International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

Comparative Analysis of Logistic Regression and XGBoost for Depression Detection from Reddit Posts

Saransh Tripathi[1], Pramod Singh[2], Nikhil Tripathi[3]
Department of Computer Science, AKS University, SATNA, India

***Abstract*-- Depression is a major mental health disorder that affects millions of individuals worldwide and often remains undiagnosed due to social stigma and limited access to professional care. The increasing use of social media platforms such as Reddit provides an opportunity to analyze textual expressions that may contain indicators of depressive behavior. This study investigates the effectiveness of machine learning techniques for depression detection using textual data collected from Reddit posts.**

A Reddit Depression Dataset containing 7,731 posts was analyzed through exploratory data analysis, text preprocessing, TF-IDF feature extraction, and VADER sentiment analysis. The extracted features were evaluated using two machine learning classifiers: Logistic Regression and XGBoost. Performance was assessed using accuracy, precision, recall, F1-score, confusion matrix analysis, and ROC-AUC.

Experimental results demonstrated strong classification performance. Logistic Regression achieved an accuracy of 94.44% and an AUC of 0.986, while XGBoost achieved a slightly higher accuracy of 94.70%. The findings indicate that TF-IDF lexical features provide substantial predictive information for distinguishing depressive and non-depressive posts. Sentiment analysis further revealed noticeable differences in emotional polarity between the two classes.

The study presents a reproducible and computationally efficient framework for depression detection using publicly available data and open-source tools. The proposed workflow is suitable for academic research and educational environments where interpretability, simplicity, and reproducibility are important considerations.

***Keywords*-- Machine Learning, Depression Detection, Natural Language Processing, TF-IDF, XGBoost, Logistic Regression, Reddit, Mental Health Analytics**

I. INTRODUCTION

Mental health disorders represent one of the most significant global health challenges of the twenty-first century. Among these, depression, formally classified as Major Depressive Disorder (MDD), is a mood disorder characterized by persistent sadness, loss of interest in daily activities, and impaired social and occupational functioning [1]. According to the World Health Organization (WHO), depression affects approximately 280 million people worldwide and remains one of the leading causes of disability across all age groups [2].

Despite its widespread prevalence, depression often remains underdiagnosed and undertreated due to factors such as social stigma, limited access to mental healthcare services, and challenges associated with conventional clinical assessment procedures [2]. Traditional diagnosis typically relies on clinician interviews and standardized screening instruments such as the Patient Health Questionnaire (PHQ-9), which, although effective, may be resource-intensive and difficult to deploy at large population scales [1].

The rapid growth of social media platforms has created new opportunities for studying human behavior and mental health at scale. Platforms such as Reddit, Twitter, and Facebook allow users to openly share personal experiences, emotions, and daily challenges through textual communication. Previous research has shown that individuals experiencing depression often exhibit distinctive linguistic patterns, including increased use of first-person singular pronouns, negative emotional vocabulary, and expressions of hopelessness or social withdrawal [3]. These observable language characteristics provide valuable signals that can be analyzed using computational techniques for the early identification of depression-related content [4].

Natural Language Processing (NLP) has emerged as a powerful approach for analyzing mental health indicators from social media text. Early studies employed lexical feature extraction techniques such as n-grams and Term Frequency-Inverse Document Frequency (TF-IDF), combined with traditional machine learning algorithms including Support Vector Machines (SVM) and Logistic Regression for depression detection [5]. While such approaches may not fully capture complex contextual nuances such as sarcasm, irony, or implicit emotional expression, they offer important advantages including interpretability, computational efficiency, and reproducibility [6]. These characteristics make classical machine learning approaches particularly suitable for educational research environments and resource-constrained settings where transparency and ease of implementation are important considerations.

This study investigates depression detection from Reddit posts using a combination of exploratory data analysis, TF-IDF feature extraction, sentiment analysis, and machine learning classification.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

The objective is to develop a computationally efficient and reproducible workflow that can be implemented using publicly available datasets and open-source tools.

Two machine learning classifiers, Logistic Regression and XGBoost [7], are evaluated using TF-IDF representations of Reddit posts. In addition, VADER sentiment analysis [8] is employed to examine the emotional characteristics of depressive and non-depressive content. The resulting framework emphasizes simplicity, interpretability, and reproducibility while maintaining strong predictive performance.

II. Related Work

2.1 Depression Detection from Social Media

Automated detection of depression from social media text has been an active area of research over the past decade. Early studies by De Choudhury et al. [4] demonstrated that behavioral and linguistic patterns extracted from Twitter could be used to identify indicators of depression, establishing the feasibility of social media-based mental health monitoring. Subsequently, Reddit emerged as a valuable resource for mental health research due to its community-based structure and the presence of dedicated support forums such as r/depression and r/SuicideWatch, where users often discuss mental health experiences openly [5].

Shen and Rudzicz [5] employed machine learning techniques incorporating textual and user-level information from Reddit data, demonstrating that contextual information can improve detection performance. Yates et al. [6] introduced models that utilized user posting histories for depression and self-harm risk assessment, highlighting the value of information distributed across multiple posts. Orabi et al. [7] investigated deep learning approaches for depression detection from social media text and reported competitive performance compared with traditional machine learning methods. Collectively, these studies demonstrate the feasibility and importance of computational depression detection while motivating the exploration of effective and reproducible feature representations.

2.2 Sentiment Analysis in Mental Health

Sentiment analysis has been widely applied in mental health text mining as a means of capturing affective states that may be associated with psychological conditions. Lexicon-based approaches, particularly VADER (Valence Aware Dictionary and sEntiment Reasoner) [8], have gained considerable popularity due to their effectiveness on informal social media text and their ability to operate without supervised training.

VADER generates four sentiment measures, including positive, negative, neutral, and compound sentiment scores, thereby providing a multidimensional representation of emotional polarity.

Coppersmith et al. [9] investigated linguistic and emotional patterns expressed by individuals experiencing mental health conditions on social media and reported significant differences in affective language compared with control populations. Mowery et al. [10] further demonstrated the usefulness of linguistic and sentiment-related indicators for identifying depression-related content in online discussions. Collectively, these findings support the use of sentiment analysis as a complementary exploratory tool for understanding the emotional characteristics of depression-related text.

III. Methodology

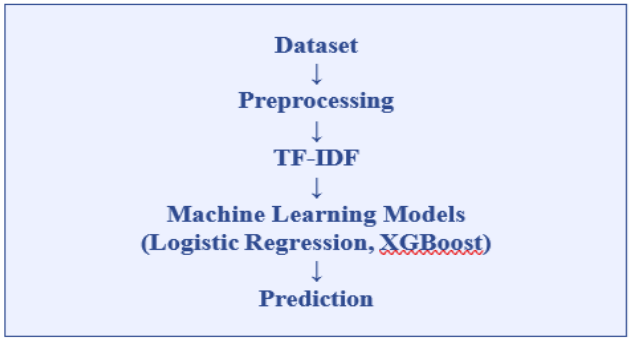

Figure 1. Overview of the Depression Detection Workflow.

3.1 Dataset

The dataset employed in this study was obtained from the publicly available Depression Reddit Cleaned Dataset available on Kaggle [11]. The dataset consists of Reddit posts collected from depression-related subreddits (e.g., r/depression and r/suicidewatch) and control subreddits representing non-depressive general discussion. Reddit has become a widely used resource for computational depression detection research due to its self-disclosure nature and the availability of community-level information that can serve as a proxy for mental health status [5].

Posts labeled as depressive are drawn from users who have explicitly sought mental health support or discussed depression-related experiences within dedicated subreddits. Control posts are sampled from general discussion communities to provide a balanced representation of non-depressive language. All posts are anonymized, and no personally identifiable information (PII) is retained in the dataset. Dataset statistics are reported in Table 1 following the completion of initial data analysis.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

Table 1. Dataset Statistics

Dataset	Total Samples	Positive (Depressive)	Negative (Control)	Source
Reddit Depression	7731	3831	3900	Reddit API / Kaggle

3.1.1 Exploratory Dataset Analysis

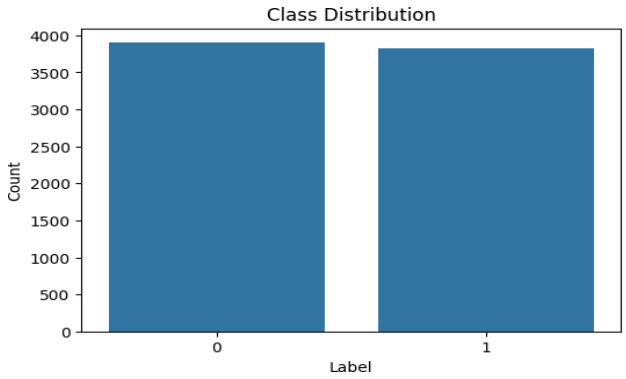

Figure 2. Class Distribution

Figure 2 illustrates the class distribution of the Reddit Depression Dataset. The dataset contains 7,731 samples, comprising 3,831 depressive posts and 3,900 non-depressive posts. The near-balanced class distribution minimizes the risk of classification bias toward a dominant class and eliminates the immediate need for oversampling techniques.

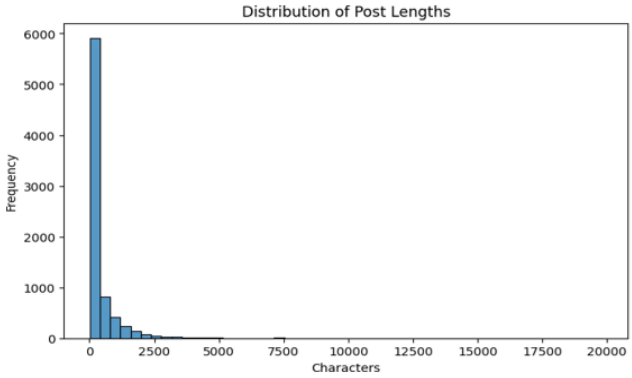

Figure 3. Distribution of Post Lengths

Figure 3 presents the distribution of post lengths measured in characters. The dataset exhibits a positively skewed distribution, where most posts are relatively short while a small number of posts contain substantially longer narratives. The average post length is 361 characters, with a median length of 110 characters. This variation in post length highlights the importance of feature extraction techniques capable of handling both concise and lengthy textual expressions of emotional states.

3.2 Data Preprocessing

The raw Reddit posts were preprocessed through a sequence of standard text-cleaning operations, including URL removal, HTML decoding, username and subreddit reference removal, lowercasing, special character filtering, stopword removal, lemmatization, and short-text filtering. These preprocessing steps reduced textual noise and improved feature quality for subsequent machine learning analysis.

Standard English stopwords were removed using the Natural Language Toolkit (NLTK) stopword corpus [12], while important negation terms such as *not* and *never* were retained due to their relevance in expressing emotional and psychological states. Furthermore, lemmatization was performed using the spaCy natural language processing library [13] to reduce words to their base forms and minimize vocabulary dimensionality.

3.3 TF-IDF Feature Extraction

Term Frequency-Inverse Document Frequency (TF-IDF) was employed to convert textual Reddit posts into numerical feature representations suitable for machine learning models. A TF-IDF vectorizer from the scikit-learn library [14] was configured with a maximum vocabulary size of 5,000 features and English stopword removal. After fitting the vectorizer on the dataset, a sparse feature matrix of dimension (7,731 × 5,000) was generated, where each row corresponds to a Reddit post and each column represents a weighted lexical feature.

Analysis of the extracted TF-IDF features revealed that terms such as "depression", "feel", "life", "anxiety", and "people" were among the most prominent words within the corpus, indicating that the dataset contains substantial mental health-related linguistic content. These lexical representations provide interpretable indicators of depressive language patterns and form the primary feature representation used for machine learning classification in this study.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

3.4 VADER Sentiment Features

VADER (Valence Aware Dictionary and Sentiment Reasoner) [8] was employed to extract sentiment-based features from Reddit posts. For each document, four sentiment indicators were generated: positive score, negative score, neutral score, and compound sentiment score. These features provide a lightweight and interpretable representation of emotional polarity and were utilized to explore the emotional characteristics of depression-related text within the dataset.

Statistical analysis of the extracted sentiment scores revealed an average compound sentiment value of - 0.166, indicating an overall tendency toward negative emotional expression within the dataset. The mean negative sentiment score (0.173) exceeded the mean positive sentiment score (0.116), further supporting the emotional characteristics expected in depression-related content.

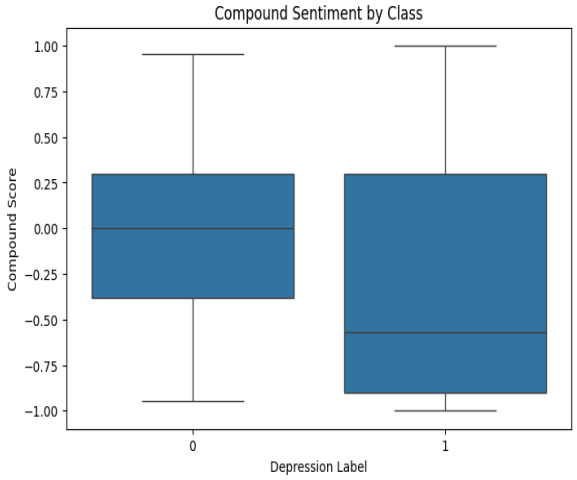

Figure 4. Comparison of compound sentiment scores between depressive and non-depressive classes.

Figure 4 compares compound sentiment scores across depressive and non-depressive posts. The depressive class exhibits substantially lower median sentiment values than the non-depressive class, indicating notable differences in emotional expression between the two groups. This observation supports the exploratory sentiment analysis conducted in the study.

3.5 Machine Learning Classification

Following feature extraction, machine learning classifiers were trained to distinguish depressive and non-depressive Reddit posts. Two classification algorithms were evaluated in this study: Logistic Regression and XGBoost [7].

Logistic Regression was selected as a strong and interpretable baseline model for text classification. The classifier was trained using TF-IDF feature representations generated from the preprocessed Reddit posts.

XGBoost was employed as a gradient-boosted tree-based classifier capable of learning complex decision boundaries from high-dimensional feature spaces. The model was trained using the same TF-IDF feature representation to enable a fair comparison with Logistic Regression.

The performance of both classifiers was evaluated using an identical train-test split and assessed through accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis.

Logistic Regression was implemented using default scikit-learn settings with random_state=42.

XGBoost was implemented using the XGBoost library with random_state=42.

IV. Experimental Setup

All experiments were conducted using Python and open-source machine learning libraries within the Google Colab environment. The workflow included data preprocessing, TF-IDF feature extraction, sentiment analysis using VADER [8], model training, and performance evaluation.

The Reddit Depression Dataset consisting of 7,731 samples was divided into training and testing sets using an 80:20 stratified split. TF-IDF feature extraction was performed using the scikit-learn library [14] with a maximum vocabulary size of 5,000 features. VADER sentiment analysis [8] was performed to examine emotional characteristics of depressive and non-depressive posts through positive, negative, neutral, and compound sentiment scores.

Two machine learning classifiers were evaluated: Logistic Regression and XGBoost [7]. Both models were trained using the same TF-IDF feature representation to ensure a fair comparison. Model performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis.

V. Results And Discussion

Table 2. Baseline Model Performance Comparison.

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	94.44%	0.97	0.91	0.94	0.986
XGBoost (TF-IDF)	94.70%	0.99	0.91	0.94	0.981

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

5.1 Logistic Regression Baseline

A Logistic Regression classifier was trained using TF-IDF features extracted from the Reddit Depression Dataset. The dataset was divided into training and testing subsets using an 80:20 split ratio. Experimental evaluation demonstrated strong predictive performance, achieving an overall accuracy of 94.44%.

For the depression class, the model achieved a precision of 0.97, recall of 0.91, and F1-score of 0.94. These results indicate that lexical features extracted through TF-IDF provide substantial discriminatory power for depression detection in social media text and establish a strong baseline for subsequent experiments.

In addition to accuracy-based metrics, the Logistic Regression classifier achieved an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.986. This high AUC value indicates strong discriminative capability and demonstrates that the model effectively distinguishes between depressive and non-depressive posts across different classification thresholds. The result further confirms the suitability of TF-IDF lexical representations for depression detection tasks.

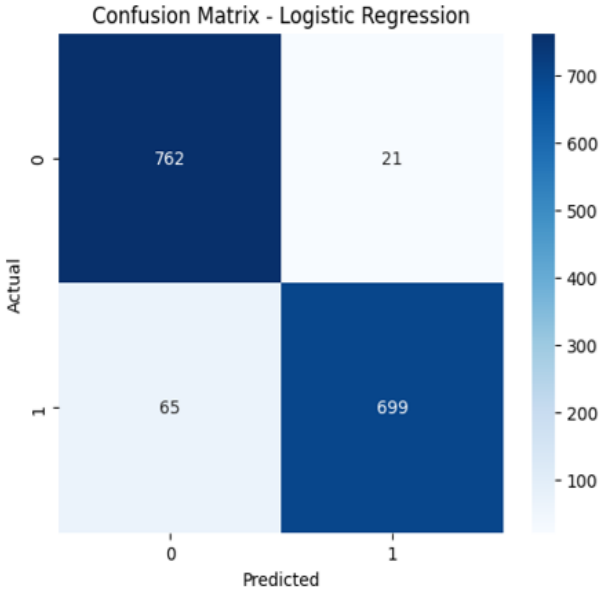

Figure 5. Confusion Matrix of the Logistic Regression Classifier.

Figure 5 illustrates the confusion matrix obtained from the Logistic Regression classifier. The majority of depressive and non-depressive samples were correctly classified, with only a limited number of false positives and false negatives.

This observation is consistent with the strong classification metrics reported in Table 4 and demonstrates the effectiveness of TF-IDF features for identifying depression-related textual patterns.

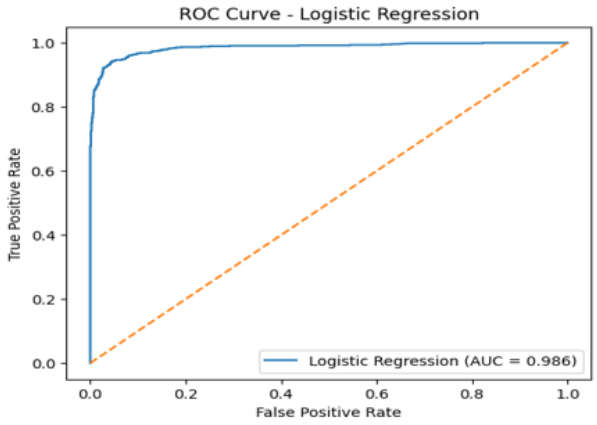

Figure 6. ROC Curve of the Logistic Regression Classifier

Figure 6 presents the Receiver Operating Characteristic (ROC) curve of the Logistic Regression classifier. The model achieved an AUC value of 0.986, indicating excellent discriminative capability between depressive and non-depressive samples. The ROC curve demonstrates that the classifier maintains a high true positive rate while keeping the false positive rate low across different decision thresholds.

5.2 XGBoost Classification Results

The XGBoost classifier was evaluated using the same TF-IDF feature representation and train-test split configuration employed for the Logistic Regression baseline. Experimental results demonstrated a classification accuracy of 94.70%, slightly outperforming the Logistic Regression model, which achieved an accuracy of 94.44%.

For the depression class, XGBoost achieved a precision of 0.99, recall of 0.91, and an F1-score of 0.94. The high precision indicates that the model generated relatively few false positive predictions while maintaining strong overall classification performance.

A comparison between Logistic Regression and XGBoost reveals that both models perform effectively on the Reddit Depression Dataset. Although XGBoost achieved the highest overall accuracy, the performance difference between the two models was relatively small (0.26%). This observation suggests that TF-IDF lexical features contain substantial predictive information for depression detection, enabling strong performance across different machine learning classifiers.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

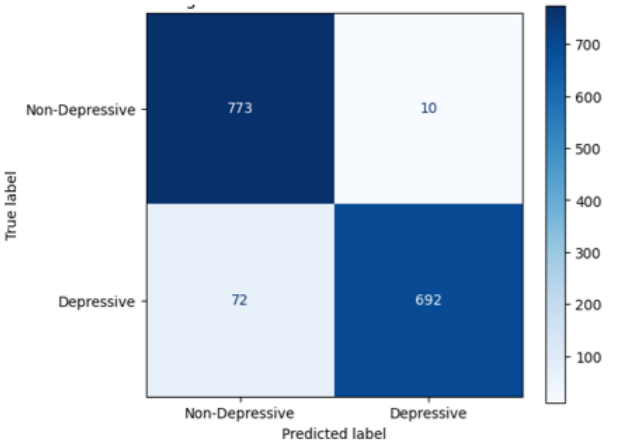

Figure 7. Confusion Matrix of the XGBoost Classifier

Figure 7 illustrates the confusion matrix obtained from the XGBoost classifier. Most depressive and non-depressive posts were correctly classified, indicating strong predictive performance. The low number of false positive and false negative predictions is consistent with the high classification accuracy reported in Table 2.

5.3 Comparative Analysis

Table 2 summarizes the performance of the evaluated machine learning models. Logistic Regression achieved an accuracy of 94.44%, while XGBoost achieved a slightly higher accuracy of 94.70%. Both models demonstrated strong classification performance, with F1-scores of 0.94 for the depression class.

The experimental results indicate that TF-IDF lexical representations provide substantial predictive information for depression detection. Although XGBoost produced the highest overall accuracy, the improvement over Logistic Regression was relatively small (0.26%). This finding suggests that the choice of feature representation may have a greater impact on performance than the selection of the classification algorithm.

Overall, both models demonstrated strong capability in identifying depression-related textual patterns within Reddit posts, confirming the effectiveness of machine learning approaches for social media-based depression detection.

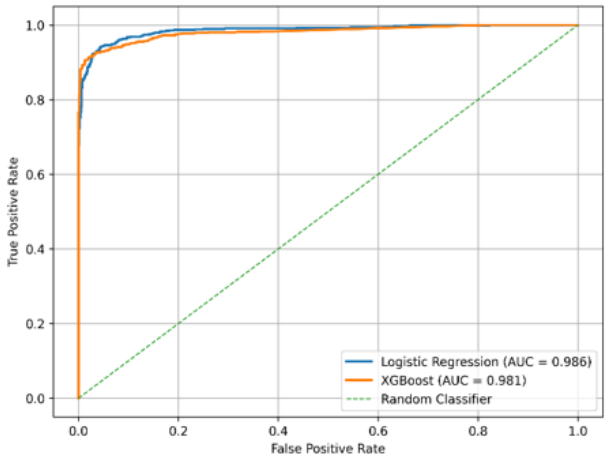

Figure 8. ROC Curve Comparison between Logistic Regression and XGBoost

Figure 8 compares the ROC curves of Logistic Regression and XGBoost. Both models achieved excellent discriminative performance with AUC values above 0.98. Although XGBoost achieved slightly higher classification accuracy, Logistic Regression achieved a marginally higher ROC-AUC score. The results indicate that both models are highly effective for depression detection using TF-IDF features.

VI. Limitations And Ethical Considerations

Several limitations should be considered when interpreting the findings of this study. The dataset consists exclusively of Reddit posts and may not generalize to other social media platforms or clinical settings. Dataset labels are derived from subreddit participation rather than clinically verified diagnoses.

Furthermore, TF-IDF features may not fully capture contextual language nuances such as sarcasm or implicit emotional expression. The dataset used in this study consists of publicly available Reddit posts that were anonymized prior to analysis. No personally identifiable information was retained, and the proposed system is intended solely for research purposes rather than clinical diagnosis.

VII. Conclusion

This study investigated the use of machine learning techniques for depression detection from Reddit posts. A publicly available Reddit Depression Dataset containing 7,731 samples was analyzed using exploratory data analysis, TF-IDF feature extraction, and VADER sentiment analysis. The dataset exhibited a balanced class distribution and a wide variation in textual content length, making it suitable for evaluating machine learning approaches for depression detection.

Experimental evaluation demonstrated that TF-IDF features provide strong predictive capability for identifying depression-related textual patterns. Logistic Regression achieved an accuracy of 94.44% with an AUC of 0.986, while XGBoost achieved a slightly higher accuracy of 94.70%. The small performance difference between the two models suggests that feature representation plays a significant role in classification effectiveness.

The results indicate that machine learning models can effectively distinguish between depressive and non-depressive Reddit posts using lexical features extracted from textual content. The study contributes a reproducible framework that can be implemented using publicly available datasets and open-source tools, making it suitable for academic and educational research environments.

Future work may investigate contextual language models such as DistilBERT, feature fusion strategies combining lexical and sentiment representations, and explainable artificial intelligence techniques to further improve model interpretability and predictive performance.

REFERENCES

[1] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders (DSM-5), 5th ed. Arlington, VA, USA: American Psychiatric Publishing, 2013.

[2] World Health Organization, "Depressive disorder (depression)," WHO Fact Sheet, Sep. 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

[3] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," Annu. Rev. Psychol., vol. 54, pp. 547-577, 2003.

[4] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. 7th Int. AAAI Conf. Weblogs and Social Media (ICWSM), 2013, pp. 128-137.

[5] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 2968-2978.

[6] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in Proc. 5th Workshop on Computational Linguistics and Clinical Psychology (CLPsych), 2018, pp. 88-97.

[7] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 785-794.

[8] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM), Ann Arbor, MI, USA, 2014, pp. 216-225.

[9] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying Mental Health Signals in Twitter," in Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (CLPsych), Baltimore, MD, USA, 2014, pp. 51-60.

[10] D. L. Mowery, C. Bryan, and M. Conway, "Feature Studies to Inform the Classification of Depressive Symptoms from Twitter Data for Population Health," in Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology (CLPsych), Vancouver, Canada, 2017, pp. 1-12.

[11] InfamousCoder, "Depression Reddit Cleaned Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/infamouscoder/depre ssion-reddit-cleaned. [Accessed: May 2026].

[12] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA, USA: O'Reilly Media, 2009.

[13] M. Honnibal and I. Montani, "spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing," 2017. [Online]. Available: https://spacy.io

[14] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.