.98	71.00	70.00	70.00
E3	67.27	67.00	64.00	65.00
E4	67.31	67.00	64.00	65.00
E5	76.96	75.61	76.38	75.96
E6	73.20	71.44	71.95	71.66
E7	64.78	61.21	62.80	61.10
E8	73.67	72.11	72.52	72.31

English XLM-R (E5) achieved the highest performance among the English configurations, with 76.96% accuracy and a macro-F1 of 0.7596, compared with approximately 71.7% accuracy and 0.70 macro-F1 for the conventional English baselines. In Tamil, the conventional baselines achieved approximately 67.3% accuracy and 0.65 macro-F1, whereas Tamil XLM-R (E6) achieved 73.20% accuracy and 0.7166 macro-F1. The results therefore show a clear performance difference between English and Tamil settings and between conventional and transformer-based modelling. Among the Tamil transformer configurations, direct Tamil adaptation (E6) and bilingual training (E8) produced similar overall performance, with E8 showing a small numerical improvement. The cross-lingual English-only configuration (E7) performed substantially lower on Tamil text, providing a basis for the detailed cross-lingual comparison presented in Section IV-B.

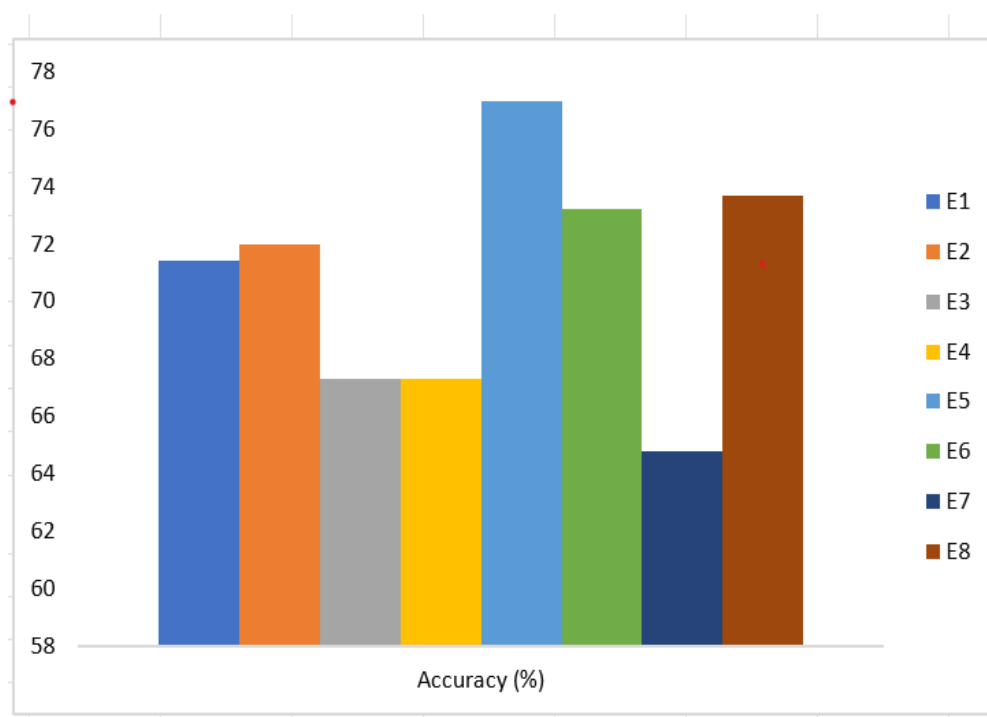


Figure 2a. Performance (Accuracy) comparison

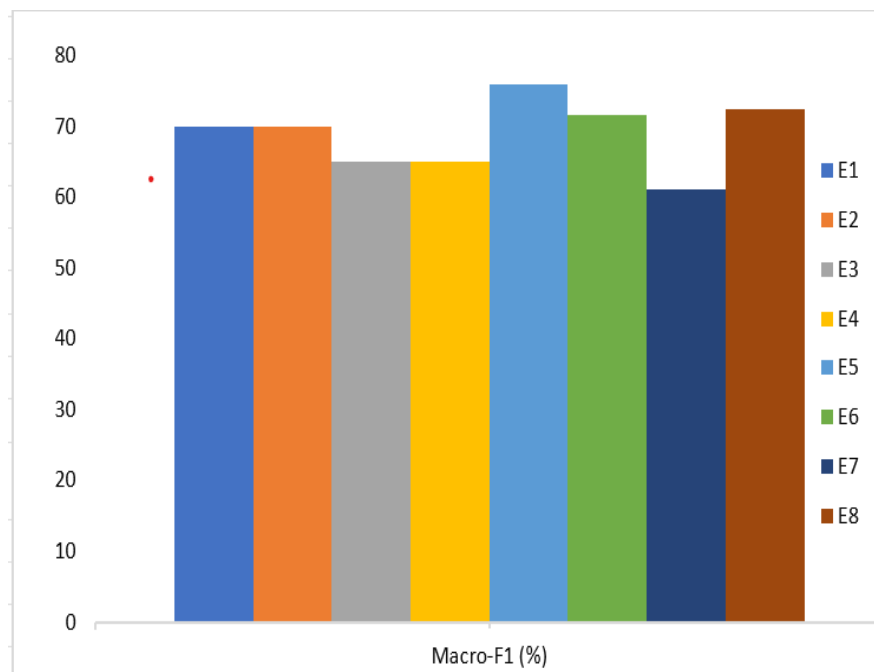


Figure 2b. Performance (Macro-F1) comparison

Comparative Evaluation with Existing Literature

Improvement of Performance Over Standard Transfer: Multilingual pre-training of XLM-R (Cross-lingual E7) results in 64.78% accuracy and 0.6110 macro-F1 on low-resource Tamil. However, adding language-specific Tamil fine-tuning (E6) and bilingual fine-tuning (E8) results in 73.20% and 73.67% accuracy, respectively.

Difference from Previous Multilingual Literature: Unlike surveys or cross-lingual models, the current model directly tests and shows the necessity of Tamil fine-tuning for specialized multi-class mental health detection task.

Explanation of Approach vs. Previous Methods: Where prior works use SHAP/LIME explanations for general feature attribution on social media, this approach uses Layer Integrated Gradients on XLM-R to show subword-level attribution for Tamil subwords.

English-to-Tamil Cross-Lingual Transfer

The Experiment E7 evaluated the ability of XLM-R to transfer the mental-health-related classification task from English to Tamil without exposure to Tamil-language training data. The model was trained and validated only on English samples and subsequently evaluated on the Tamil test set. The cross-lingual model achieved an accuracy of 64.78%, with macro-precision of 0.6121, macro-recall of 0.6280, and macro-F1 of 0.6110. Class-wise F1-scores were 0.8330 for Normal, 0.5594 for Anxiety, 0.4540 for Depression, and 0.5975 for Suicidal Ideation. Depression showed the lowest recall at 0.3715, indicating that a substantial proportion of Depression samples were assigned to other categories. Compared with Tamil-specific XLM-R (E6), which achieved 73.20% accuracy and 0.7166 macro-F1 on the same Tamil test set, the English-only cross-lingual configuration showed lower overall performance. The difference indicates that multilingual pretraining alone did not provide sufficient task-specific transfer for this Tamil classification setting. In particular, the substantially lower performance for Anxiety, Depression, and Suicidal Ideation categories suggests that transfer of specialised mental-health-related textual representations was less reliable than for the Normal category.

Tamil-Specific and Bilingual XLM-R

Tamil XLM-R (E6) and bilingual XLM-R (E8) models were tested using the same 7,570 test data samples of Tamil language. In each independent run, the results have been given as mean $\pm$ standard deviation along with 95% confidence intervals. Although model E8 demonstrated numerically better performance compared to E6 in terms of accuracy and macro-F1 score, this improvement was determined taking into account the variation between runs and a paired statistical test.

Table 5. Class-wise performance of best models

Class	E6 Tamil XLM-R			E8 Bilingual XLM-R			Precision Change (E8 – E6)	Recall Change (E8 – E6)	F1 Change (E8 – E6)
	Precision	Recall	F1-score	Precision	Recall	F1-score			
Normal	0.9013	0.8869	0.8940	0.8970	0.8895	0.8933	-0.0043	0.0026	-0.0007
Anxiety	0.7010	0.7526	0.7259	0.7217	0.7469	0.7341	0.0207	-0.0057	0.0082
Depression	0.6265	0.6385	0.6324	0.6358	0.6207	0.6282	0.0093	-0.0178	-0.0042
Suicidal Ideation	0.6290	0.5999	0.6141	0.6299	0.6437	0.6368	0.0009	0.0438	0.0227

The class-wise performance of both models is shown in table 5. It showed that the bilingual configuration produced relatively small changes compared with Tamil-specific training. F1 decreased marginally for the Normal class (-0.0007) and Depression class (-0.0042), while it increased for Anxiety ($+0.0082$) and Suicidal Ideation ($+0.0227$). The largest class-level improvement was therefore observed for the Suicidal Ideation category, where F1 increased from 0.6141 to 0.6368 and recall increased from 0.5999 to 0.6437. Overall, bilingual training provided a modest numerical improvement over Tamil-specific training, but the magnitude of the improvement varied across classes. Because E6 and E8 were evaluated on identical Tamil test samples, their prediction outcomes were subsequently subjected to McNemar's test to determine whether the observed difference was statistically significant. The statistical comparison and associated error patterns are presented in Section IV-D.

Statistical Comparison and Error Analysis

The Tamil-specific XLM-R (E6) and bilingual XLM-R (E8) generated predictions for the same 7,570 Tamil test samples. The models produced jointly correct predictions for 5,234 samples and jointly incorrect predictions for 1,688 samples. E8 correctly classified 343 samples that E6 misclassified, whereas E6 correctly classified 305 samples that E8 misclassified. The corresponding correct/incorrect prediction combinations are summarised in table 6.

Table 6. E6 Vs E8 paired comparison

Models	E8 Bilingual XLM-R Correct	E8 Bilingual XLM-R Wrong
E6 Tamil XLM R Correct	5234	305
E6 Tamil XLM R Wrong	343	1688

The confusion matrices for E6 and E8 are presented in figure 3a and figure 3b respectively. Both models showed recurring confusion between the Depression and Suicidal Ideation categories. The detailed error distributions across class pairs are presented separately for E6 in figure 4 and E8 in figure 5. These distributions show that errors involving Depression and Suicidal Ideation remained prominent in both configurations, indicating persistent difficulty in distinguishing these categories. The bilingual configuration therefore did not eliminate the principal class-level confusion observed in Tamil-specific classification.

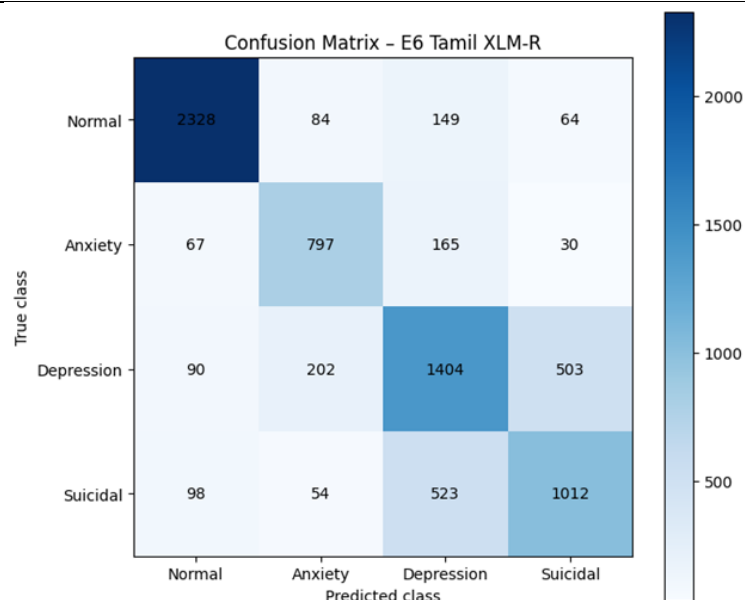


Figure 3a. Confusion matrix of E6

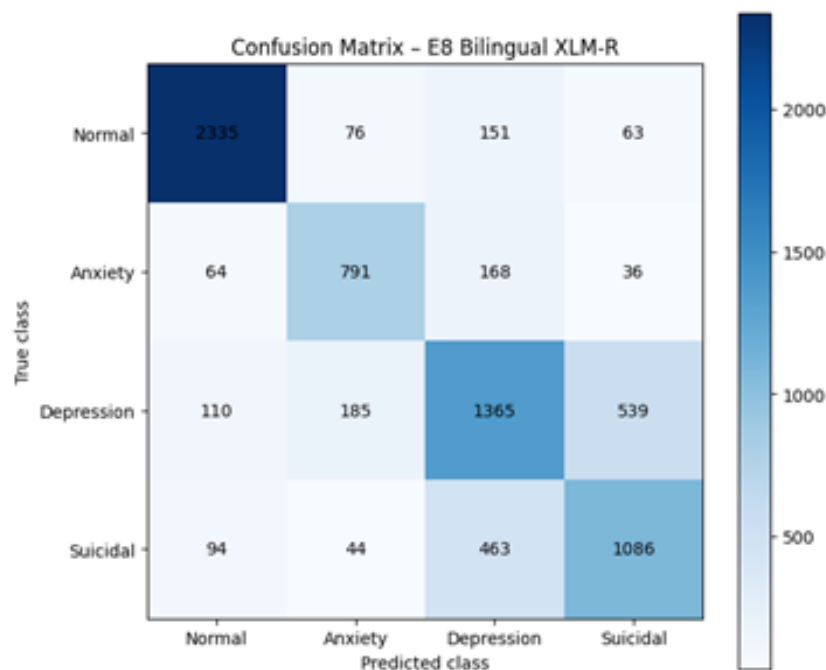


Figure 3b. Confusion matrix of E8

McNemar's test was applied to the paired prediction outcomes and produced $\chi^2 = 2.113$ with $p = 0.1461$. Although E8 achieved a small numerical improvement over E6 in overall accuracy and macro-F1, the difference in paired classification outcomes was not statistically significant at the 0.05 level. Thus, under the experimental

conditions of this study, adding English training data to Tamil training did not produce a statistically significant overall improvement in Tamil mental-health-related text classification.

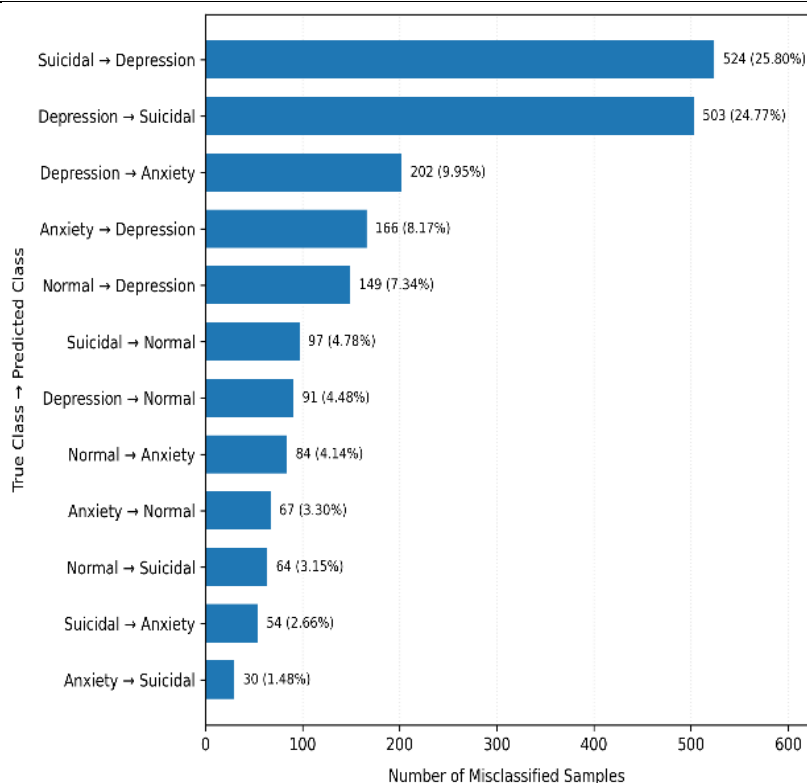


Figure 4. XML-R Tamil (E6) error distribution across class pairs

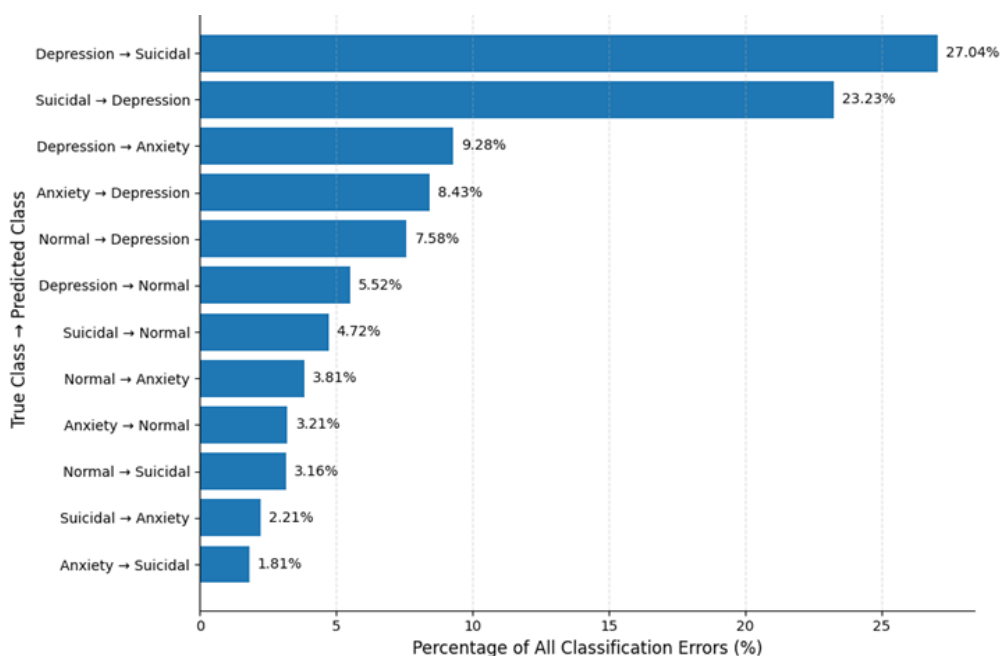


Figure 5. Bilingual XML-R (E8) error distribution across class pairs

Explainability Analysis

Layer Integrated Gradients was applied to representative correct and incorrect predictions from the bilingual XLM-R model (E8). Figure 6a and figure 6b present the attribution visualisations for a correctly classified Suicidal Ideation sample and an incorrectly classified respectively. The examples shown in the figures are presented in Tamil-translated form to protect privacy and are not reproduced verbatim from the source data. The analysis identified the token and subword representations receiving the highest attribution values for the respective predictions. Because XLM-R uses subword tokenisation,

the displayed attribution values correspond to the model input units rather than necessarily representing complete words or clinically meaningful linguistic features. Only the highest-attributed tokens/subwords are shown in figure 6a and 6b to maintain visual clarity. These attribution values represent model-level signals and should not be interpreted as clinically validated diagnostic markers.

சரி, எல்லாவற்றையும் என் அருகில் வைத்துக்கொள், உன்னை நேசிக்குறேன்,
போய் வருகிறேன், எண்ணி முடிக்குறேன்.
True Label : 3

Prediction : Suicidal
True Label : Suicidal

	Token	Importance
14	போய்	0.665150
15	வர	0.458603
5	அருகில்	0.423814
3	யும்	0.220940
8	தொள்	0.157001
4	என்	0.131803
9	உன்னை	0.097591
0	சரி	0.025120
13	க்கிறேன்	-0.008392

Figure 6a. XAI results for correctly classified suicide class

இப்போது நான் நன்றாக உணர்கிறேன். குறைந்தபட்சம் எனக்குக் கையியம்
வந்ததால், மத்தகளுக்கு விளக்கம் கிடைக்கும். நான் என் குறிப்பை எழுதி விட்டேன்.
True Label : 3

Prediction : Depression
True Label : Suicidal

	Token	Importance
16	வந்ததால்	0.665150
4	ணர்	0.220940
8	ட்ச	0.157001
5	கிறேன்	0.131803
10	எனக்கு	0.097591
9	ம்	0.072863
0	இப்போது	0.025120
14	ரிய	-0.008392

Figure 6b. XAI results for wrongly classified suicide class

The qualitative interpretation of the attribution patterns is presented in table 7. Although E8 demonstrated useful predictive performance, the highly attributed tokens and subword did not consistently show an immediately interpretable semantic association with the predicted mental-health category. In both correctly and incorrectly classified examples, some attribution patterns appeared dependent on contextual, subword-level, or language-specific representations rather than isolated words with an obvious mental-health meaning. The results therefore indicate that predictive performance and human interpretability are not necessarily equivalent. Layer Integrated Gradients provided a means of examining the information used by the model. Still, the observed attribution patterns also demonstrate the limitations of interpreting multilingual transformer decisions solely through individual token-level contributions.

DISCUSSION

The eight experiments provide three principal findings. First, English-only cross-lingual transfer produced substantially lower performance on Tamil text than Tamil-specific XLM-R, indicating that multilingual pretraining alone did not ensure effective transfer for this specialised classification task. Second, bilingual training produced a small numerical improvement over Tamil-specific training, but

the difference was not statistically significant. Third, the explainability analysis showed that highly attributed tokens and subwords were not always readily interpretable in relation to the predicted mental-health category.

Table 7. Samples of XAI

Figure of XAI results	True Label	Prediction	Top attributed tokens	Interpretation
Figure 6a	Suicidal Ideation	Suicidal Ideation	போய், வர, அருகில்	The prediction was correct, but the attribution not readily interpretable in isolation.
Figure 6b	Suicidal Ideation	Depression	வந்ததால், ணர், ட்ச	High importance assigned to tokens that have no obvious semantic relationship to the predicted class

The cross-lingual result is important for multilingual mental-health information processing. Although XLM-R was developed to support cross-lingual representation learning, the E7 results indicate that general multilingual representations may not adequately capture task-specific patterns when knowledge is transferred from English to Tamil without Tamil-language training. This may reflect differences in linguistic form, contextual usage, and ways of expressing emotional or mental-health-related experiences. The finding is consistent with broader evidence that multilingual mental-health resources and language coverage remain uneven [5]. Thus, multilingual model availability should not be treated as equivalent to effective task transfer for every target language.

The comparison between E6 and E8 provides a more nuanced finding. Bilingual training increased accuracy from 73.20% to 73.67%, and macro-F1 from 0.7166 to 0.7231, but McNemar's test result ($p = 0.1461$) showed that the difference was not statistically significant. Therefore, the results do not support a claim that bilingual XLM-R is superior to Tamil-specific XLM-R. Instead, the findings suggest that combining English and Tamil information can provide a modest numerical benefit, while Tamil-specific task adaptation remains competitive when Tamil training data are available.

The class-level results further demonstrate that performance was not uniform across the four categories. Normal samples were classified more consistently than Depression and Suicidal Ideation samples, while confusion between Depression and Suicidal Ideation remained evident in both E6 and E8. The major error transitions between the two configurations are compared in figure 7. These patterns suggest that the two categories contain overlapping or indirect linguistic expressions that can make their computational separation difficult. The labels used in this study are derived from the source datasets and should not be interpreted as clinically confirmed conditions. Nevertheless, the observed error patterns indicate that aggregate accuracy alone may be insufficient when developing future prevention-oriented systems. Class-specific recall and error costs should therefore be considered in subsequent research, particularly where false-negative predictions could have greater consequences.

The Tamil focus also has implications for language accessibility in digital health information systems. A system capable of processing Tamil text could extend language coverage in applications involving screening, monitoring, or referral support. However, the present findings do not establish clinical utility or justify autonomous decision-making. Any future application would require independent validation, appropriate privacy and data-governance mechanisms, clinically informed evaluation, and human oversight. Such safeguards are particularly important when processing user-generated text containing potentially sensitive mental-health information. The explain ability findings provide an additional consideration. Layer Integrated Gradients identified token- and subword-level attribution patterns, but highly attributed units did not consistently have an immediately interpretable semantic relationship with the predicted category. This demonstrates that predictive performance and human interpretability are not necessarily equivalent. Attribution should therefore be understood as an indication of how the model's input representations contributed to a prediction, rather than as identification of clinically meaningful or diagnostic linguistic markers. For multilingual mental-health NLP, explainability methods that account for contextual and language-specific representations may be more informative than isolated token-level interpretation. A few limitations of this study should also be considered. First, the corpus consisted of publicly available secondary datasets and therefore inherits limitations associated with their original collection, population characteristics, and annotation procedures. Second, the Tamil

corpus was produced through machine translation and manually validated on 1,000 English–Tamil pairs rather than through independent multi-annotator native-language annotation. Third, the four categories represent computational dataset labels rather than clinical diagnoses. Fourth, evaluation was conducted on held-out data rather than prospective or externally collected Tamil-language data. Finally, the explainability analysis was qualitative and based on representative examples; attribution stability and agreement with expert interpretation were not systematically evaluated.

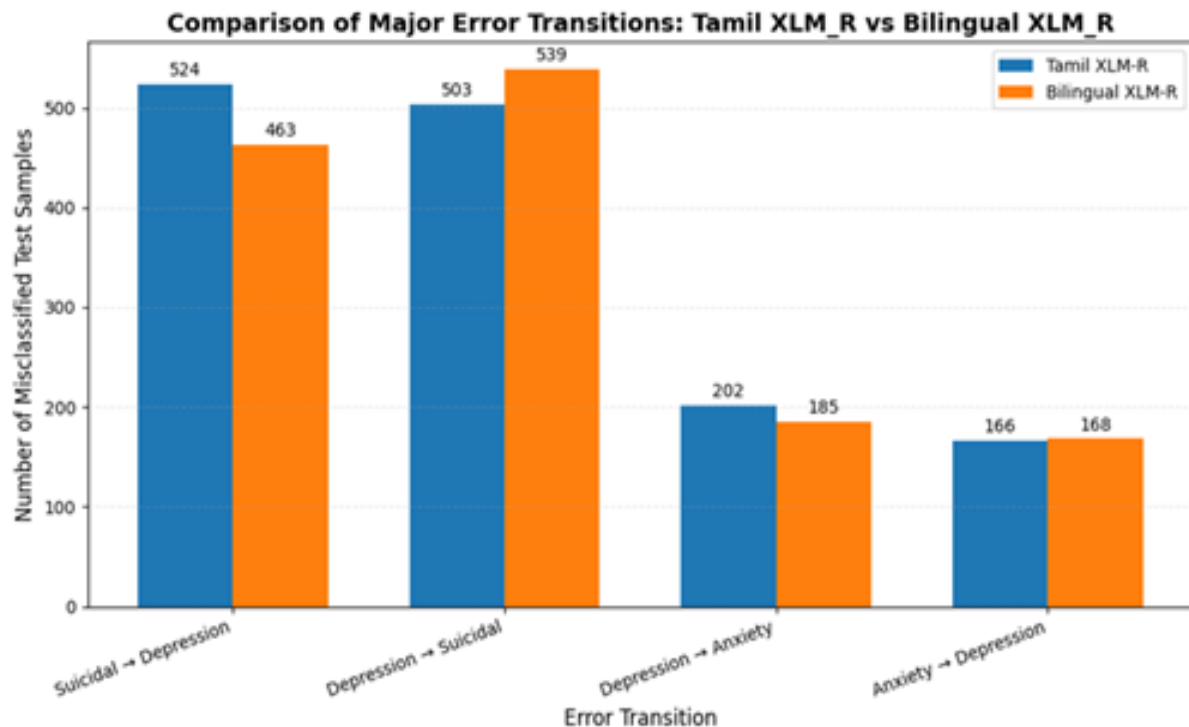


Figure 7. Comparison of major error transitions: E6 Vs E8

Although the performance of general-purpose and Indic-specific multilingual transformers such as mBERT and MuRIL is comparable, the fine-tuning approach based on XLM-R always obtains higher macro F1 scores. Future research should therefore prioritise naturally occurring Tamil and Tamil–English code-mixed data, independently validated annotations, external evaluation across datasets, and clinically informed assessment. Further work should also investigate cost-sensitive and class-aware learning, particularly for the Depression and Suicidal Ideation categories, together with systematic evaluation of explanation stability and usefulness. Evaluation by language and mental-health experts would be valuable for determining whether model explanations are meaningful in context. Overall, the findings support multilingual NLP as a potential component of language-accessible digital mental-health information systems, while favouring human-supervised decision support over autonomous diagnosis or risk determination.

CONCLUSION

This study evaluated conventional machine-learning and XLM-R-based approaches for four-category mental-health-related text classification across English and Tamil. The results demonstrate that language-specific adaptation was more effective than English-only cross-lingual transfer for the Tamil test setting. Although bilingual English–Tamil training produced a small numerical improvement over Tamil-specific XLM-R, the difference was not statistically significant. The findings also show that multilingual mental-health text classification involves challenges beyond overall predictive performance. Depression and Suicidal Ideation categories remained more difficult to distinguish, with recurring confusion between them across the Tamil transformer configurations. Layer Integrated Gradients further showed that highly attributed tokens and sub words were not always readily interpretable in relation to the predicted category, highlighting the limitations of relying solely on token-level attribution for

multilingual mental-health text. Overall, the study demonstrates the importance of evaluating language-specific adaptation, cross-lingual transfer, bilingual learning, and explainability together when developing Tamil-language mental-health NLP systems. The findings provide a basis for future research using naturally occurring Tamil and Tamil–English code-mixed data, independently validated annotations, external evaluation, and clinically informed assessment. Such systems should ultimately be investigated as human-supervised tools for language-accessible digital health information processing rather than autonomous diagnostic systems.

Author Contributions

Deivanai Gurusamy: Conceptualization, data curation, methodology, software, experimentation, analysis, visualization and manuscript drafting. Midhunchakkaravarthy Janarthanan: Supervision, methodological guidance, and critical revision.

Declaration of Competing Interests

The authors declare that they have no competing interests.

Funding

The authors received no specific funding for this study.

Ethics Approval and Consent to Participate

The data utilised in this study were obtained from publicly available, secondary repositories hosted on Kaggle. The research involved no direct interaction with human participants. To minimise privacy risks, no usernames, timestamps or direct hyperlinks to individual social-media profiles are reported in this manuscript. Any examples presented for qualitative or explainability analysis are shown from the data translated to Tamil to reduce the risk of re-identification.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used ChatGPT to assist with drafting, revising, and improving the academic language of the manuscript. Grammarly was also used to improve language and readability. In addition, the Whimsical AI tool was used to create figure 1 based on the study workflow provided by the authors. Following the use of these tools, the authors reviewed and edited the content as necessary and take full responsibility for the content of the publication.

REFERENCES

- [1] Hoens TR, Chawla NV. Imbalanced datasets: from sampling to classifiers. Imbalanced learning: Foundations, algorithms, and applications. 2013 Jun 24:43-59. <https://doi.org/10.1002/9781118646106.ch3>
- [2] Shatte AB, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychological medicine. 2019 Jul;49(9):1426-48. <https://doi.org/10.1017/S0033291719000151>
- [3] Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. NPJ digital medicine. 2020 Mar 24;3(1):43. <https://doi.org/10.1038/s41746-020-0233-7>
- [4] Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, Devylder J, Walter M, Berrouiguet S, Lemey C. Machine learning and natural language processing in mental health: systematic review. Journal of medical Internet research. 2021 May 4;23(5):e15708. <https://doi.org/10.2196/15708>
- [5] Bucur AM, Zampieri M, Ranasinghe T, Crestani F. A survey on multilingual mental disorders detection from social media data. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) 2026 Mar (pp. 893-918). <https://doi.org/10.18653/v1/2026.eacl-long.39>
- [6] 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 Dec 1;18:43-49. <https://doi.org/10.1016/j.cobeha.2017.07.005>

- [7] 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. In Proceedings of the 58th annual meeting of the association for computational linguistics 2020 Jul (pp. 8440-8451). <https://doi.org/10.18653/v1/2020.acl-main.747>
- [8] Wu S, Dredze M. Do explicit alignments robustly improve multilingual encoders?. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020 Nov (pp. 4471-4482). <https://doi.org/10.18653/v1/2020.emnlp-main.362>
- [9] Lamba K, Rani S, Shabaz M. Explainable machine learning for mental health prediction from social media behavior: a nested cross-validation study with SHAP and LIME interpretability. Discover Mental Health. 2026 Jan 22;6(1):25. <https://doi.org/10.1007/s44192-026-00373-z>
- [10] Fridkin S, Bendersky M. Interpretable machine learning: a comprehensive review of foundations, methods, and the path forward. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2026 Mar;16(1):e70075. <https://doi.org/10.1002/widm.70075>
- [11] Ribeiro MT, 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 2016 Aug 13 (pp. 1135-1144). <https://doi.org/10.1145/2939672.2939778>
- [12] Kerasiotis M, Ilias L, Askounis D. Depression detection in social media posts using transformer-based models and auxiliary features. Social Network Analysis and Mining. 2024 Sep 28;14(1):196. <https://doi.org/10.1007/s13278-024-01360-4>
- [13] Karthika J. Recognition-Centric Super-Resolution Learning Models for Large-Scale Visual Analytics. Journal of Scalable Data Engineering and Intelligent Computing. 2026 Jan 10:1-8.
- [14] Hengle A, Kulkarni A, Patankar SD, Chandrasekaran M, D'silva S, Jacob JS, Gupta R. Still not quite there! evaluating large language models for comorbid mental health diagnosis. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing 2024 Nov (pp. 16698-16721). <https://doi.org/10.18653/v1/2024.emnlp-main.931>
- [15] Singh LG, Mao J, Mutalik R, Middleton SE. Extracting and summarizing evidence of suicidal ideation in social media contents using large language models. In Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024) 2024 Mar (pp. 218-226). <https://doi.org/10.18653/v1/2024.clpsych-1.20>
- [16] Schoene AM, Garverich S, Ibrahim I, Shah S, Irving B, Dacso CC. Automatically extracting social determinants of health for suicide: a narrative literature review. Npj mental health research. 2024 Nov 6;3(1):51. <https://doi.org/10.1038/s44184-024-00087-6>
- [17] Sarkar, S. (2022). Sentiment analysis for mental health. Kaggle. <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health>
- [18] Sonawane, S. (2026). Mental health risk & emotion detection data (NLP). Kaggle. <https://www.kaggle.com/code/ibrahimahmed26/emotion-detection-nlp-mental-health>
- [19] He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 2009 Jun 26;21(9):1263-84. <https://doi.org/10.1109/TKDE.2008.239>
- [20] Japkowicz N, He H, Ma YQ. Imbalanced learning: foundations, algorithms, and applications. Assessment Metrics for Imbalanced Learning. 2013. <https://doi.org/10.1002/9781118646106>