

Early Detection of Multilingual Mental Health Depression Using Pretrained Transformers and Machine Learning

Ali Sami Azeez^{1*}, Osama Abduljaleel Ali², Nawar Abbood Fadhil³,
and Dr. Ali Mohammed Sahan⁴

^{1*}Department of Information Technology Management, Technical College of Management,
Middle Technical University, Baghdad, Iraq. ali.sami@mtu.edu.iq,
<https://orcid.org/0000-0003-3433-7907>

²Computer Center, Al-Muthanna University, Al-Muthanna, Iraq. osama@mu.edu.iq,
<https://orcid.org/0000-0002-0711-5025>

³Department of Information Technology Management, Technical College of Management,
Middle Technical University, Baghdad, Iraq. nawar@mtu.edu.iq,
<https://orcid.org/0000-0002-7741-2965>

⁴Department of Information Technology Management, Technical College of Management,
Middle Technical University, Baghdad, Iraq. dralimohammed2@gmail.com,
<https://orcid.org/0000-0001-5161-4756>

Received: September 29, 2025; Revised: November 22, 2025; Accepted: December 26, 2025; Published: March 31, 2026

Abstract

The social media is producing vast amounts of user-generated text, which can serve as a great indicator of initial mental health diagnosis. This paper develops a scalable, multilingual depression classifier based on classical machine learning (ML) methods and state-of-the-art, pretrained transformer-based models to overcome the weaknesses of language-specific and binary-only methods in previous studies. In a contrast to the majority of the studies, the work is a systematic exploration of bilingual and multilingual depression recognition in the context of Arabic, English, Russian, and Spanish data in a single pipeline. TF-IDF is used to represent textual information to conventional ML classifiers, such as SVM, Random Forest, Naive Bayes and AdaBoost, and transformers, such as XLM-RoBERTa and XLNet are used to train contextual semantic representations. Decades of experiments demonstrate that models using transformers always perform better in comparison to traditional models of machine learning. XLM-RoBERTa provided 94.33% accuracy, 0.94 F1-score, and 0.99 AUC, which outperforms SVM (93% accuracy) and means a lot in terms of performing XLNet (72.36% accuracy). XLM-RoBERTa achieved 99.5% accuracy in Russian, 98% in English, 96% in Arabic, and 85.9% in Spanish in single-language tests, which shows that it is strong in various languages. The findings reveal the usefulness of pretrained multilingual transformers to identify subtle cases of depression, which offers a dependable, language-independent approach to screening early cases of digital depression in mental-health monitoring systems in the real world.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA),
volume: 17, number: 1 (March - 2026), pp. 274-293. DOI: 10.58346/JOWUA.2026.11.016

*Corresponding author: Department of Information Technology Management, Technical College of Management, Middle Technical University, Baghdad, Iraq

Keywords: Multilingual Depression Detection, Mental Health Analytics, Social Media Mining, Transformer Models, XLM-RoBERTa, Machine Learning, Natural Language Processing, Digital Mental Health.

1 Introduction

Mental health is now a significant issue in medical and social sciences because of the growing number of mental illnesses, of which approximately 1 billion people in the world are affected. Covid-19 pandemic increased the mental health problems even in people who were previously healthy, and exacerbated the symptoms of people who already had a mental disorder (Moreno et al., 2020; Akbari et al., 2014). This was particularly true of teenagers, who developed depression as an essential social health issue, associated with suicidal ideation and suicide attempts. Mental disorders also cause a trillion of dollar economic loss worldwide with the most prevalent disorders being anxiety and depression (Health, 2020). Approximately half of all depressed persons in Latin America do not receive appropriate medication, and this is mainly because they are not well diagnosed. Online solutions, such as those based on internet-detectives, are under investigation, especially among the Latino and Spanish-speaking communities. Nevertheless, it requires multilingual care programs and in particular, to immigrant populations. Self-reporting and questionnaires are some of the traditional diagnostic techniques, which are unstable and therefore it is essential to detect and treat them in time in order to be effective in treatment (Adeshina et al., 2025; Pool-Cen et al., 2023). Computational models, such as machine learning (ML) and deep learning (DL) are actively employed to identify mental disorders based on genetic, demographic, and linguistic information (Chauhan & Desai, 2022; Le Glaz et al., 2021; Squarcina et al., 2021), and NLP models, such as the BERT, are critical in the tasks associated with mental health detection (Jain et al., 2019).

The current work is devoted to the under-researched topic of multilingual methods of depression recognition with the help of Natural Language Processing (NLP) in text messages in social media (Kim et al., 2021). Its goal is to create a classification model with TF-IDF, word embedding, machine learning, and transformer-based models to identify depression in various languages on social media networks.

The structure of the paper is as follows: Section 1: Introduction covers the problem of mental health, the effects of COVID-19, and the difficulties of diagnosing the issue, along with the necessity to identify depression in multilingual persons. Section 2: A literature review of the past studies of machine learning, deep learning, hybrid and multilingual detection of depression. Section 3 contains Materials and Methods which contain datasets, preprocessing, embeddings, and ML/transformer models. Section 4 involve results of the experiment with multilingual and language-specific testing, Word Cloud Analysis. Section 5 provides the discussion of this paper and lastly it has a conclusion.

2 Literature Review

The section provides a review of recent studies that determine depression with classical machine learning (ML) and deep learning (DL) methods when dealing with different sources of data. The conventional ML algorithms such as Random Forest, SVM, and Logistic Regression have demonstrated weaker performance in terms of F1-score and accuracy and have been effective in finding patterns but cannot cope with the specifics of language and context in textual data.

2.1 Machine Learning Based Models

Recent research based on social media and health questionnaires presents high performance of traditional ML in recognizing depression. XGBoost scored 83.87% on the D1 and 96.4 on English Twitter data and maximum of 86.45 on Twitter D2 via logistic regression. Random Forest also performed at 90.3% on Bengali Twitter, 77.0% on English twitter, and 82.39% on Arabic twitter with other models such as SVM and Liblinear also performing well. Multi-model TF-IDF were 90.4, RF 94.87 and CNN 95. The study of 2023 reported lower accuracies of SVM and RF (59% and 57%) as well. These ML methods of predicting depression are summarized in table 1.

Table 1: Summary of ML approaches for depression detection

Ref.	Language	Prediction	Method	Best Accuracy
Victor et al., (2020)	English	5 Levels of Depression	RFT, XGBoost, LR, SVM	D1: XGBoost (83.87), D2: LR (86.45)
Skaik & Inkpen, (2020)	Bengali	Depression	DT, RF, SVM, LR, NB, KNN	RF: (90.3)
Azam et al., (2021)	English	Depression	SVM, LR, RF, GBDT, XGBoost	XGBoost: (96.4)
Musleh et al., (2022)	English	Depression	RF, SVM	RF: (77.0)
Zhou et al., (2021)	Arabic	Depression	SVM, RF, LR, KNN, AdaBoost, NB	RF: (82.39)
Kim et al., (2020)	English	Depression	Multi-Model + TF-IDF features	LR: (90.4)

2.2 Deep Learning-Based Models

The section evaluates the use of deep learning methods to detect depression, which are efficient in detecting the patterns of data that are complex. A 2020 study on Reddit claimed CNN as the most accurate (75.13), with a Twitter-based Sense Mood system having 88.39. LSTM-MDL-finetune made 87.14 per cent. on Twitter data in 2021. In 2022, Twitter and Google Trends were tested with CNN, BiGRU, and Distil BERT achieving 85.8 and 84.9%. BERT made 67% on Hindi-English Twitter data, and CNN-based methods on Indian Twitter data made 98% accuracy. These deep learning studies have been found to summarize on social media depression detection, as shown in table 2.

Table 2: Summarizing of related works used deep learning approaches

Ref.	Language	Prediction	DL Approach	Highest Accuracy (%)
Lin et al., (2020)	English	Depression and non-depression	XGBoost, CNN	CNN :75.13
Ghosh & Anwar, (2021)	English	Depression	SenseMood system	88.39
Basiri et al., (2021)	English	Depression	LSTM-MDL-finetuner	87.14
Pradhan & Sharma, (2023)	English	Positive or Negative Opinions about COVID-19	CNN, BiGRU, FastText, NBSVM, DistilBERT	CNN:85.8
Kute, (2022)	Hindi-English	Depression	LSTM, BERT,	BERT:67
Zogan et al., (2022)	Indian	Depression	CNN, LSTM, Bi-LSTM	CNN: 98.00

2.3 Hybrid Deep Learning Models

Deep learning models based on hybrid models, which integrate two or more architectures, have positively performed in detecting depression. A 2022 study on MDHAN on Twitter had an accuracy of 89.5% and a 2023 CNN-BiLSTM predicting normal, depression, and anxiety had an accuracy of 88.93. In 2021, NBTree hybrid model has reached 97.31% accuracy on the D1 and D2 datasets. In 2023, a study of Portuguese twitter with the use of LogReg, LSTM, CNN, and BERT found that BERT outperformed other models, achieving 63% accuracy on depression or 61% on anxiety. These hybrid DL studies are summarized in table 3.

Table 3: Summarizing of related works used hybrid DL approaches-based depression detection

Ref.	Language	Prediction	Hybrid Approach	Highest Accuracy (%)
Bendebane et al., (2023)	English	Depression	MDHAN	89.5
Govindasamy & Palanichamy, (2021)	English	Normal, Depression, and Anxiety	CNN-BiLSTM	88.93
Santos et al., (2024)	English	Depression	NBTree	D1: [97.31] D2: 97.31]

2.4 Multilingual Based Models

Multilingual depression detection has been investigated in several studies based on speech and text. In identified speech characteristics like the rate of speaking and the intensity, authors consider them as MDD biomarkers. A CNN model was able to perform 0.85 accuracy with Chinese speech data, and 0.74 bilingual training. Research employed a collection of LSTM and GRU using Fast Text embeddings on eight India languages. As suggested in, federated learning under multilingual models-maintained privacy and f1-scores of 76.5% (Korean), 98.7% (Arabic), 98.9% (Russian), 85.8% (Spanish), and 69.2% (English).

This literature demonstrates how much progress has been made in using deep learning and conventional ML models to identify depression from user-generated text. The choice of model is still heavily influenced by the data source, language, and particular mental health problem being predicted, even though deep learning model especially hybrid approaches consistently perform better in terms of accuracy than traditional models.

3 Materials and Methods

This section describes the proposed methodology for developing a single, multilingual depression detection system that employs cutting-edge transformers and machine learning (ML) algorithms. The system seeks to classify social media postings into two categories depression, which identifies a mental health issue, and non-depression, which designates normal content, using attributes retrieved from the social media post content. The methodology incorporates dataset collection, data preprocessing, the application of state-of-the-art transformers, classification models using ML techniques, evaluation metrics, and results analysis. The framework for this methodology is exemplified in figure 1 below.

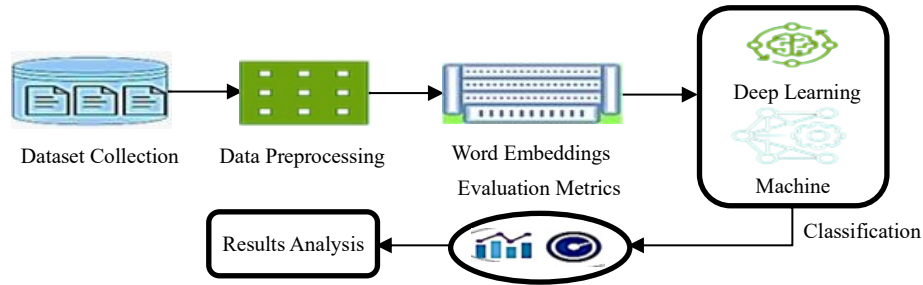


Figure 1: A framework of the suggested methodology

The figure 1 provides the general workflow of the suggested multilingual depression detection system, which demonstrates the process of work with the social media data collection and preprocessing, followed by the feature extraction, model training according to the machine learning and transformer-based methodology, and the ultimate performance evaluation and results analysis.

3.1 Dataset Collection

In the data collection stage, four single-language depression datasets Arabic, Spanish, Russian, and English were pooled into one data frame to analyze multilingual depression. These data sets are composed of twitter posts of social media with unequal sizes in each language. Table 4 is a summarized version of the total samples, division in classes of depression and non-depression, and the number of samples that were chosen to be analyzed. An equal sample of 7,000 samples was taken of Arabic, Russian, Spanish whereas the whole English data of 7,731 was taken.

Table 4: The detailed description and analysis of the datasets

Dataset name	Total size	Depression class	Non-depression class	Selected Samples used
Arabic	10000	5000	5000	7000
Russian	64039	32021	32018	7000
Spanish	173956	96470	77486	7000
English	7731	3900	3831	7731

3.2 Data Preprocessing

The preprocessing of multilingual depression data will consist of some significant steps: the use of punctuation marks and posting aggregation into a new column of positive and negative points, where the absent values are substituted with empty entities. The contents are then transformed to lowercase, trailing/leading whitespaces eliminated and the newline characters substituted with spaces to have uniformity.

3.3 Word Embedding

TF-IDF was employed to transform the text in posts into numeric feature vectors to be used in classical ML models, and word embedding algorithms of the XLM-RoBERTa and XLNet models were used to generate context-sensitive multilingual embeddings that capture language-specific fine-grained linguistic patterns to be analyzed further.

3.4 Classification Models

This section inspects classification models used to predict social media user-generated posts as depression or non-depression. It comprises conventional machine learning models such as AdaBoost,

Naive Bayes, SVM, and RF, all of which focus on TF-IDF feature-based learning. Additionally, cutting-edge transformer models such as XLM-RoBERTa and XLNET are provided, which use built-in attention mechanisms to improve context understanding in multilingual depression tasks such as classification.

3.4.1 ML Models

3.4.1.1 SVM

SVM is a supervised classification tool that divides classes with the help of a hyperplane in high-dimensional space, and it is the best tool to identify the patterns in multilingual social media depression data sets. Equation 1 illustrates the formulation of the model.

$$f(x) = \text{sign}(w \cdot x + b) \quad (1)$$

Here, w represents the weight vector, x denotes the feature vector and b represents the bias term. The SVM function is employed in order to maximize the distance or margin between dissimilar classes, so it is the model that divides the classes in the most straightforward way that it can.

3.4.1.2 Naive Bayes (NB)

Naive Bayes is a probabilistic classifier, which is as well founded on the Bayesian theorem, but it assumes that features are independent of each other. It provides practical calculation and adequate predictive accuracy, which is appropriate to multilingual text classification. Equation 2 demonstrates the formulation of the model.

$$P(C | X) = \frac{P(C) \cdot P(X|C)}{P(X)} \quad (2)$$

Where $P(C | X)$ characterizes the probability of class C assumed features X , $P(C)$ is the preceding, $P(X | C)$, is the probability, and $P(X)$ is the indication.

3.4.1.3 AdaBoost Model

AdaBoost is an ensemble algorithm, a combination of weak classifiers to form a better one, through repeated trials of putting extra weight on the misclassified cases. This will improve multilingual depression detection. Equation 3 illustrates the formulation of the model.

$$F(x) = \sum_{t=1}^T \alpha_t h_t(x) \quad (3)$$

Where α_t is the weight of the weak classifier $h_t(X)$ and T is the total number of classifiers.

3.4.1.4 RF Model

Random Forest trains the individual decision trees using a random subset of data and this minimizes overfitting and enhances the robustness and accuracy of multilingual depression classification. The formulation of the model is presented in equation 4.

$$F(x) = \frac{1}{N} \sum_{i=1}^N T_i(x) \quad (4)$$

Where N represents the set and number of decision trees and $T_i(X)$ is the estimates of the i -th tree.

3.4.2 Transformers Models

This subsection presents the description of state art multilingual large language models applied in this research work for diagnosis and classification multilingual depression detection. These models are XLMROBERTA and XLNET transformers. Further, each model has different deep learning architecture for processing cross word embedding features contexts extracted from the social media post contents.

3.4.3 XLMROBERTA Model Description

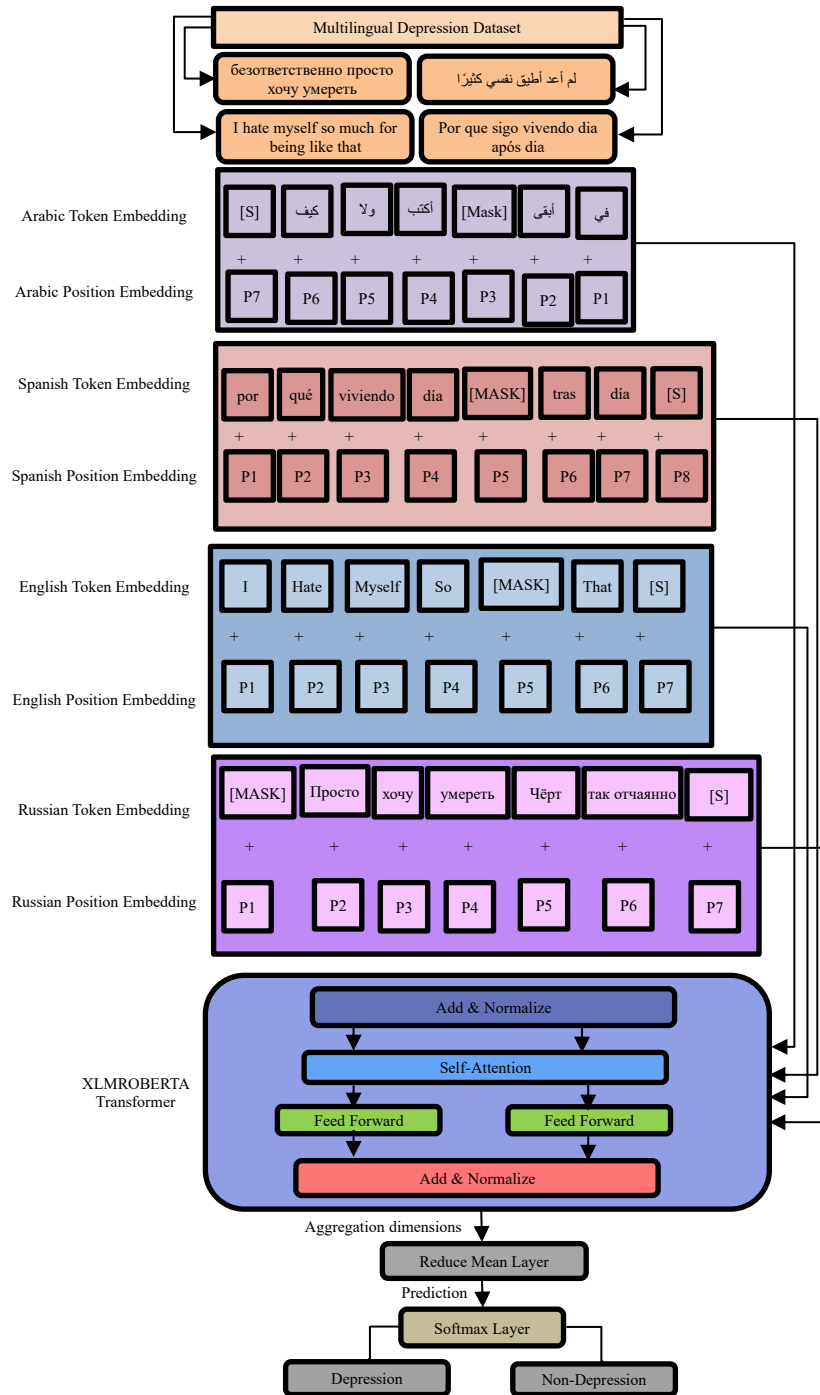


Figure 2: Demonstrating of the XLMROBERTA architecture used

XLM-RoBERTa is a cross-lingual language model, which has well-optimized bidirectional encoder representations. Its multilingual architecture of detection of depression based on social media post is illustrated in figure 2 it processes Arabic, Spanish, English, and Russian posts. The input sentences are processed into token embeddings by being tokenized and then positional embeddings are added to generate the final input representations.

Embeddings are then fed through XLM-RoBERTa which uses normalization layers, self-attention and feed-forward networks in order to encode semantic and syntactic interrelations between languages. The outputs are summed together using a mean reduction layer and are inputted into a SoftMax classification head to make binary Depression or Non-Depression predictions. This model is effective when dealing with multilingual samples, and table 5 shows the hyperparameters of the model.

Table 5: XLM-RoBERTa model architecture parameters

Parameter	Value	Description
Embedding Dimension	768	Dimensionality of token embeddings produced by the XLM-RoBERTa model.
Maximum Sequence Length	80	Maximum number of tokens considered for each input sequence.
Batch Size	16	Number of samples processed together in one training iteration.
Epochs	5	Number of complete passes through the training dataset.
Learning Rate	1e-5	Step size for gradient descent optimization.
Optimizer	Adam Optimizer	Algorithm used for updating model weights during training.
Dropout Rate	0.1	Regularization technique to prevent overfitting by dropping connections during training.
Embedding Layers	Token and Position Embeddings	Converts text tokens and the positions into numerical representations for transformer input.
Classification Layer	Dense (softmax activation)	Final layer for predicting class probabilities (depression or non-depression).

This table 5 provides the main parameters and architecture specifications of the depression detection model with XLM-RoBERTa. It involves training parameters of embedding dimension, sequence length, batch size, epochs, learning rate, optimizer, dropout rate, and model elements of embedding layers of token and position representations and a dense classification layer with SoftMax activation to predict depression or non-depression.

3.4.2.2 XLNET Transformer Model Description

XLNet is an autoregressive language model that builds on the Transformer architecture and has next-token prediction, permutation-based training, segment-level recurrent model and two-stream attention model. To identify multilingual depression, XLNet is trained on Arabic, Spanish, English, and Russian posts, where the information is converted into numerical values through token and position embeddings to extract semantic meaning and maintain sequence of information to read and understand it in context. The XLNet architecture was used as shown in figure 3.

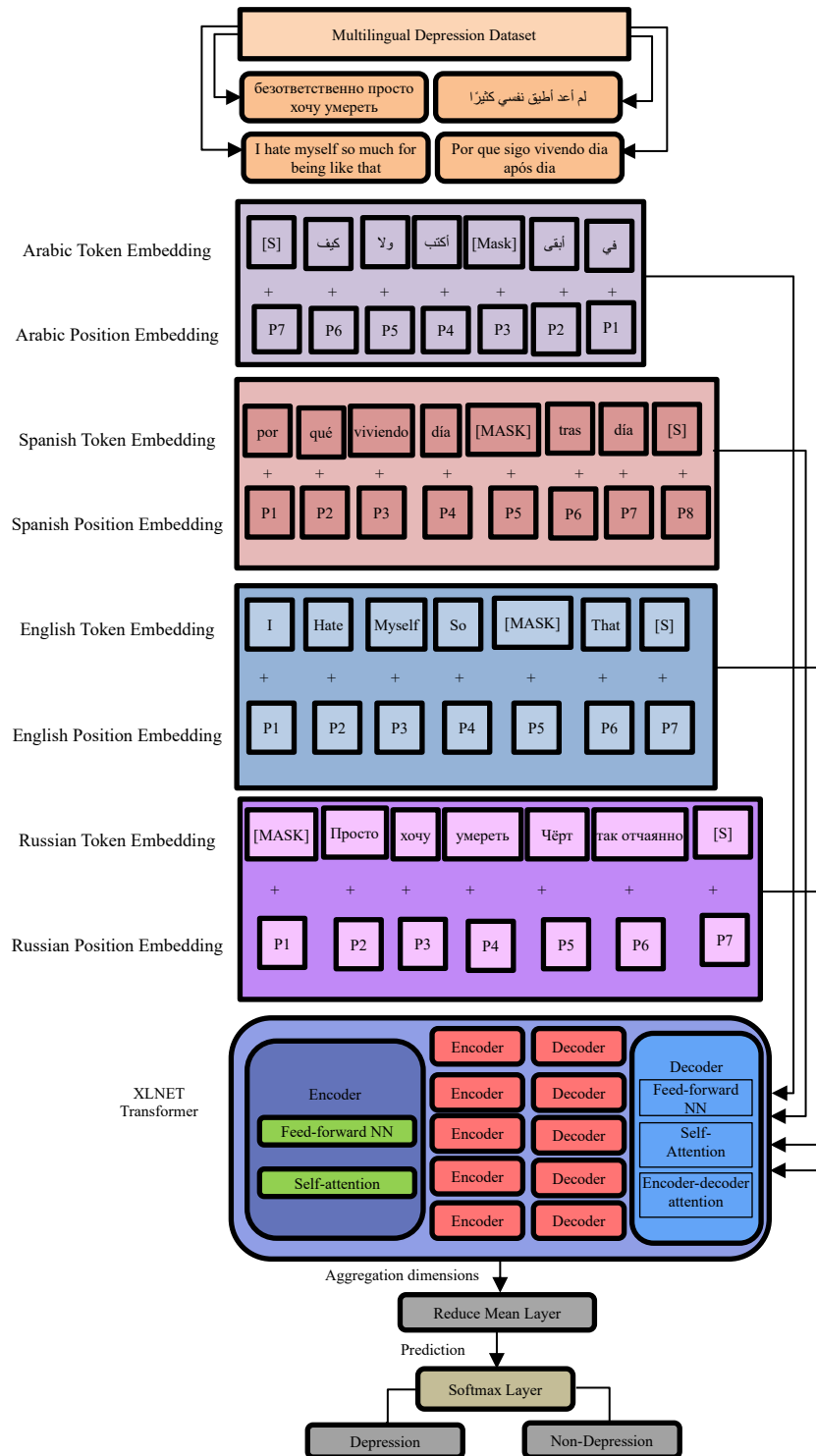


Figure 3: The XLNET architecture used for multilanguage depression detection

In XLNet, the encoder is a transformer based on self-attention in order to learn the dependencies between tokens, then a decoder with cross-lingual attention in order to identify patterns of depression across languages. The output is pooled into a fixed-length vector, and it is sent to a SoftMax layer to predict the post as depression or non-depression. The design makes use of the multilingual embeddings

and cross-lingua attention, which are effective in managing the linguistic diversity and contextual relevance. The parameters in the model are listed in table 6.

Table 6: Summarization of hyperparameters used in XLNET model architecture

Parameter	Value
Embedding Dimension	768
Maximum Sequence Length	80
Batch Size	16
Epochs	5
Learning Rate	1e-5
Optimizer	Adam Optimizer
Dropout Rate	0.1
Classification Layer	Dense (softmax activation)

This table 6 is a summary of the essential hyperparameters and model setting of the depression detection system, like embedding dimension, sequence length, batch size, number of epochs, learning rate, optimizer, dropout rate, and classification layer adopted in predicting depression or not.

4 Experimental Results

This part provides the classification performance of machine learning models (SVM, RF, Naive Bayes, AdaBoost), and transformer models (XLM-RoBERTa, XLNet) on the social media posts on multilingual depression detection. Arabic, Russian, English, and Spanish are analyzed with the help of TF-IDF and word embeddings. Model effectiveness is measured using evaluations measures of accuracy, precision, recall, F1-score, and specificity. Transformer experiments included datasets divided 80/ 20 in ML models and 60/ 20/ 20 training- testing- validation.

The models of depression detection in this paper were run in Python with the help of such libraries as scikit-learn which supports the use of such traditional machine learning models as SVM, Naive Bayes, Random Forest, AdaBoost, and transformers which include XLM-RoBERTa and XLNet. XLM-RoBERTa parameter settings were an embedding dimension equal to 768, a maximum sequence length of 80 tokens, a batch size of 16 and five training epochs. The Adam optimizer was used with the learning rate of 1e-5 and dropout rate of 0.1 in order to avoid overfitting. In the case of machine learning models, default parameters were popular and these were the tree depth, regularization, and the number of estimators where the feature extraction was performed using TF-IDF in combination with the machine learning models.

Evaluation Metrics

1. **Accuracy:** Accuracy determines the accuracy of the model in general, the percentage of correct predictions (positive and negative) of all projections.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

2. **Precision:** Precision is measuring the accuracy of positive predictions, which is the percentage of true positives in all of the predicted positive cases.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

3. **Recall:** Recall is used to estimate the capability of a particular model to detect all positive cases that are relevant, which gives the percentage of true positives in relation to all actual positives.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (7)$$

4. **F1-Score:** The F1-Score is the harmonic mean of precision and recall and is a way of trading off between the two, particularly when the data set is not balanced.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

For equation (5-8), Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

4.1 Testing Results of ML Models

Random Forest was the most effective ML model in the multilingual detection of depression experiment based on TF-IDF features, with its precision, recall, and F1-score of 0.92, and specificity of 0.94. Naive Bayes was somewhat moderate, and its accuracy is 0.86 with a recall of 0.85 and F1-score of 0.85 and specificity of 0.76 which is lower as compared to the other two models. The summary of the ML testing is presented in table 7.

Table 7: ML models testing results for multilingual depression detection

Model	Precision	Recall	F1 Score	Specificity
RF	0.92	0.92	0.92	0.94
NB	0.86	0.85	0.85	0.76
SVM	0.93	0.93	0.93	0.95
AdaBoost	0.89	0.89	0.89	0.93

SVM demonstrated good results of 0.93 precision, recall and F1-score, and 0.95 specificity, and was effective in both distinguishing between depression and non-depression. AdaBoost had 0.89 precision, recall, and F1-score and 0.93 specificity, which is not as good as RF and SVM but still good.

4.2 Testing Results of the XLMROBERTA Model

4.2.1 Multilingual Depression Detection Testing Results

Table 8 indicates that the XLM-RoBERTa model has an accuracy, and 94.33% with a low-test loss of 0.2198 and precision, recall, and F1-score of 94,0.94, respectively, for multilingual depression. It is effective as illustrated by its 93% specificity in the recognition of non-depressive content. These findings show that the model is reliable and can be scaled to different languages.

Table 8: XLM-RoBERTa testing results for multilingual depression detection

Overall Performance		Class-wise Performance		Depression	Non-depression
Metric	Value	Metric	Value	Value	Value
Test Loss	0.2198	Precision	0.96	0.92	
Test Accuracy	0.9433	Recall	0.93	0.96	
F1-Score	0.94	F1-Score	0.94	0.94	
Specificity	0.93	Support	2148	2052	
Precision	0.94				

Analysis of XLM-RoBERTa by classes demonstrates good results on multilingual depression. Precision, recall, and F1-score are 96, 93 and 94 respectively when used with non-depression and 93, 96 and 94 respectively when used with depression; this shows balanced and accurate classifications. The confusion matrix and ROC curve can be found in figure 4.

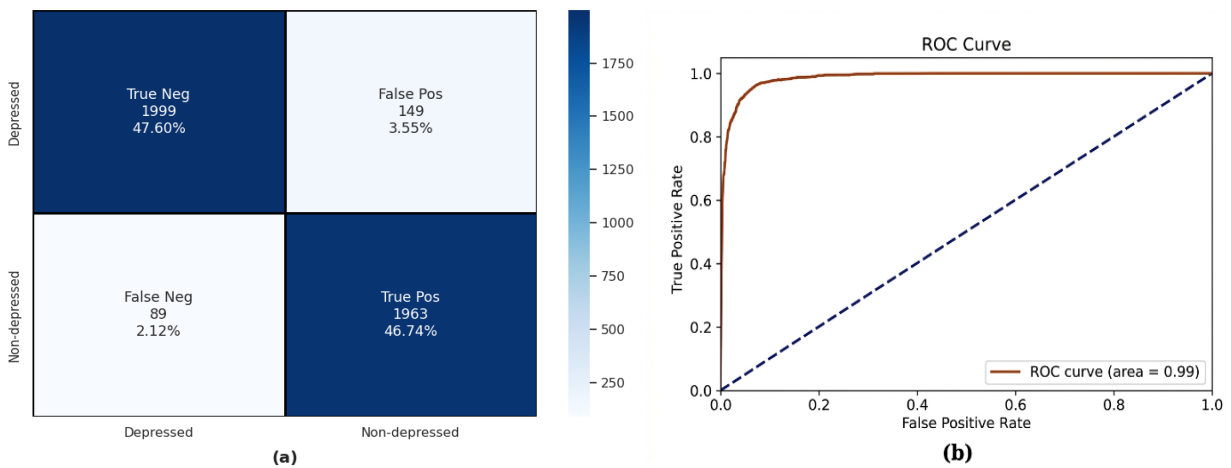


Figure 4: (a) CM and (b) ROC curve of XLM-RoBERTa in multilingual experiment

Figure 4-a presents the confusion matrix where XLM-RoBERTa has correctly identified 47.6% and 46.74% of the cases of depressed and non-depressed respectively, and the misclassification rates are low (3.55% false positives, 2.12% false negatives). The curve of ROC is depicted in figure 4-b and has an AUC of 0.99, which implies high sensitivity and low false positive. These findings reveal reliability and strength of the model and figure 5 represents training and validation performance.

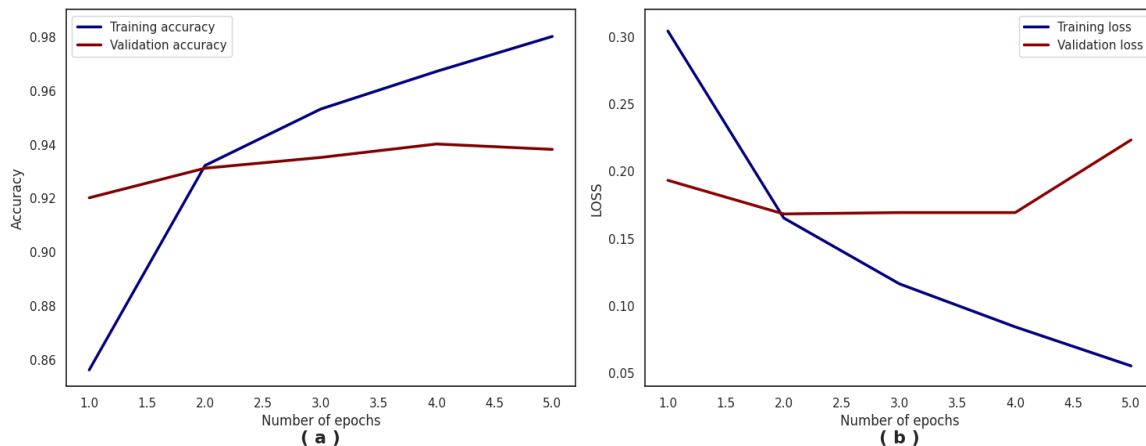


Figure 5: Shows (a) Training and validation and (b) Losses accuracies of the model

Figure 5 demonstrates the training and validation performance of XLM-RoBERTa during five epochs. Accuracy of training increases to 98 and the validation accuracy stabilizes at 95 hence constant generalization. The training loss decreases steadily and validation loss decreases at the beginning, but increases slightly in the third epoch, which can be regarded as a slight overfitting.

4.2.2 Testing Results of the XLMROBERTA Model Using Russian Language

This subsection presents the testing results of the XLM-RoBERTa model for detecting depression in Russian language using word embedding features extracted from social media post content. The model achieved an impressive test accuracy of 99.5%, with a minimal test loss of 0.0219. It demonstrated near-perfect classification performance, with a precision of 1.00 and a recall of 0.99 for depression detection. Table 9 presents testing results of the XLMROBERTA model using Russian language data.

Table 9: Testing classification results of the XLMROBERTA model using Russian language

Overall performance		Class-wise performance		
Metric	Value	Metric	Depression	Non-depression
Test Loss	0.0219	Precision	0.99	1.00
Test Accuracy	0.995	Recall	1.00	0.99
F1-Score	0.99	F1-Score	1.00	0.99
Recall	0.99	Support	687	714
Specificity	0.99			

The model obtained a balanced F1-score of 0.99, a precision of 1.00 and a recall of 0.99 on non-depression, which indicates effective prediction of depression in the Russian-language posts on the social media using the least number of false positives.

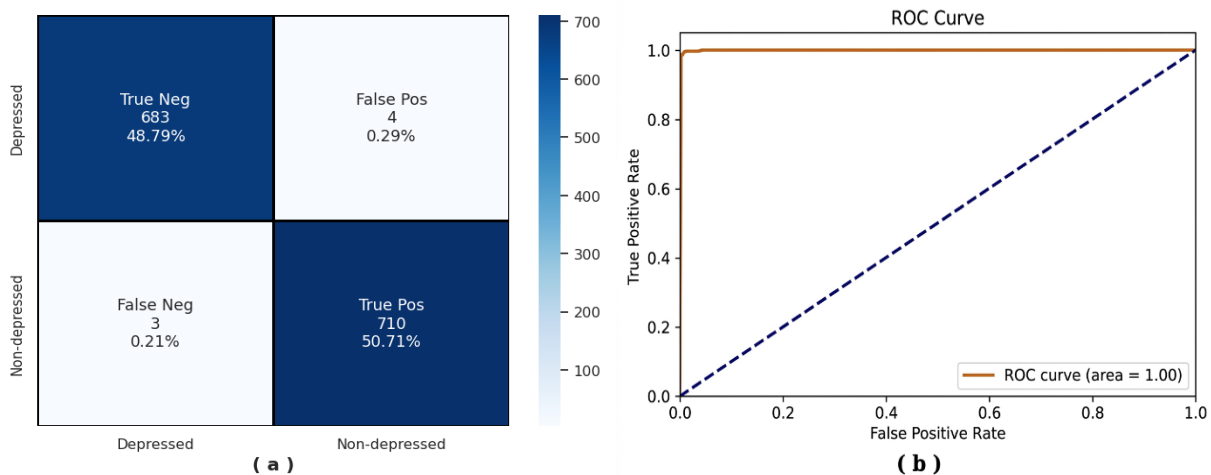


Figure 6: (a) CM and (b) ROC curve of XLM-RoBERTa on Russian data

In figure 6-a, it is observed that XLM-RoBERTa accurately labeled 48.79 % of depressed and 50.71% non-depressed Russian posts, and the misclassifications were very few. The AUC of the ROC curve (Figure 6-b) is 1.00, and this means that the ROC curve has virtually perfect discrimination, high sensitivity, and low false positives.

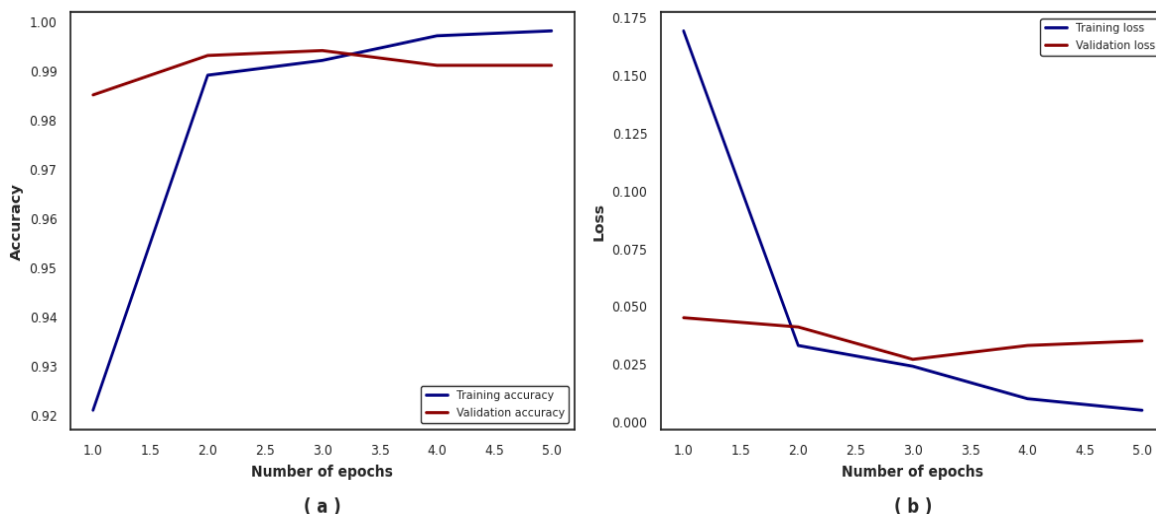


Figure 7: (a) Accuracy and (b) Loss plots of XLM-RoBERTa on Russian data

The performance of the XLM-RoBERTa model on the Russian data is shown in figure 7: (7-a) Accuracy Plot shows training accuracy reached 100% and validation accuracy reached 99%, indicating excellent generalization; (7-b) Loss Plot shows a consistent decline in training loss and high, stable validation loss, which means excellent and efficient depression classification.

4.2.3 Testing Results of the XLMROBERT Model Using Spanish Language Data

In this sub-section, the results of the XLM-RoBERT a model on Spanish-language data (Table 10) are presented. The model is sensitive to both cases of depression and non-depression with a compromise of accuracy and recall. It demonstrates a slightly better recall in depression and a better precision in non-depression which implies that there is a need to enhance recall in non-depression detection to achieve better performance.

Table 10: XLM-RoBERTa testing results on Spanish data

Overall Performance		Class-wise Performance		Depression	Non-depression
Metric	Value	Metric	Value	Value	Value
Test Loss	0.364	Precision	0.81	0.92	
Test Accuracy	85.9%	Recall	0.93	0.79	
F1-Score	0.86	F1-Score	0.87	0.85	
Specificity	0.93	Support	687	713	
Precision	0.87				

Figure 8 presents the confusion matrix (8-a) and ROC curve (8-b) of XLM-RoBERTa using the data in the Spanish language. The model reasonably identified 638 true negatives (45.57) and 565 true positives (40.36) where 49 equals false positives and 148 equals false negatives. The ROC curve shows that the model is strongly performing with an AUC of 0.93 which shows excellent discrimination between depressed and non-depressed posts and also the model is effective in the Spanish-speaking population.

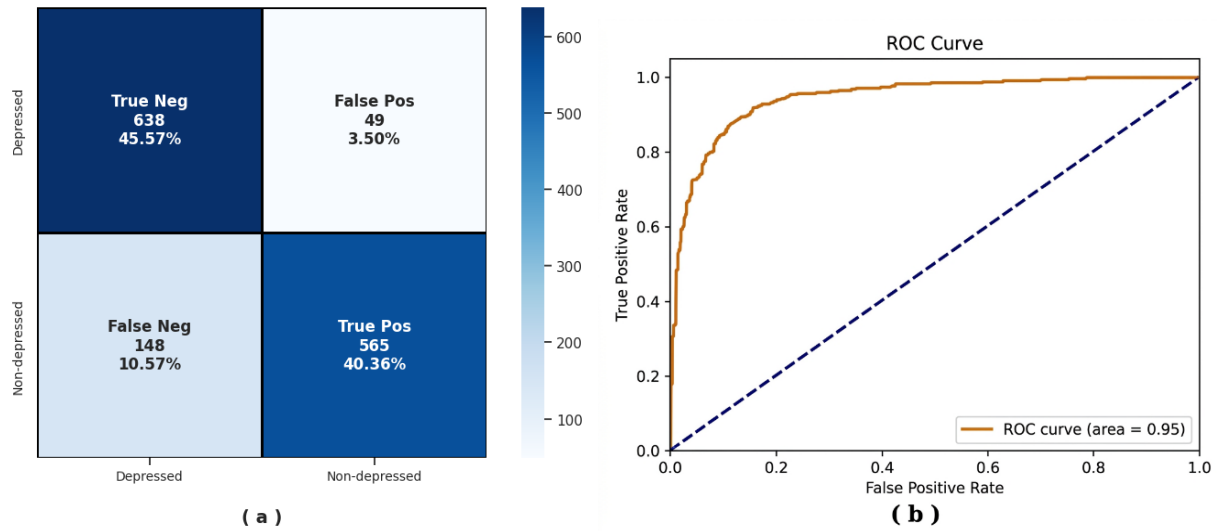


Figure 8: (a) CM and (b) ROC curve of XLM-RoBERTa on Spanish data

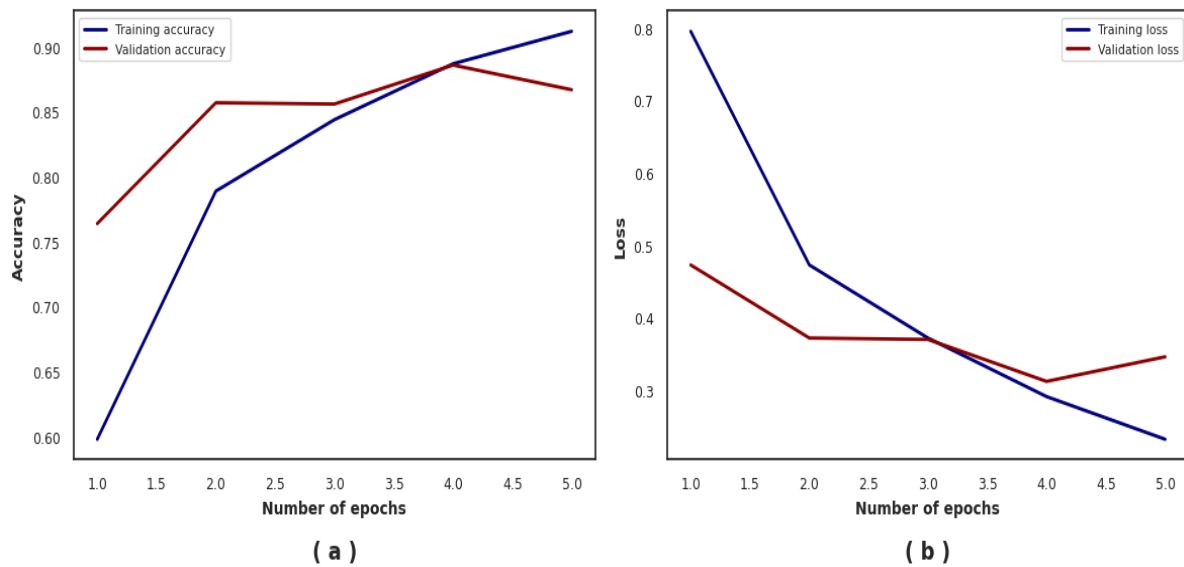


Figure 9: (a) Accuracy and (b) Loss plots of XLM-RoBERTa on Spanish data

Figure 9 demonstrates that XLM-RoBERTa can perform on Spanish data in 5 epochs. The training and validation accuracies increase in the accuracy plot (9-a) to almost 90 and 90.4, respectively, and thus, there is slight overfitting. The loss plot (9-b) indicates that the training loss is slowly decreasing and the validation loss is also decreasing, but at first, it is stabilizing, which means successful learning but requires more optimization of generalization to the unknown data.

4.3 Testing the Results of the XLNET Model Based Multilingual Depression Detection

Table 11 indicates that the XLNet model had an accuracy of 72.36 and F1-score of 0.76 on multilingual depression detection. The model achieves very high recall in detecting cases of depression but its poor accuracy shows that it has low false positives, hence there can be better ways of minimizing classification errors.

Table 11: XLNet testing results for multilingual depression detection

Overall Performance		Class-wise Performance		Depression	Non-depression
Metric	Value	Metric	Value	Value	Value
Test Loss	0.5445	Precision	0.87	0.66	
Test Accuracy	72.36	Recall	0.54	0.91	
F1-Score	0.76	F1-Score	0.67	0.76	
Specificity	0.72	Support			
Precision	0.71		2148	2052	

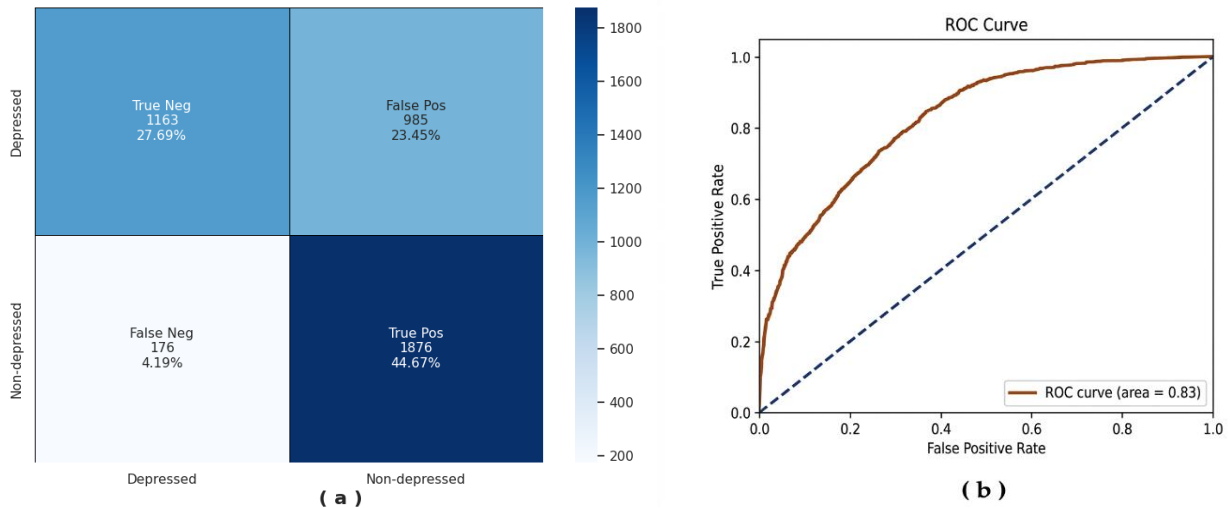


Figure 10: (a) CM and (b) ROC curve of XLNet for depression detection

Figure 10 shows the model’s performance with (a) a confusion matrix and (b) a ROC curve. The confusion matrix reports 1,876 true positives (44.67%) and 1,163 true negatives (27.69%), with 985 false positives and 176 false negatives, indicating moderate sensitivity but lower specificity. The ROC curve (AUC = 0.83) reflects moderate discrimination between depression and non-depression, suggesting the need for optimization to reduce false positives and improve balance.

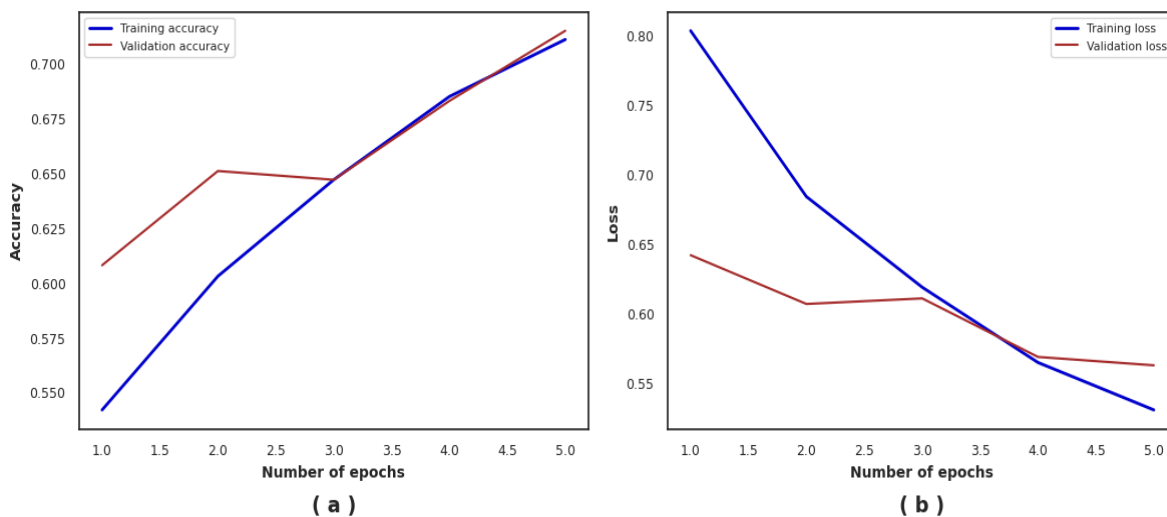


Figure 11: (a) Training and validation accuracies and (b) Losses plots of XLNET model using multilingual depression dataset

5 Discussion

The paper illustrates that transformer-based models especially XLM-RoBERTa can be effectively used in detecting multilingual depression with an accuracy of 94.33 with high performance on all languages including Russian (99.5%), English (98), and Arabic (96). Although there was a slight decrease in accuracy (85.9%) on Spanish, the high specificity, recall, F1-score, and precision of the model means that it can be trusted to recognize posts which have depressive and non-depressive content. Comparatively, XLNet exhibited fair performance (72.36% accuracy, F1 = 0.76), indicating the better performance of the cross-lingual attention systems of XLM-RoBERTa. More complex semantic and syntactic dependencies are essential because traditional ML models, such as SVM, Random Forest, and AdaBoost, were surpassed by transformer models, which were adequate but not limited to simple associations. The findings confirm the hypothesis of the proposed framework as a scalable, language-independent approach to early detection of depression on social media, and it can be applied to monitor public health. The way of work in the future should be aimed at covering the language and standardizing datasets, consideration of hybrid models to increase the accuracy and the strength.

6 Conclusion

Depression, which is a persistent sadness, lack of interest and poor performance, is a condition that should be detected and interceded early, in order to be managed. The current paper presents both single and multilingual depression detection models, which are built based on the traditional machine learning algorithms (SVM, Naive Bayes, Random Forest, AdaBoost) and the transformer models (XLM-RoBERTa, XLNet), compared on diverse datasets with such measures as accuracy, recall, F1-score and privacy. SVM was the most accurate (93%), and XLM-RoBERTa was stronger than XLNet with 94% and proves to be strong in multilingual settings like Arabic English Russian and Spanish. The research paper indicates the benefits of transformer models compared to conventional classifiers in generalization and contextual learning. The future direction of work will be to enhance generalization of the model, increase language coverage and to tackle either ethical or privacy issue in mental health inference; privacy protective solutions will be adopted to ensure that sensitive information is securely taken care of.

References

- [1] Adeshina, A. M., Adeleye, O., & Razak, S. F. A. (2025). Predictive Model for Healthcare Software Defect Severity using Vote Ensemble Learning and Natural Language Processing. *Journal of Internet Services and Information Security*, 15(1), 437-450. <https://doi.org/10.58346/JISIS.2025.11.029>
- [2] Akbari, M., Mehrabi, A., Karamkhani, J., Shahbaz-poor, H. R., Hemmati, M., Ranjbari, M., & Moradi, A. (2014). Identifying the effect of life skills training on addicts' mental health who referred to the methadone addiction recovery centers. *International Academic Journal of Social Sciences*, 1(2), 91-96.
- [3] Azam, F., Agro, M., Sami, M., Abro, M. H., & Dewani, A. (2021, April). Identifying depression among twitter users using sentiment analysis. In *International Conference on Artificial Intelligence (ICAI)* (pp. 44-49). <https://doi.org/10.1109/ICAI52203.2021.9445271>
- [4] Basiri, M. E., Nemati, S., Abdar, M., Asadi, S., & Acharya, U. R. (2021). A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. *Knowledge-Based Systems*, 228, 107242. <https://doi.org/10.1016/j.knosys.2021.107242>

- [5] Bendebane, L., Laboudi, Z., & Saighi, A. (2023, September). Mental Disorders prediction from twitter data: application to syndromic surveillance systems. In *Novel & Intelligent Digital Systems Conferences* (pp. 140-145). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44097-7_14
- [6] Chauhan, S., & Desai, M. (2022). Machine Learning-Based Predictive Maintenance Scheduling in Industrial Settings. *International Academic Journal of Science and Engineering*, 9(4), 9–12. <https://doi.org/10.71086/IAJSE/V9I4/IAJSE0929>
- [7] Ghosh, S., & Anwar, T. (2021). Depression intensity estimation via social media: A deep learning approach. *IEEE Transactions on Computational Social Systems*, 8(6), 1465-1474. <https://doi.org/10.1109/TCSS.2021.3084154>
- [8] Govindasamy, K. A., & Palanichamy, N. (2021, May). Depression detection using machine learning techniques on twitter data. In *2021 5th international conference on intelligent computing and control systems (ICICCS)* (pp. 960-966). IEEE. <https://doi.org/10.1109/ICICCS51141.2021.9432203>
- [9] Health, T. L. G. (2020). Mental health matters. *The Lancet. Global Health*, 8(11), e1352. [https://doi.org/10.1016/S2214-109X\(20\)30432-0](https://doi.org/10.1016/S2214-109X(20)30432-0)
- [10] Jain, S., Narayan, S. P., Dewang, R. K., Bhartiya, U., Meena, N., & Kumar, V. (2019, May). A machine learning based depression analysis and suicidal ideation detection system using questionnaires and twitter. In *2019 IEEE students conference on engineering and systems (SCES)* (pp. 1-6). IEEE. <https://doi.org/10.1109/SCES46477.2019.8977211>
- [11] Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1), 11846. <https://doi.org/10.1038/s41598-020-68764-y>
- [12] Kim, J., Uddin, Z. A., Lee, Y., Nasri, F., Gill, H., Subramanieapillai, M., ... & McIntyre, R. S. (2021). A systematic review of the validity of screening depression through Facebook, Twitter, Instagram, and Snapchat. *Journal of affective disorders*, 286, 360-369. <https://doi.org/10.1016/j.jad.2020.08.091>
- [13] Kute, R. (2022). Mental health analyzer for depression detection based on textual analysis. *Journal of Advances in Information Technology*. 13(1). 67-77. <https://doi.org/10.12720/jait.13.1.67-77>
- [14] Le Glaz, A., Haralambous, Y., Kim-Dufor, D. H., Lenca, P., Billot, R., Ryan, T. C., ... & Lemey, C. (2021). Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, 23(5), e15708. <https://doi.org/10.2196/15708>
- [15] Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., & Leung, H. (2020, June). Sensemood: depression detection on social media. In *Proceedings of the 2020 international conference on multimedia retrieval* (pp. 407-411). <https://doi.org/10.1145/3372278.3391932>
- [16] Moreno, C., Wykes, T., Galderisi, S., Nordentoft, M., Crossley, N., Jones, N., ... & Arango, C. (2020). How mental health care should change as a consequence of the COVID-19 pandemic. *The lancet psychiatry*, 7(9), 813-824. [https://doi.org/10.1016/S2215-0366\(20\)30307-2](https://doi.org/10.1016/S2215-0366(20)30307-2)
- [17] Musleh, D. A., Alkhales, T. A., Almakki, R. A., Alnajim, S. E., Almarshad, S. K., Alhasaniah, R. S., ... & Almuqhim, A. A. (2022). Twitter arabic sentiment analysis to detect depression using machine learning. *Computers, Materials, & Continua*, 71(2), 3463-3477. <https://doi.org/10.32604/cmc.2022.022508>
- [18] Pool-Cen, J., Carlos-Martínez, H., Hernández-Chan, G., & Sánchez-Siordia, O. (2023, April). Detection of depression-related tweets in Mexico using crosslingual schemes and knowledge distillation. In *Healthcare* (Vol. 11, No. 7, p. 1057). MDPI. <https://doi.org/10.3390/healthcare11071057>
- [19] Pradhan, R., & Sharma, D. K. (2023). Retracted Article: An ensemble deep learning classifier for sentiment analysis on code-mix Hindi–English data. *Soft Computing*, 27(15), 11053-11053. <https://doi.org/10.1007/s00500-022-07091-y>

- [20] Santos, W. R. D., de Oliveira, R. L., & Paraboni, I. (2024). SetembroBR: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*, 58(1), 273-300. <https://doi.org/10.1007/s10579-022-09633-0>
- [21] Skaik, R., & Inkpen, D. (2020, December). Using twitter social media for depression detection in the canadian population. In *Proceedings of the 2020 3rd Artificial Intelligence and Cloud Computing Conference* (pp. 109-114). <https://doi.org/10.1145/3442536.3442553>
- [22] Squarcina, L., Villa, F. M., Nobile, M., Grisan, E., & Brambilla, P. (2021). Deep learning for the prediction of treatment response in depression. *Journal of affective disorders*, 281, 618-622. <https://doi.org/10.1016/j.jad.2020.11.104>
- [23] Victor, D. B., Kawsher, J., Labib, M. S., & Latif, S. (2020, November). Machine learning techniques for depression analysis on social media-case study on bengali community. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1118-1126). IEEE. <https://doi.org/10.1109/ICECA49313.2020.9297436>
- [24] Zhou, J., Zogan, H., Yang, S., Jameel, S., Xu, G., & Chen, F. (2021). Detecting community depression dynamics due to COVID-19 pandemic in Australia. *IEEE Transactions on Computational Social Systems*, 8(4), 982-991. <https://doi.org/10.1109/TCSS.2020.3047604>
- [25] Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1), 281-304. <https://doi.org/10.1007/s11280-021-00992-2>

Authors Biography



Ali Sami Azeez obtained the B.Sc. degree from Diyala University, Diyala, Iraq, and the M.Sc. degree from Bharati Vidyapeeth University, Pune, India. He is currently a Lecturer at the Middle Technical University (MTU) in Baghdad, Iraq, and a PhD researcher specializing in artificial intelligence and Natural Language Processing.



Osama Abduljaleel Ali received the bachelor's degree from Al-Rafidain University College in Baghdad, Iraq, and the master's degree from South Ural State University in Chelyabinsk, Russia. He is currently working as a lecturer at Al-Muthanna University. His research interests focus on natural language processing and artificial intelligence, with particular attention to computational methods for text analysis.



Nawar Abbood Fadhil holds a bachelor's degree from Middle Technical University and a master's degree in information technology from Universiti Utara Malaysia (UUM). He currently works as a lecturer in the Department of Information Technology at the Technical College of Management, Baghdad, Middle Technical University. His research interests include information technology, data analysis, and modern digital systems.



Dr. Ali Mohammed Sahan is a Professor at the Department of technical information management, Middle Technical University, Baghdad, Iraq. His research interest included Artificial Intelligent, image and video processing, He has published and reviewed several research articles in Scopus and Clarivate journals.