WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning Approach for Depression Detection from Code-Mixed (Tanglish) and English Social Media Text

Dr. B. Karthikeyan [1], MS. S. Jayalakshmi [2]

[1] Dr.B. Karthikeyan, Assistant Professor [SG], Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli-620 017, Tamil Nadu, India, bkarthikeyanphd@gmail.com.

[2] MS. S. Jayalakshmi, Student, II MSc, Department of Computer Science, Bishop Heber college (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli-620 017, Tamil Nadu, India, jaylakshmi2k3@gmail.com.

ABSTRACT :

The emerging social media usage and its effect on mental health, paved the way for detecting depression using social posts and various analytical techniques. While much research has been conducted with social media text for mental health evaluations, the research on code-mixed language of social media text is unexplored. This paper addresses the issue by utilizing a Tanglish dataset, which captures the linguistic nuances of bilingual datasets that monolingual datasets may overlook. This paper works with two datasets each of varying sizes, both containing Tanglish and English text, which is to evaluate the performance of traditional machine learning models in code-mixed environments. Three machine learning models—Support Vector Machine (SVM), XGBoost, and Logistic Regression (LR)—were used to classify depressed or not depressed on-related posts. Among these SVM outperformed with highest accuracy on Dataset I, while LR performed best on Dataset II with high accuracy. This study provides a foundation for early depression detection to facilitate timely support for individuals affected by depression.

Keywords: Tanglish, Code-Mixed Languages, Depression Detection.

Introduction :

According to the World Health Organization (WHO), the World Health Organization considers depression as one of the major mental illnesses affecting over 280 million people in the world [1]. Depression is a main reason for suicide, which is the fourth cause of death in those aged between 15 and 29 years. Depression accounts for 77% of all suicides in the developing world. Alarming fact: 75% of all depression cases are untreated in the world. The primary cause for the large percentage of untreated depression is a scarcity of mental health resources, such as qualified personnel and facilities. Early detection mechanisms can mitigate the effects of depression and lead to reduced suicide rates.

Recently, social media has become one of the major sources of data in mental health research. The sources include Facebook, Twitter, Threads, and Instagram. People post their feelings, thoughts and experiences online, which makes for a good repository of text data that can be analysed to identify mental health indicators. Users in South Asian multilingual regions like South Asia tend to communicate more frequently by mixing Tamil with English - Tanglish. Since informal in nature and non-standard in spelling, the language proves quite hard for processing with the techniques available in classical works.

Different computational methods that have been utilized in identifying depression from the social media post. Traditional machine learning approaches are Support Vector Machines (SVM), Naive Bayes, Logistic Regression (LR) and Advanced techniques-deep learning models like Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) which can capture contextual as well as sequential information. These are usually challenged by code-mixed languages because they lack specialized datasets and are also very hard to handle such complexities for language switching.

By focusing on challenges in depression detection systems, this paper utilizes social media text in Tanglish and English, the work focuses on how language mixing in online communication influences the expression and identification of depression. Incorporating linguistic features from both the Tanglish and English, it will enhance detection systems for their better accuracy while making them inclusive. The findings could lead to more effective mental health detection tools, especially in multilingual environments and provide deeper insights into mental health challenges in diverse linguistic and cultural contexts. This paper aims to improve mental health support in regions where code-mixed languages are widely used.

Literature Review :

Prajwal Rai et al. [2] performed a study focused on identifying suitable machine learning models for the early detection of individuals with suicidal tendencies by comparing the various machine learning algorithms like SVM, Naive Bayes and Logistic Regression. The result of this paper shows that SVM outperformed with high accuracy of 95.45% [2024].

Rehmani et al. [3] addressed the challenge of detecting depression by utilizing social media posts in non-structural and structural language. The non-structural language (Roman Urudu) is combined with a structural language (English) to classify the depression risk as not depresses, moderate and severe by utilizing SVM, Random Forest, SVM Radial Basis Function and BERT. The finding reveals that SVM well performed with high accuracy of 84%. The finding reveals that SVM well performed with high accuracy of 84% [2024].

Mohammad Fattah et.al [4] predicted mental health with the help of Twitter text using machine learning models. Support vector machine (SVM) was very effective in predicting depression from tweets with an accuracy of 82%. Logistic regression (LR), XGBoost and random forest (RF) are similar in that they all report very comparable accuracy scores at 79% [2024].

Anuj Kumar [5] conducted a study to detect depression based on Twitter posts using machine learning algorithms. For this, Support Vector Machine (SVM), Random Forest, and Logistic Regression were some of the machine learning algorithms used for depression identification in tweets. Out of all algorithms, Random Forest achieved the maximum prediction accuracy of 88%, which goes to show how effectively it can detect depression-related tweets . Both LR and SVM achieved same accuracy of 74%.[2023].

Suyash Dabhane et.al [6] Implemented various Machine Learning Algorithms to detect depression using Social Media post. Support Vector Machine (SVM) outperforms by achieving 85.04% accuracy, K-Nearest Neighbours (KNN) achieved 73.29% accuracy, Logistic Regression (LR) with an accuracy of 84.86%, Naïve Bayes Classifier with 83.04% accuracy, Decision Tree achieving 80.53% accuracy, Multi-Layer Perceptron (MLP) achieves 78.65% accuracy [2021].

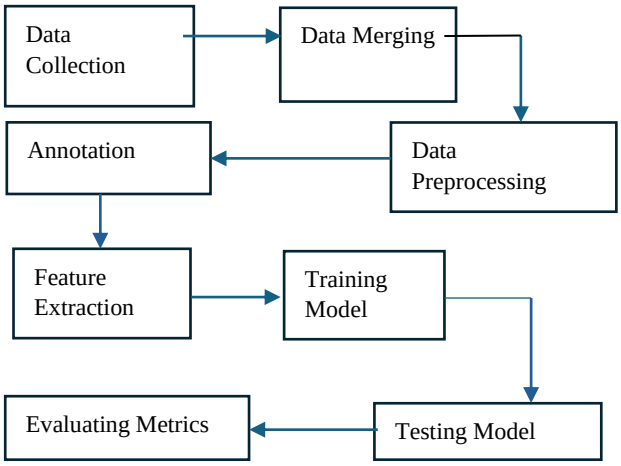
Proposed system :

The proposed system enhances depression detection by utilizing two datasets, both comprising a mix of Tanglish and standard English text. While the datasets share the same linguistic structure, they differ in size, allowing for an analysis of how dataset size impacts the system's performance.

In comparison to existing studies, the proposed system implements new feature extraction methods, enhanced preprocessing techniques and different algorithms. These improvements ensure better handling of informal language, non-standard spellings and linguistic nuances commonly found in Tanglish and English posts.

Methodology :

The process begins with Data Collection, followed by Data Merging, Data Preprocessing, Annotation, Feature Extraction, Training Model, Testing Model and concludes with Evaluating Metrics. Figure 1 represents the methodology used in this paper.



Representation of Methodology

Data Collection :

The two types of data - Tanglish and English data has been collected for this work. These data are then combined to create two separate datasets of varied sizes. The first dataset contains a smaller volume of Tanglish and English text. The second dataset includes a larger volume of the Tanglish and English text with differ sizes.

Case 1

The dataset is used in Case I is referred to as Dataset I (Smaller Dataset). This dataset is composed of a combination of Tanglish and English data.

1.1) Tanglish Data.

The Tanglish dataset includes 4,749 entries, with 2,081 manually collected from social media platforms like Facebook, Twitter, Instagram, and Threads, and 2,668 sourced from Kaggle. This diverse combination enriches the dataset for depression detection.

1.2) English Data

The English data is sourced entirely from Kaggle, comprising a total of 8,287 social media posts.

Case 2

In case 2, Another Dataset II (Larger Dataset) is created by adding 1972 Tanglish data, while retaining same amount of English data. This dataset is to Monitor the performance and accuracy of model when code-mixed language (Tanglish) data is increased.

Sample Records - Tanglish Data

S.NO	Text	Labels
1	Enna vazhkada idhu	Positive
2	Enanandhalum pakthuklam, vidu!!	Negative
3	En Enaku mattu ipidi nadakudhu	Positive
4	friends meet pannadhum enka stress fulla poiduchu	Negative
5	Life ipidiye poiduma sir	Positive

Table Sample records – english data

S.NO	Text	Label
1	I feel so lost lately. Nothing makes sense anymore.	Negative
2	No one would notice if I wasn’t here.	Negative
3	Had a productive day at work, feeling accomplished!	Positive
4	weather amazing today Perfect Walk	Negative
5	I feel unworthy useless	Positive

Data Merging

Data merging essentially merged data - Tanglish (code-mixed) and English (monolingual) data in case1 and case2, allowing the model to learn from code-mixed (Tanglish) and monolingual (Standard English) text in a manner that it would be able to detect depression-related content across diverse linguistic situations.

After Data Merging, there are 13036 data in Dataset I and 15008 data in Dataset II.

Data Preprocessing

In this paper, several preprocessing techniques were applied to both dataset to ensure that the model can effectively process and analyse the text for depression detection. The following six preprocessing techniques were applied in this paper:

- Contraction Expansion
- Punctuation, Special Character
- Lowercase Conversion
- Stopword
- Tokenization

Annotation

The dataset from two different case were annotated using a binary labelling system: A label of 0 was assigned to instances representing "Negative"- indicates as not depressed while a label of 1 was used for "Positive" - indicates as depressed text. This annotation method allows the machine learning model to effectively differentiate between the two classes and improve the accuracy of depression detection.

In Dataset 1 there are 5882 data labelled as 1(depressed) and 7154 data labelled as 0(non-depressed). Likewise, in Dataset II there are 7507 data labelled as 1(depressed) and 7501 data labelled as 0(non-depressed).

Feature Extraction

In this particular paper the technique of feature extraction adopted is TF-IDF.

Term frequency (TF) measures how often a term appears in the given dataset; inverse document frequency measures how rare a term is across the entire dataset; common terms are less informative than rare ones. Therefore, the application of TF-IDF combines both effects-this will help highlight those terms that are common for specific documents and at the same time rare throughout the complete corpus.

Training and Testing the Model

The dataset is split in an 80:20 ratio—80% for training to optimize model parameters and 20% for testing its accuracy on unseen data.

EvaluatingMetrics

Evaluating metrics are significant to assess the performance of a machine learning model. These metrics are efficient for identify areas where the model performs well or where improvements might be needed. The evaluation metrics included in this paper are the confusion matrix, accuracy, precision, recall and F1 score.

Confusion Matrix

Confusion matrix is a table which is to Evaluate the number of correct and incorrect prediction made by the model. It serves as the foundation for other evaluation metrics.

Table Sample records – english data

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

- True Positives (TP): The number of positive instances correctly predicted as positive.
- True Negatives (TN): The number of negative instances correctly predicted as negative.
- False Positives (FP): The number of negative instances incorrectly predicted as positive
- False Negatives (FN): The number of positive instances incorrectly predicted as negative

Accuracy

It measures the proportion of correctly predicted instances out of the total instances.

Correctly predicted Instances: TP + FP

Total Instances: TP + FP + TN + FN

Accuracy= (correctly predicted instance) / (total instance)

Precision

It measures the proportion of predicted positives that are correct.

Precision= TP / (TP+FP)

Recall

It measures the proportion of actual positives that were correctly identified by the model.

Recall= TP / (TP+FN)

F1 Score

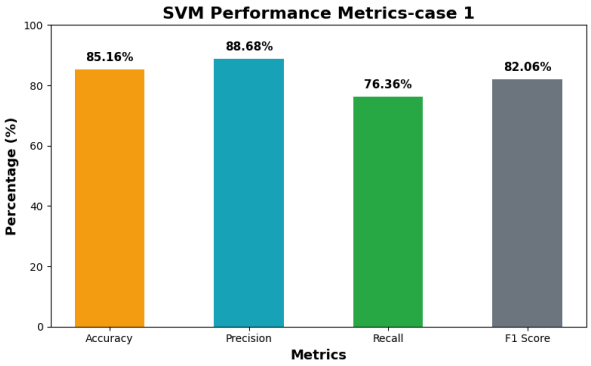
The F1 Score is a metric used to evaluate the balance between Precision and Recall in a model's performance.

F1 =2* (precision * recall) / (precision + recall)

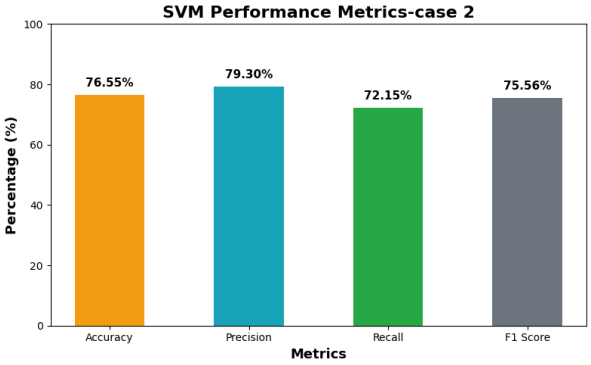
Result and Discussion :

Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm demonstrated exceptional performance, making it highly effective for depression detection using the code-mixed and standard language dataset. Figure 2 shows the SVM performance metrics for Dataset I and Figure 3 Shows the SVM performance metrics for Dataset II.



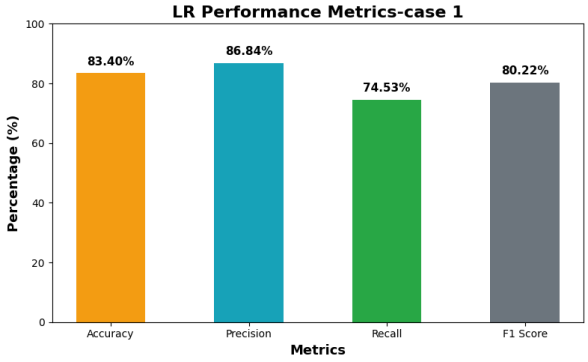
Performance Metrics of SVM (Dataset I)



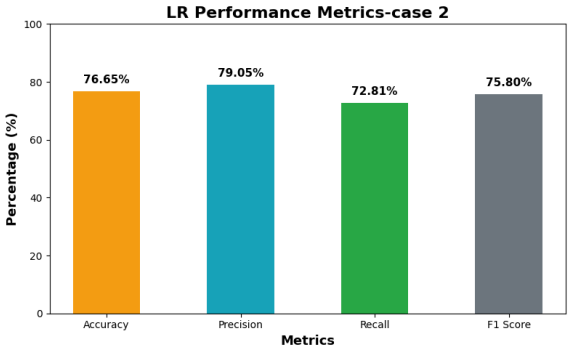
Performance Metrics of SVM (Dataset II)

Logistic Regression (LR)

Logistic Regression, a widely used baseline algorithm, showcased competitive results in this study. Figure 4 represents its classification breakdown in Dataset I and Figure 5 represents its classification breakdown in Dataset II.



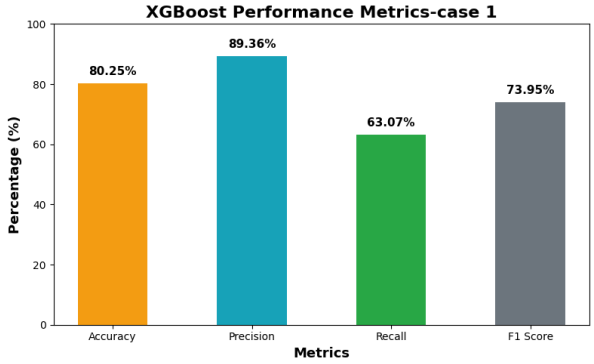
Performance Metrics of LR (Dataset I)



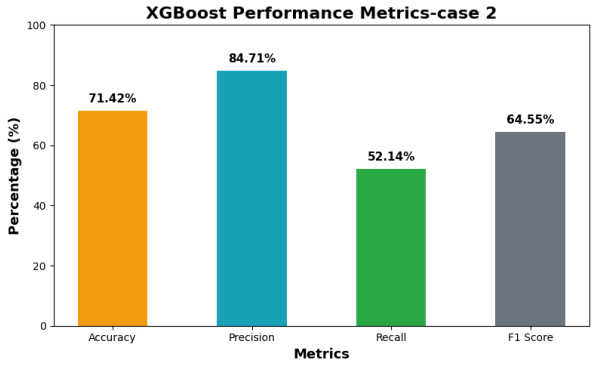
Performance Metrics of LR (Dataset II)

XGBoost

XGBoost, a robust and widely adopted learning technique, exhibited high precision. Figure 6 illustrates Performance Metrics of XGBoot in Dataset I and Figure 7 illustrate Performance Metrics of XGBoot in Dataset II.



Performance Metrics of XGBoost (Dataset I)



Performance Metrics of XGBoost (Dataset II)

Comparison of Performance Metrics

Model	Case	Accuracy	Precision	Recall	F1Score
SVM	Case I	85.16	88.68	76.36	82.06
LR	Case I	83.40	86.84	74.53	80.22
XGBoost	Case I	80.25	89.36	63.07	73.95
SVM	Case II	76.55	79.30	72.15	75.56
LR	Case II	76.65	79.05	72.81	75.80
XGBoost	Case II	71.42	84.71	52.14	64.55

Results show that smaller Dataset in Case I yield higher accuracy, with SVM achieving the highest accuracy of 85.16% and Logistic Regression performing well at 83.40%. XGBoost, despite its high precision, had the lowest accuracy, 80.25%.

For larger dataset in Case II – when more tanglish data added, Logistic Regression achieved the highest accuracy 76.65%, followed by SVM at 76.55%, while XGBoost had the lowest accuracy 71.42 The reason for the low accuracy may be the unstructured spelling in the Tanglish data and utilizing traditional models to evaluate the code-mixed language may lead to low accuracy.

Conclusion :

This paper analysis the impact of dataset size on the performance of machine learning algorithms for depression detection using datasets containing a mix of Tanglish and standard English text. The analysis reveals that smaller datasets result in higher accuracy and balanced performance, while larger datasets (contain more Tanglish data) lead to low accuracy. These findings highlight the inverse relationship between dataset size and model accuracy while incorporating code-mixed language Tanglish and English data.

Limitations :

Tanglish data tends to have inconsistent spellings and colloquial grammar, which can add noise, and it will impact the model accuracy This paper highlights the performance drop in larger datasets but does not explore advanced deep learning techniques, such as Transformers, which could mitigate this issue.

The work does not implement preprocessing techniques for processing Tanglish data, which could enhance model performance.

Future Enhancement :

Future research can aim at integrating more advanced methods to enhance performance. Transformers or neural networks can be investigated to improve the capture of patterns in Tanglish text. Cross-validation techniques can improve the reliability of evaluation. Preprocessing Tanglish text independently of English will maintain its distinctive structure.

REFERENCES :

1. World Health Organization, "Depression," *WHO*, Jan, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.
2. Prajwal Rai, Kumar Prasun, Gajendra Sharma, Yubraj Bhattarai, "Comparison of naïve bayes, logistic regression and support vector machine for predicting suicidal tendency from social media content," *International Journal of Artificial Intelligence Research*, vol. 7, no. 1, Mar. 2024.
3. Filza Rehmani, Qaisar Shaheen, Muhammad Anwar, Muhammad Faheem, "Depression detection with machine learning of structural and non-structural dual languages," *Healthcare Technology Letters*, vol. 11, no. 4, Jun. 2024.
4. Mohammed Fattah, Mohd Anul Haq, "Tweet Prediction for Social Media using Machine Learning," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, Jun. 2024.
5. Anuj Kumar "Journal of Propulsion Technology", International Journal of Research Publications, vol. 44, no. 4, Dec. 2023.
6. S. Dabhane and P. M. Chawan, "Depression Detection on Social Media using Machine Learning Techniques," *International Journal for Scientific Research & Development*, vol. 9, no. 4, Jun. 2021.
7. S. Jayanthi, R. Pradeep, and K. Madhavan, "Analyzing Depression Symptoms Using Social Media Posts with Machine Learning Techniques," Journal of Biomedical Informatics, 2023.
8. Kaggle, "Sentiment Analysis on Code-Mixed Tamil-English Data," 2024.https://www.kaggle.com/datasets/tamil-eng-sentiment-analysis.
9. Kaggle, "Multilingual Depression Detection Dataset," 2024. https://www.kaggle.com/datasets/multilingual-depression-detection.
10. Sentiment140 Dataset. "Sentiment140 - A Large Dataset for SentimentAnalysis,"https://www.kaggle.com/datasets/kazanova/sentiment140.
11. GeeksforGeeks, "Data Preprocessing in Machine Learning using Python," 2023. https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/.