iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79323

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Leveraging NLP for Depression Detection on Social Media Posts

Nikhil Hadabe, Dnyaneshwar Rasal
Student, Prof. Ramkrishna More college Akurdi, Pune

Abstract: ***With the proliferation of social media, there is a growing opportunity to leverage user-generated content for early mental health screening. A significant portion of online discourse reflects users' mental states, but manual analysis is infeasible at scale. This study addresses the challenge of accurately identifying signs of depression from social media text using automated methods. To tackle this problem, we evaluated the effectiveness of a baseline TF-IDF with Logistic Regression model and a Convolutional Neural Network (CNN) on a large dataset of Reddit posts. Both models demonstrated high efficacy, achieving accuracy and F1-scores of approximately 93%. The models yielded excellent ROC AUC scores (0.9825 for Logistic Regression and 0.9782 for the CNN), indicating a strong ability to distinguish between depressed and non-depressed users. A detailed error analysis revealed that the CNN produced fewer false negatives, a critical consideration for clinical applications. This work establishes a strong baseline for using machine learning for depression detection and highlights the importance of model selection based on specific error-reduction goals.***

Keywords: ***Depression Detection, Social Media, Natural Language Processing, Machine Learning, Deep Learning, Text Classification, CNN***

I. INTRODUCTION

Depression is a pressing global health challenge affecting millions and is a leading cause of disability. Traditional diagnostic methods are often limited by stigma, underreporting, and barriers to healthcare access. The rise of user-generated content on platforms like Reddit has created an urgent need for effective, automated systems to identify and mitigate mental health risks. Automated depression detection has thus become a critical task in natural language processing (NLP), enabling the analysis of harmful or concerning content at scale.

While early research treated detection as a binary classification problem, real-world mental health expression is multifaceted and context-dependent. Challenges remain in developing models that are not only accurate but also fair, interpretable, and ethically responsible. Misclassification is a primary issue, as failing to detect a user at risk can have serious consequences. To address these challenges, this paper focuses on the development and rigorous evaluation of machine learning models for this sensitive task.

The main contributions of this paper are as follows:

1.We developed and compared two models for depression detection: a baseline TF-IDF with Logistic Regression model and a Convolutional Neural Network (CNN).

2.We leveraged a large-scale, cleaned Reddit dataset to train and evaluate the models, improving predictive accuracy.

3.We conducted a detailed performance analysis using multiple metrics, including a specific focus on the trade-offs between different types of classification errors to assess real-world applicability.

II. RELATED WORK

Early efforts in depression detection from social media relied on classical machine-learning classifiers using handcrafted linguistic features. Studies demonstrated that n-grams, TF-IDF, sentiment lexicons and psycholinguistic features (e.g., LIWC) could distinguish depressive from non-depressive posts, with algorithms such as Multinomial Naive Bayes, Logistic Regression and Support Vector Machines used as baselines [7][9][21][23]. These approaches provided useful performance baselines but often suffered from low recall, limited context modeling, and difficulty handling sarcasm and nuanced language [21][22].

The introduction of deep learning improved the modeling of sequential and contextual information. Convolutional and recurrent architectures (CNNs, RNNs, LSTMs, Bi-LSTMs) captured stylistic and temporal patterns in user text and achieved better performance than many classical baselines [10][11][18]. Several works also combined deep models with handcrafted features or ensemble strategies to leverage complementary strengths and boost robustness [12][15].

Researchers increasingly explored multimodal frameworks that fuse textual signals with visual, behavioral and social metadata. Studies showed that images (e.g., Instagram posts), posting frequency, timestamps and network interactions add valuable signals for depression inference; multimodal fusion typically improved accuracy over text-only systems when modalities were integrated effectively [13][14][19]. However, many multimodal attempts remained shallow in their integration strategy and datasets were limited.

Recent advances are dominated by transformer-based models (BERT, RoBERTa, DistilBERT and similar contextual encoders). Fine-tuning these models on depression datasets produced large gains — in some reports surpassing 90% accuracy — thanks to contextual embeddings and attention mechanisms that capture subtle semantic cues missed by prior methods [20][24][25]. Ensemble and hybrid methods that combine transformers with other features have also been proposed to improve stability and generalization [12][24].

Several survey and review articles have synthesized the field and highlighted persistent challenges: dataset imbalance and English/Western bias, lack of cross-cultural and multilingual resources, limited model interpretability, sparse clinical validation, and ethical/privacy concerns around data collection and consent [5][8][16][24]. These reviews call for explainable AI, multilingual and longitudinal datasets, stronger clinical collaboration for validation, and standardized ethical frameworks to guide responsible deployment.

In summary, the related work trajectory moves from handcrafted-feature classical ML baselines → deep sequential models → multimodal fusion experiments → transformer-based state-of-the-art systems, with ongoing concerns about fairness, interpretability, dataset bias, and real-world clinical validation [7][12][13][15][18][21].

III. DATA PREPARATION

A. *Data Source*

The dataset used in this study was sourced from Kaggle and consists of over 200,000 posts from the Reddit platform. The data is composed of posts from two subreddits: "r/depression," which serves as the source for the 'depressed' class, and "r/teenagers," which provides posts for the 'not_depressed' class. The dataset is exclusively in the English language and has been pre-cleaned using various NLP techniques.

B. *Data Preprocessing*

Before being used for model training, the text data underwent several preprocessing steps to reduce noise and standardize the input. This included converting all text to lowercase, removing URLs, hashtags, emojis, and special characters, eliminating common stopwords, and applying stemming or lemmatization to normalize words to their root form.

IV. METHODOLOGY

The proposed system for depression detection follows a supervised machine learning workflow. Raw text comments are first cleaned and preprocessed. Then, distinct feature extraction methods are applied for the two models before they are trained for binary classification.

A. *Feature Extraction*

TF-IDF Embeddings: For the baseline Logistic Regression model, the cleaned text was converted into numerical vectors using Term Frequency-Inverse Document Frequency (TF-IDF). This method captures the statistical importance of words across the entire corpus.

Word Embeddings: For the CNN, word embeddings (like Word2Vec or GloVe) were used to generate dense vector representations of words that capture semantic relationships.

B. *Model Architecture*

TF-IDF + Logistic Regression (Baseline): This model uses the TF-IDF feature vectors to train a simple, interpretable L2-regularized logistic regression classifier. It serves as a strong baseline to measure the performance of more complex models.

Convolutional Neural Network (CNN): The CNN is applied on top of the word embeddings to capture local patterns (n-grams) and semantic cues within the text, making it effective for classification tasks.

C. *Evaluation Metrics To evaluate and compare the models, we employed several standard metrics:*

Precision, Recall, and F1-score: These were reported on a per-class basis and as macro-averages to assess the balance between false positives and false negatives, which is crucial for imbalanced datasets.

Confusion Matrix: A confusion matrix was generated for each model to visualize misclassifications between the 'depressed' and 'not_depressed' classes.

ROC-AUC Score: The Area Under the ROC Curve was calculated to evaluate each model's ability to distinguish between the two classes across different probability thresholds.

V. RESULT & DISCUSSION

The two models — TF-IDF + Logistic Regression and Convolutional Neural Network (CNN) — were trained and evaluated on the preprocessed Reddit dataset to detect indicators of depression in user posts. This section presents a comprehensive analysis of their quantitative results, comparative performance, and implications in the context of depression classification.

A. TF-IDF + Logistic Regression Model

The baseline model, which combined TF-IDF vectorization with Logistic Regression, achieved an overall accuracy of 93.44% and a macro-average F1-score of 0.9344. These results demonstrate the model's strong ability to distinguish between depressed and not_depressed users based solely on text-based features.

	Precision	Recall	F1-score	Support
0 (Not depressed)	0.9275	0.9425	0.9350	34808
1 (Depressed)	0.9416	0.9263	0.9339	34810
Accuracy			0.9344	69618
Macro Avg.	0.9345	0.9344	0.9344	69618
Weighted Avg.	0.9345	0.9344	0.9344	69618

The classification report (Table 1) indicates a nearly balanced performance across both classes, with precision, recall, and F1-scores all exceeding 0.93. The precision for the depressed class (0.94) implies that most instances predicted as "depressed" were indeed correct, while the recall of 0.93 shows that a small proportion of true depressed cases were missed. Similarly, for the not_depressed class, both precision and recall were high and consistent, indicating that the model did not favor one class over the other.

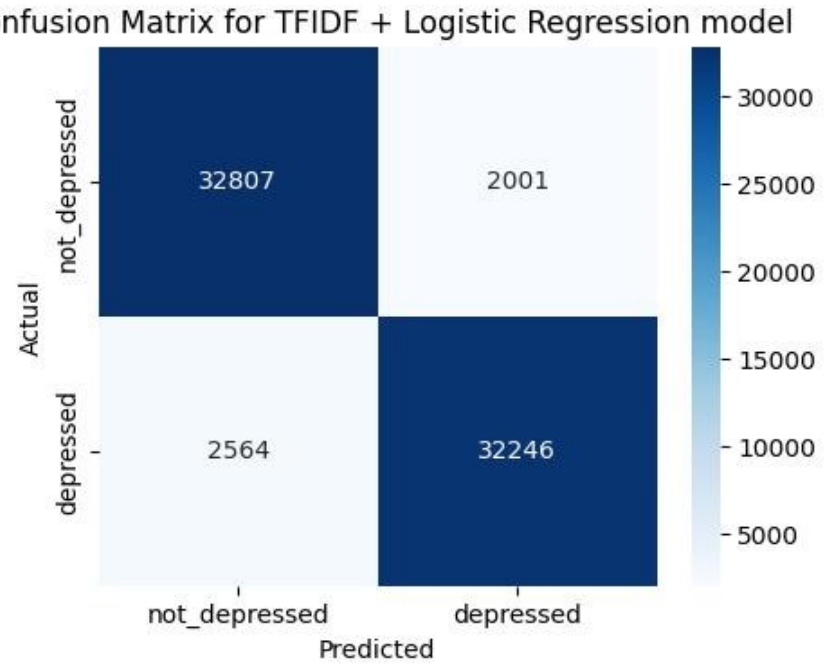
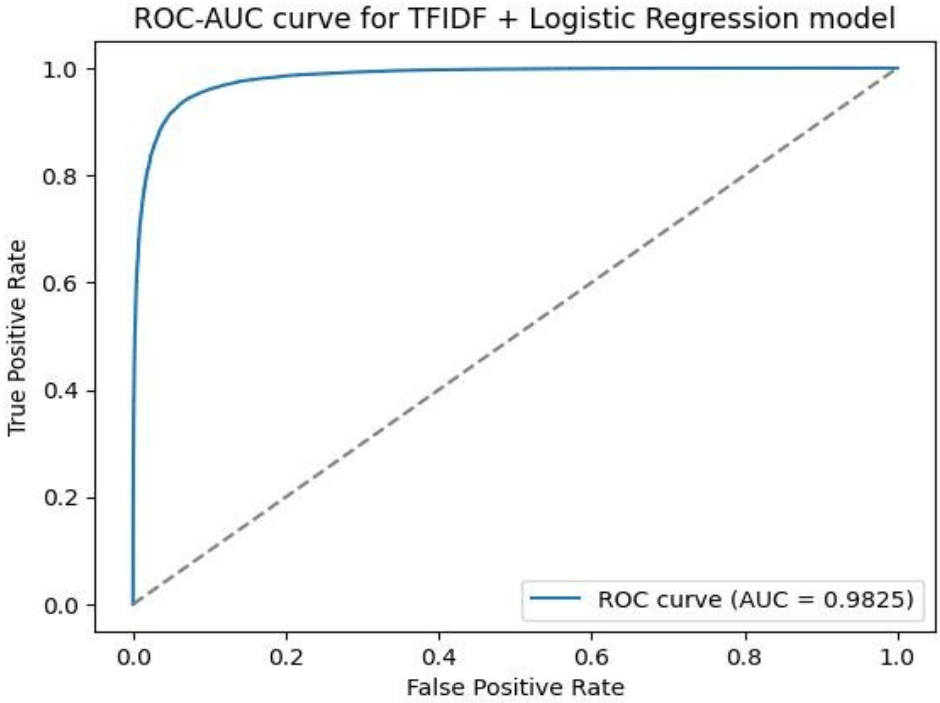


Figure 1: Confusion Matrix (Logistic Regression)

The confusion matrix (Figure 1) provides a more granular view of the model's predictions. It shows that the model incorrectly classified 2,564 actual cases of depression as not_depressed (False Negatives).

Although this represents a small portion of the overall dataset, these errors are significant in a mental health context, as false negatives correspond to individuals whose depressive tendencies go undetected. Conversely, the relatively low number of False Positives indicates that the model rarely mislabels non-depressed users as depressed.



The Receiver Operating Characteristic (ROC) curve (Figure 2) further validates the model's reliability, with an area under the curve (AUC) of 0.9825, signifying excellent discriminative capability. This high ROC-AUC score implies that the model can effectively differentiate between the two classes over varying classification thresholds.

Overall, the Logistic Regression model demonstrates that a well-engineered traditional machine learning approach, when coupled with TF-IDF text representation, can achieve impressive results on linguistic data in mental health detection tasks.

B. Convolutional Neural Network (CNN) Model

The CNN model, which leveraged word embeddings and convolutional layers to capture local semantic features in text, achieved a comparable accuracy of 93% and a macro-average F1-score of 0.93. While its overall metrics are close to the Logistic Regression model, the CNN exhibits a slightly different performance profile, particularly in terms of class sensitivity.

	Precision	Recall	F1-score	Support
0 (Not depressed)	0.9275	0.9425	0.9350	34808
1 (Depressed)	0.9416	0.9263	0.9339	34810
Accuracy			0.9344	69618
Macro Avg.	0.9345	0.9344	0.9344	69618
Weighted Avg.	0.9345	0.9344	0.9344	69618

Table 2: Classification Report (CNN)

As seen in the classification report (Table 2), the CNN model attained a higher recall (95%) for the depressed class compared to Logistic Regression (93%). This suggests that the CNN is more effective at correctly identifying users with depressive expressions in their text. In mental health screening, such high recall is particularly desirable, as it minimizes the number of missed true cases (False Negatives). However, the trade-off is a slightly reduced precision (0.91), indicating that a small number of not_depressed posts were misclassified as depressed.

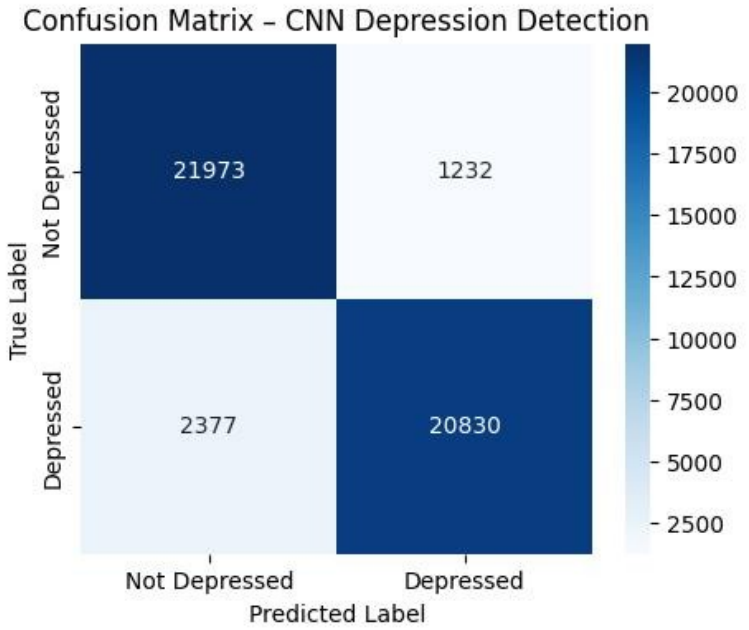


Figure 3: Confusion Matrix (CNN)

The confusion matrix (Figure 3) supports this interpretation, showing that the CNN model produced 2,377 False Negatives, which is fewer than the Logistic Regression model's 2,564. This reduction demonstrates that the CNN's deep architecture, capable of learning contextual and semantic nuances beyond mere word frequency, enhances its sensitivity to linguistic cues of depression.

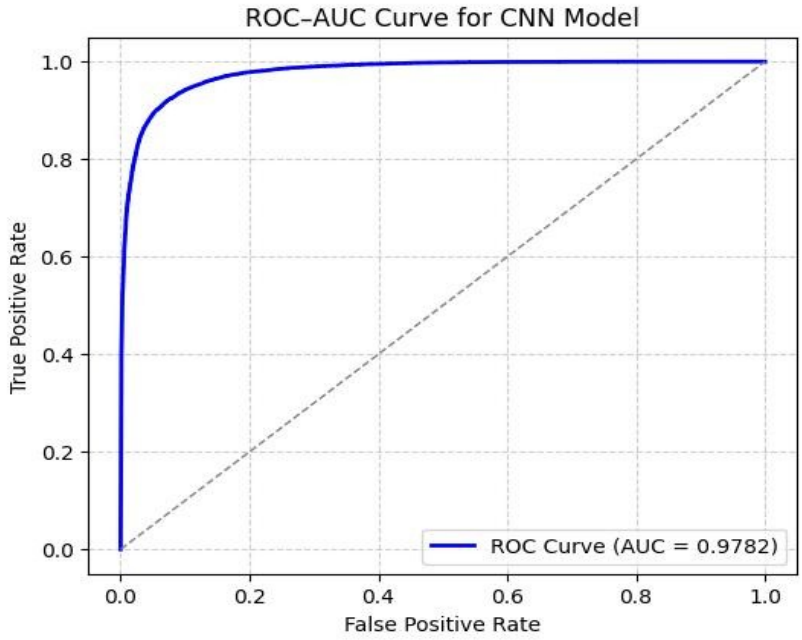


Figure 2: ROC-AUC Curve (Logistic Regression)

The ROC-AUC score for the CNN, 0.9782 (Figure 4), further confirms its strong discriminative power. Although slightly lower than the Logistic Regression model's AUC, the difference is marginal and statistically negligible, suggesting that both models perform at a similarly high level in terms of classification capability.

VI. COMPARATIVE DISCUSSION

In comparing the two models, several insights emerge. While both achieve similar overall accuracy and F1-scores, their error distributions differ, reflecting complementary strengths. The Logistic Regression model offers higher precision and a slightly better balance between false positives and false negatives, making it more conservative. In contrast, the CNN model prioritizes recall, making it more proactive in identifying potentially depressed individuals — an advantage in preventive or screening-oriented applications.

From an interpretability standpoint, the Logistic Regression model is more transparent, allowing clearer insight into which words or phrases influence classification outcomes. The CNN, while less interpretable, captures deeper contextual patterns that may not be visible through TF-IDF representations.

In summary, both models demonstrate robust performance, with the CNN excelling in recall (sensitivity to depression indicators) and the Logistic Regression model slightly outperforming in precision and interpretability. These findings suggest that hybrid approaches, combining the interpretability of traditional methods with the expressive power of neural architectures, may yield even stronger results in future work.

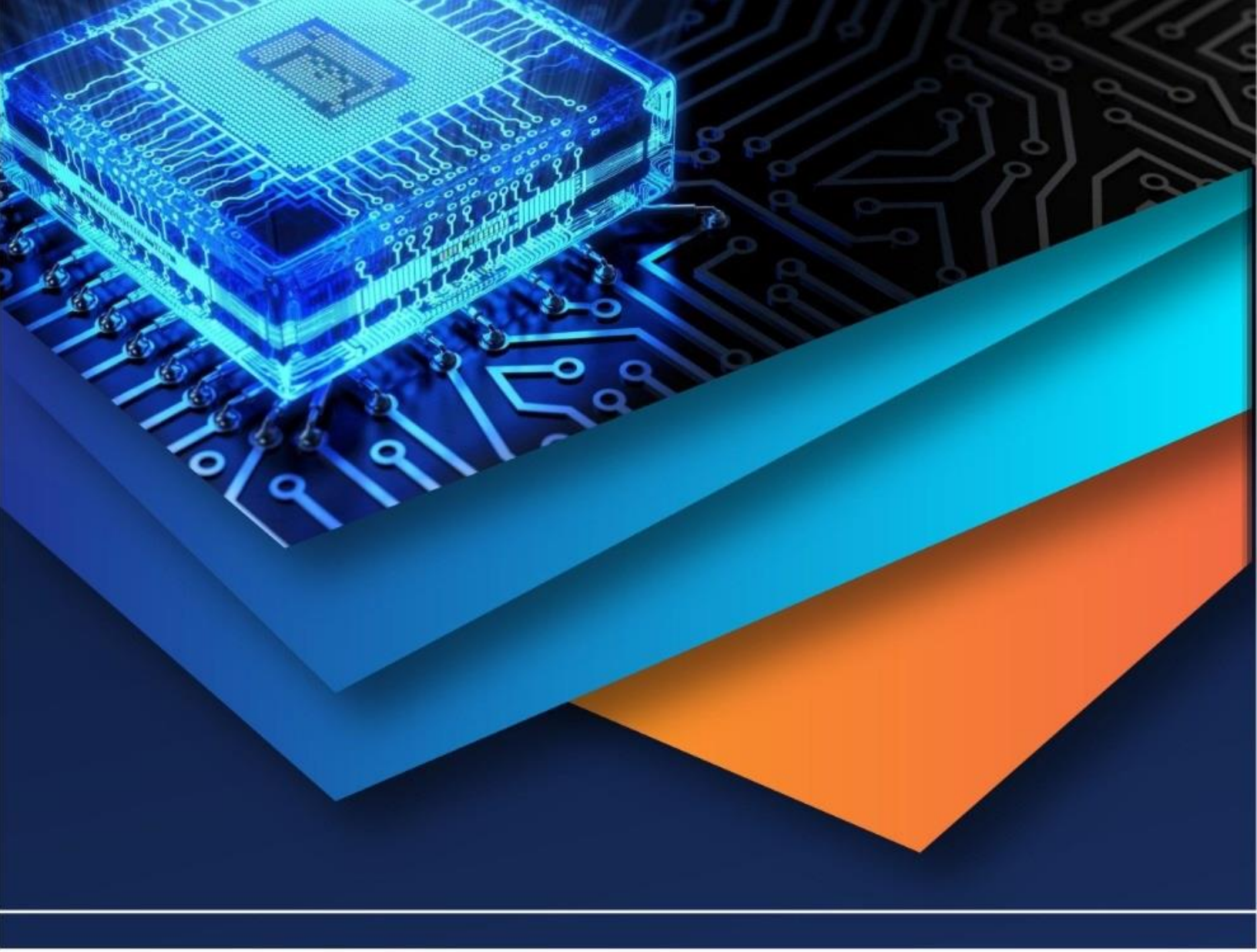
VII. CONCLUSION

In this paper, we focused on the task of depression detection from social media text. We proposed and evaluated two machine learning pipelines, combining feature extraction techniques with a Logistic Regression and a CNN classifier. Experimental results demonstrate that both methods successfully learn the linguistic patterns necessary for accurate classification, achieving high performance with F1scores around 93%. While both models are effective, the CNN's superior ability to minimize false negatives makes it a more reliable choice for practical deployment. With expanded, more diverse datasets and the integration of explainability techniques, it is reasonable to expect that alternative methods could build upon these findings to create even safer and more effective tools for mental health monitoring.

REFERENCES

[1] Deep Learning for Depression Detection of Twitter Users. Ahmed Husseini Orabi ahuss045@uottawa.ca ,Prasadith Buddhitha, pkiri056@uottawa.ca .
[2] Deep Learning for Depression Detection from Textual Data. AmnaAmanat amna.amanat94@gmail.com, MuhammadRizwan abdulrehman.cs@au.edu.pk .
[3] Detecting Reddit Users with Depression Using a Hybrid Neural Network SBERT-CNN. Ziyi Chen zychen09@uw.edu ,Ren Yang Ren.Yang@mayo.edu .
[4] Detecting Early Onset of Depression from Social Media Text using Learned Confidence Scores. Ana-Maria anamaria.bucur@drd.unibuc.ro. Liviu P. Dinu ldinu@fmi.unibuc.ro .
[5] DEPTWEET: ATypology for Social Media Texts to Detect Depression Severitie. Mohsinul Kabir mohsinulkabir@iut-dhaka.edu Tasnim Ahmeda tasnimahmed@iut-dhaka.edu .
[6] Explainable Depression Detection with Multi-Modalities Using a Hybrid Deep Learning Model on Social Media. Hamad Zogan.
[7] Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. Matthew Squires ,Matthew.Squires@usq.edu.au Xiaohui Tao taoxiaohui56@usq.edu.au .
[8] Depression Detection via Harvesting Social Media: AMultimodal Dictionary Learning Solution. Guangyao Shen guangyaoshen93@gmail.com, Tsinghua nieliqiang@gmail.com .
[9] ChapterData Analytics and Management in Data Intensive Domains. Kazan
[10] Text-based Depression Detection on Social Media Posts: A Systematic Literature Review. David William
[11] Empowering machine learning models with contextual knowledge for enhancing the detection of eating disorders in social media posts. José Alberto jbena@unileon.es .
[12] Early Detection of Mental Health Issues Using Social Media Posts. Qasim Bin Saeed binsaeeq@myumanitoba.ca .
[13] Deep Learning-Based Early Depression Detection Using Social Media. Tejas Vaidya. [14]Detection of Depression in Social Media Posts using Emotional Intensity Analysis. M. Kiran Mayee kinu92@gmail.com .
[14] Depression detection from Social Media Bangla Text Using Recurrent Neural Networks. Sultan Ahmed IL66977@umbc.edu .
[15] NarrationDep: Narratives on Social Media For Automatic Depression Detection. Hamad Zogan, hamad.zogan@gmail.com , Imran Razzak, imran.razzak@unsw.edu.au .
[16] Depression detection in social media posts using affective and social norm features. Ilias Triantafyllopoulos hliastriant1@gmail.com ,Georgios Paraskevopoulos geopar,potam@central.ntua.gr .
[17] Depression Detection on Twitter Social Media Platform using Bidirectional Long-Short Term Memory. Andre Agasi Simanungkalit, Warih Maharani, & Prati Hutari Gani.
[18] Predicting Depression Levels Using Social Media Posts: A Comprehensive Survey. Shreya Shitole shreyashitole282@gmail.com ,Anushka Chillal anushkachillal23@gmail.com .
[19] Depression detection in social media comments data using machine learning algorithms. Zannatun Nayem Vasha zannatun15-12939@diu.edu.bd .
[20] Depression Detection from Social Media Posts Using Multinomial Naive Theorem. Rajeev manit Rajeevmanit12276@gmail.com .
[21] Emotion-Based Reinforcement Attention Network for Depression Detection on Social Media:Algorithm Development and Validation. Jian Wang, wangjian@dlut.edu.cn .

[22] Detection of Depression Severity Using Bengali Social Media Posts on Mental Health:Study Using Natural Language Processing Techniques. Muhammad Khubayeeb Kabir muhammad.khubayeeb.kabir@g.bracu.ac.bd .
[23] Novel Transformer Based Contextualized Embedding and Probabilistic Features for Depression Detection from Social Media. Kashif Munir kashif.munir@kfueit.edu.pk and Nagwan Abdel Samee nmabdelsamee@pnu.edu.sa .
[24] Depression Detection Technology Based on Multimodal Social Media Data. Hongyi Pu znjcd@ldy.edu.rs .



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY