2025 IEEE 7th International Conference on Sustainable Technologies for Industry 5.0 (STI 2025) 14-15 December, Dhaka

# Multiclass Mental Disorder Detection from Social Media Using BERT and Hybrid Data Balancing

Zinnia Sultana*, Kaniz Fatema Tanni* , Anirban Datta *,
Punam Nandy*, and Koushik Jamilur Rahman*
*Dept of Computer Science and Engineering, International Islamic University Chittagong, Bangladesh
Email: zinniaiiuc@yahoo.com*, kaniztanni22@gmail.com*, anirbandattadurjoy@gmail.com*,
punamnandy1273@gmail.com*, koushikjamil001@gmail.com*

***Abstract*—Mental illnesses are complex disorders that significantly affect individuals worldwide. In the digital era, social media has become an important medium for individuals to express their emotions and seek support. Machine learning (ML) and deep learning (DL) approaches have been used to identify mental health problems from online posts, yielding promising results. However, most research focuses on binary classification and often struggles with imbalanced datasets. This study aims to develop a BERT-based multiclass classification model to identify several mental disorders, including Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder. To address data imbalance, a hybrid data balancing technique combining Random Oversampling (ROS) and Random Undersampling (RUS) is applied. The model is evaluated on a Mental Health Text Dataset collected from Kaggle and compared with traditional ML algorithms, including Logistic Regression, Random Forest, XGBoost, and Naive Bayes, as well as an LSTM-based DL model. Experimental results show that the baseline BERT model achieved an accuracy of 83%, which improved to 90% after applying the hybrid balancing technique. The proposed approach demonstrates superior performance compared to traditional ML algorithms and LSTM, highlighting the effectiveness of integrating transformer-based models with data balancing techniques for more accurate and robust multiclass mental illness detection from social media data.**

***Index Terms*—Mental Health Detection, Multiclass Classification, BERT, Hybrid Data Balancing, Social Media Analysis**

## I. Introduction

Mental health disorders are prevalent worldwide and pose significant challenges to individuals and healthcare systems [1], [2]. Traditionally, diagnosis has relied on self-reporting and standardized clinical questionnaires. With the growth of social media, users increasingly share their personal experiences and emotions online, creating rich linguistic data that can be analyzed for early identification of mental health conditions [3], [4]. Detecting these patterns early enables timely interventions, potentially reducing severity and long-term complications.

Advances in Natural Language Processing (NLP) and ML have enabled the automated analysis of social media text to assess mental well-being. Many studies have explored methods such as support vector machines (SVM), naive Bayes (NB), logistic regression (LR), and random forest (RF) to detect mental health conditions from online posts [5]–[9]. More recently, DL techniques, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have been applied to capture semantic and contextual features in user-generated text [10]–[12]. The emergence of transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT), has shifted NLP toward architectures that can understand deep contextual relationships in text [13], [14]. BERT leverages pre-trained language representations fine-tuned for specific tasks, leading to superior performance in many text classification problems, including mental health detection. Despite its potential, this approach faces the following challenges:

1) Most studies focus on binary classification, but few have explored detecting multiple disorders.
2) Social media data often have uneven numbers of posts for each mental health category, which can bias predictions.

To address these limitations, this study proposes a BERT-based multiclass classification method to identify a range of mental disorders, including Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder. To handle data set imbalance, a hybrid data balancing technique is applied that combines Random Oversampling (ROS) and Random Undersampling (RUS). Furthermore, the performance of the BERT model is compared with traditional ML models, including LR, RF, XGBoost, NB, and LSTM, to evaluate the advantages of transformer-based approaches over conventional methods.

The main contributions of this study are:

1) A multiclass classification framework is developed using BERT to identify a wide category of mental disorders.
2) The effectiveness of the BERT-based model is validated by comparing its performance against traditional ML models.
3) The study investigates the impact of data balancing techniques on improving BERT's performance for multiclass classification with an unbalanced dataset.

The rest of this paper is structured as follows: Section II presents a review of related literature. Section III explains the methodology and materials in detail. Section IV discusses the results and key findings. Finally, Section V provides the conclusion.

## II. Related works

In recent years, ML and DL techniques have been widely used to detect mental health conditions. This section reviews prior studies based on the target task and methodology.

In [6], posts and comments from Facebook and YouTube were analyzed using six classifiers, with SVM achieving the best performance in distinguishing depressed and non-depressed users. Building on this approach, [15] developed five ML models, where SVM again performed best, achieving F1 scores of 96.6% for Arabic and over 87% for English binary classification. Other studies explored detection of suicidal ideation and stress. For instance, [8] found KNN most effective in predicting suicidal ideation among Bangladeshi university students, identifying 19.9% at risk and enabling early intervention. Logistic regression with Bag of Words features was applied by [16] to detect stress in Reddit academic posts, achieving 77.78% accuracy and highlighting higher stress levels among professors. DL approaches, particularly LSTM networks, have been effective in capturing long-term dependencies in text for depression detection [17], [18]. CNNs also showed strong cross-platform generalization, classifying mental illnesses across Twitter and Reddit data [19]. A hybrid CNN-BiLSTM model with attention (CBA) further improved depression detection on the CLEF2017 dataset, achieving an AUC-ROC of 0.85 and MCC of 0.77 [20]. Transformer-based techniques, including BERT and ELMo embeddings, have recently demonstrated state-of-the-art performance in detecting depression, suicidal ideation, and stress on social media datasets [14], [21], [22]. For example, SENSDeep, a stacking ensemble of six pre-trained transformers, achieved over 97% accuracy and F1-score in depression detection [14]. Similarly, TextCNN combined with linguistic and semantic features improved anxiety detection on Chinese social media [23], and LSTM models enhanced with transformer embeddings identified various mental illnesses from Reddit data [12]. BERT embeddings, combined with ensemble methods, have shown high accuracy for depression detection and early symptom identification [24]. Stacking ensembles using SBERT embeddings and sentiment features further improved early-stage depression detection, achieving F1 scores of 69% and 76% on different datasets [13].

Most existing ML and DL methods for mental health detection focus on binary classification, limiting their ability to handle multiclass problems, especially with imbalanced datasets. This reduces the accuracy for less frequent conditions. To address this gap, this study applies a hybrid data balancing technique alongside BERT to better capture patterns across multiple mental disorders, aiming to improve early and accurate identification and support timely mental health intervention.

## III. Materials and Methods

This section describes the sequential steps in this research. It begins with data collection and preprocessing. Traditional ML models are trained alongside a fine-tuned BERT model. The trained models are evaluated using standard classification metrics to compare their performance. Figure 1 shows the workflow diagram of the proposed work.

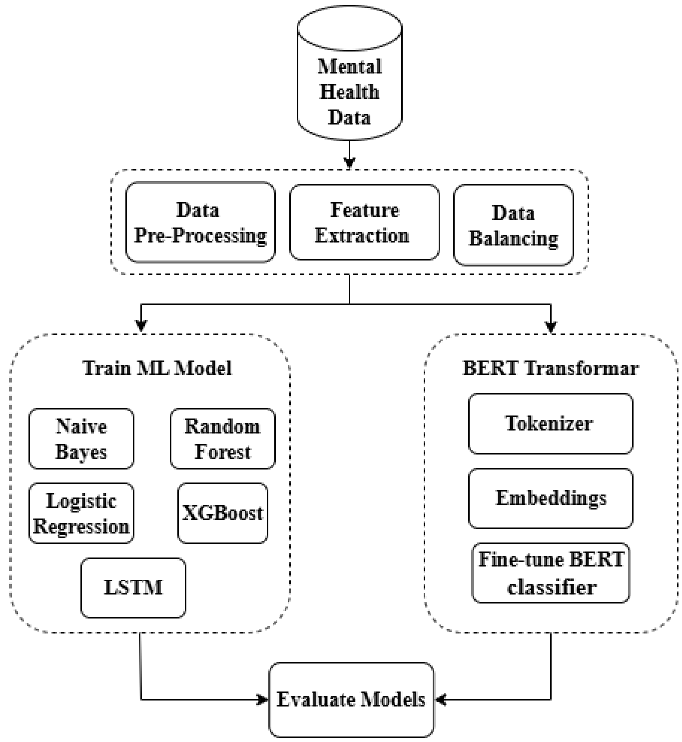


Fig. 1: Workflow Diagram

### A. Dataset Overview

This study uses a mental health text dataset collected from Kaggle [25], containing 53,043 entries labeled with seven mental health statuses: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder. After removing 362 entries with missing text, the final dataset includes 52,681 records. Each entry consists of an ID, a statement, and its label. Table I presents the number of samples and their corresponding percentages for each class. The dataset is highly imbalanced, with Normal, Depression, and Suicidal comprising over 80% of the data, while classes like Personality Disorder and Stress are rare. This imbalance shows the need for proper balancing techniques to improve model performance.

TABLE I: Status Distribution of Dataset

| **Class** | **No. of Records** | **Percentage(%)** |
|---|---|---|
| Normal | 16,343 | 31.0% |
| Depression | 15,404 | 29.2% |
| Suicidal | 10,652 | 20.2% |
| Anxiety | 3,841 | 7.3% |
| Bipolar | 2,777 | 5.3% |
| Stress | 2,587 | 4.9% |
| Personality disorder | 1,077 | 2.0% |
| **Total** | **52,681** | **100%** |

TABLE II: Transformation of Statements After Text Preprocessing

| Original Statement | Lowercased Text | Punctuation Removed | Tokenized Text | Stemmed Text |
|---|---|---|---|---|
| Oh my gosh | oh my gosh | oh my gosh | [oh, my, gosh] | oh my gosh |
| All wrong, back off dear, forward doubt. Stay ... | all wrong, back off dear, forward doubt. stay ... | all wrong back off dear forward doubt stay in ... | [all, wrong, back, off, dear, forward, doubt, stay, in..] | all wrong back off dear forward doubt stay in.. |
| I'm restless and restless, it's been a month n... | i'm restless and restless, it's been a month n... | im restless and restless its been a month n.. | [im, restless, and, restless, its, been, a, month, n..] | im restless and restless it been a month n.. |
| trouble sleeping, confused mind, restless hear... | trouble sleeping, confused mind, restless hear... | trouble sleeping confused mind restless hear ... | [trouble, sleeping, confused, mind, restless, hear..] | trouble sleep confus mind restless hear.. |

## B. Data Pre-Processing

Text preprocessing plays a vital role in transforming raw text into a form suitable for ML models, ensuring consistency and reducing noise in the data. In this study, various preprocessing methods were applied to the dataset. Table II presents sample statements demonstrating the effects of these preprocessing steps.

*1) **Lowercasing**:* This step converts all text data to lowercase, eliminating differences caused by letter casing and ensuring uniformity. For example, the phrase ”Oh my gosh” is transformed into ”oh my gosh” as shown in Table II.

*2) **Removal of Punctuation**:* Once all text was lowercased, punctuation marks and special characters were removed. This step cleans the text by deleting symbols that do not carry meaningful information as illustrated in Table II.

*3) **Tokenization**:* The cleaned text was then split into individual words or tokens. This allows the model to treat each token as a separate feature for analysis. For example, ”oh my gosh” becomes [oh, my, gosh].

*4) **Stemming**:* Finally, tokens were reduced to their root forms using the Porter Stemmer. This step merges different forms of a word to reduce vocabulary size. Like ”sleeping” becomes ”sleep,” and ”confused” becomes ”confus.” as seen in Table II.

## C. Feature Extraction

To prepare the textual data for ML models, it is essential to convert it into a numerical representation. TF-IDF vectorization was applied by initializing the TfidfVectorizer with an n-gram range of (1, 2), enabling the capture of both unigrams and bigrams. The number of features was limited to 50,000 to focus on the most informative tokens. The vectorizer was fitted on the training set and subsequently used to transform both the training and test datasets. Additionally, two numerical features total number of characters and the total number of sentences per sample, were extracted separately. These features were concatenated with the TF-IDF vectors using horizontal stacking, resulting in a combined feature vector of dimension 50,002 for each instance.

## D. Data Balancing

Table I highlights the significant class imbalance in the dataset. Inspired by [26], a hybrid sampling approach combining Random Oversampling (ROS) for minority classes and Random Undersampling (RUS) for majority classes was applied. This balances all classes without excessive duplication of minority samples. Other techniques were also evaluated and compared in the results section. After balancing, each class contains 10,000 samples, and the training and testing distribution is shown in Table III.

TABLE III: Class Distribution After Hybrid Balancing

| Class | Balanced Instances | Training Instances | Testing Instances |
|---|---|---|---|
| Normal | 10,000 | 7,690 | 1,923 |
| Depression | 10,000 | 7,702 | 1,917 |
| Suicidal | 10,000 | 7,700 | 1,920 |
| Anxiety | 10,000 | 7,688 | 1,925 |
| Stress | 10,000 | 7,695 | 1,920 |
| Bipolar Disorder | 10,000 | 7,703 | 1,917 |
| Personality Disorder | 10,000 | 7,702 | 1,918 |
| **Total** | **70,000** | **53,880** | **13,440** |

## E. Proposed BERT-Based Architecture

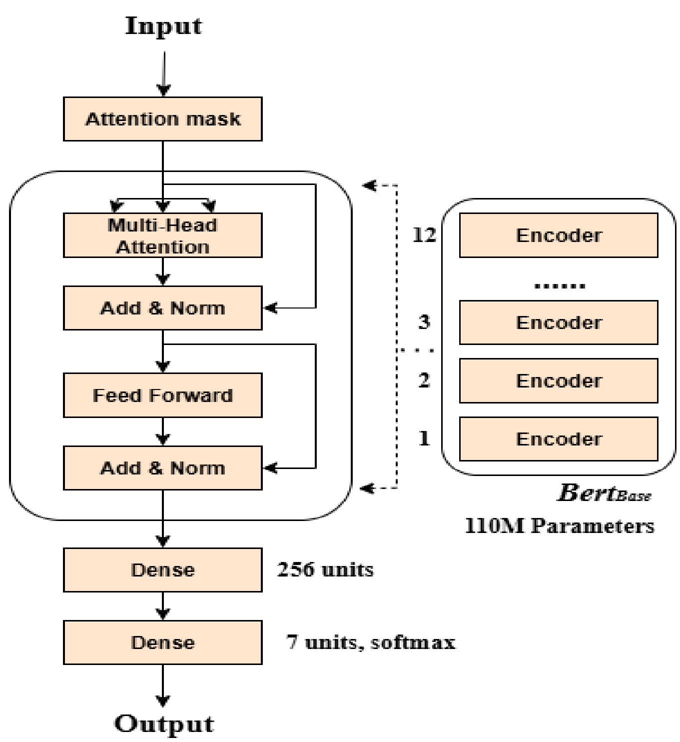


Fig. 2: Proposed Architecture of BERT

In this study, the pre-trained BERT-base-uncased model is employed as a feature extractor for multiclass mental disorder classification. Figure 2 illustrates the proposed BERT-based classification architecture. The input sentences are first tokenized using BERT's pre-trained tokenizer, which converts raw text into token IDs and attention masks. These inputs are fed into the BERT encoder, consisting of 12 transformer layers that generate deep, context-aware embeddings. The pooled output corresponding to the token serves as a summary representation of the entire input sequence. To enhance task-specific learning, the pooled output is passed through an additional fully connected dense layer with 256 neurons and ReLU activation. A dropout layer with a rate of 0.3 is applied next to mitigate overfitting. The final classification layer is a fully connected dense layer with 7 output units, corresponding to the seven mental disorder categories. A softmax activation function converts the output logits into class probabilities. The model is fine-tuned end-to-end using the Adam optimizer with a learning rate scheduler. Cross-entropy loss is employed to guide the optimization process over 10 epochs, ensuring effective learning and improved classification performance.

## IV. Results and Discussion

The experiment was carried out in three aspects, which are explained in detail below:

a) EX. A: Comparative analysis of BERT and traditional ML models.
b) Exp. B: Impact of data balancing on BERT's classification performance
c) Exp. C: Analysis of class-wise performance and error patterns.

*1) Results for Exp. A:* Table IV presents the performance comparison of various classification models on the multiclass mental disorder dataset. The metrics reported include accuracy, precision, recall, and F1-score.

TABLE IV: Performance comparison of different models

| Method | Accuracy (%) | Precision | Recall | F1-score |
|---|---|---|---|---|
| Naive Bayes | 65.00% | 0.76 | 0.55 | 0.55 |
| XGBoost | 70.82% | 0.74 | 0.65 | 0.67 |
| Logistic Regression | 75.34% | 0.77 | 0.70 | 0.72 |
| Random Forest | 73.93% | 0.79 | 0.69 | 0.71 |
| LSTM | 76.01% | 0.73 | 0.71 | 0.72 |
| **BERT** | **83.00%** | **0.81** | **0.82** | **0.81** |

BERT achieved the highest accuracy (83%), outperforming LSTM (76.01%), LR (75.34%), and RF (73.93%). Its precision, recall, and F1-score were also the highest, showing that BERT effectively captures complex contextual features in social media text for mental disorder classification. Traditional models like NB and XGBoost performed worse, with accuracies of 65.00% and 70.82%, respectively. Among traditional models, LR and LSTM performed comparably well, with balanced metrics around 0.7. These results demonstrate the effectiveness of BERT in capturing complex contextual features from social media text for mental disorder classification.

*2) Results for Exp. B:* In Experiment B, different data balancing methods were applied to the BERT model to analyze its performance. As shown in Table V, the hybrid random oversampling and undersampling approach achieved the highest accuracy of 90%, with precision, recall, and F1-score all around 0.90, demonstrating a balanced performance across metrics. In contrast, the hybrid method combining SMOTE with random undersampling resulted in lower performance, with accuracy and F1-score near 0.80. Pure undersampling techniques exhibited even lower effectiveness. These results indicate that the hybrid random oversampling and undersampling strategy is more effective in addressing class imbalance and enhancing classification performance in multiclass mental illness detection. Figure 3 shows BERT's accuracy improved by about 7% on the balanced dataset, while LR, RF, and LSTM showed increases of 5%, 4%, and 3%, respectively.

TABLE V: Performance Comparison of BERT Model with and without Data Balancing

| Data Balancing Technique | Samples per Class | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Without data balancing | (3269, 3081, 2131, 768, 517, 556, 215) | 83% | 0.81 | 0.82 | 0.81 |
| Hybrid random (over & under sampling) | 10,000 | **90%** | **0.91** | **0.90** | **0.90** |
| Hybrid (smote + random undersampling) | 8,000 | 81% | 0.81 | 0.79 | 0.80 |
| Undersampling | 750 | 78% | 0.79 | 0.77 | 0.77 |

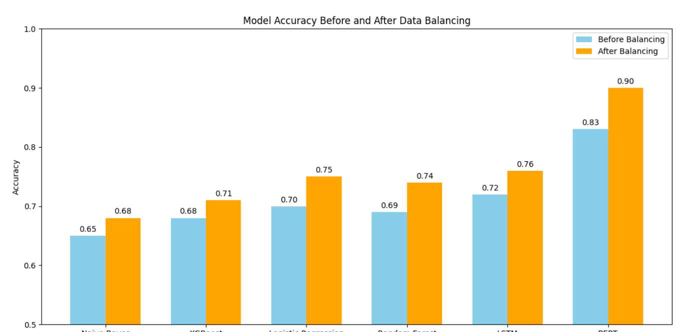


Fig. 3: Comparison of Model Performance on Balanced vs. Unbalanced Data

Figures 4 and 5 show the training and validation accuracy and loss of a BERT model on unbalanced and balanced datasets. The model trained on the balanced dataset (Fig. 5) shows superior performance, with both training and validation accuracy increasing steadily and validation loss consistently decreasing. In contrast, the model trained on the unbalanced dataset (Fig. 4) overfits, as evidenced by a large gap between

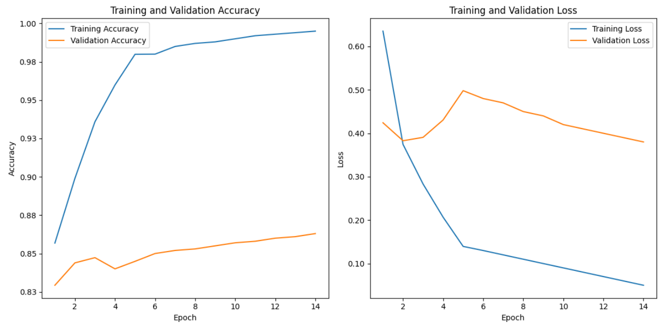


Fig. 4: Training and validation accuracy (left) and loss (right) curves of the BERT (Imbalanced Dataset)

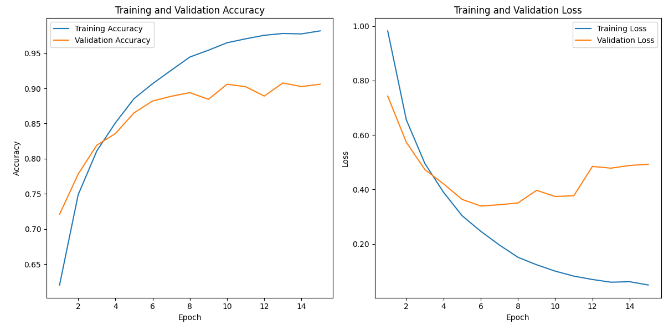


Fig. 5: Training and validation accuracy (left) and loss (right) curves of the BERT(Balanced Dataset)

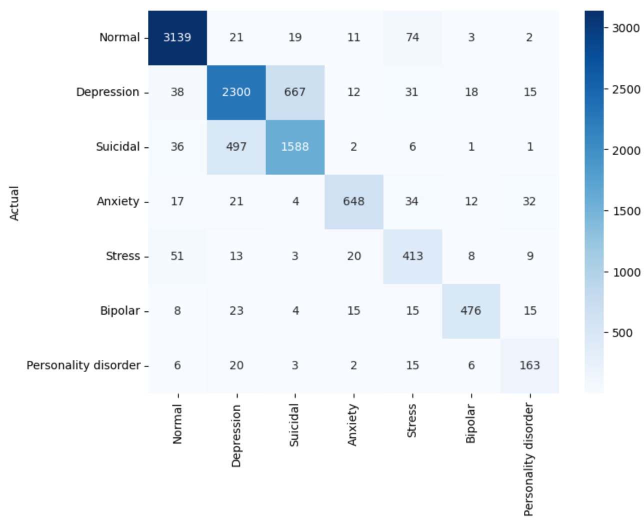


Fig. 6: Confusion matrix of BERT model without data balancing.


Fig. 7: Confusion matrix of BERT model with data balancing.

training and validation accuracy and an increasing validation loss, indicating poor generalization. This highlights the importance of addressing data imbalance for effective machine learning.

*3) Results for Exp. C:* Figures 6 and 7 present the confusion matrices of the BERT model without and with data balancing, respectively. Without data balancing, the model shows strong performance on the majority "Normal" class, as indicated by the large diagonal values for this class, while smaller classes such as Stress and Bipolar remain underrepresented. Rare classes like Personality disorder and Suicidal also exhibit low accuracy, indicating a clear bias toward the majority classes. Misclassifications frequently occur toward larger classes, with many Stress samples incorrectly labeled as Normal. In contrast, with data balancing, predictions are more evenly distributed, with substantial increases in correct classifications for Stress, Bipolar, and Personality disorder. Although Normal class accuracy decreases, this reflects reduced bias and better recognition of minority classes Furthermore, Table VI provides a detailed class-wise comparison of precision, recall, and F1-score for BERT under balanced and unbalanced scenarios. The analysis shows that data balancing generally improves precision and F1-scores for minority classes, enhancing the model's ability to correctly identify less frequent mental health conditions. On the other hand, for some majority classes such

TABLE VI: Class-wise Precision, Recall, and F1-Score of BERT with and without Balancing

| **Class** | **Without Balancing** | | | **With Balancing** | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| Normal | 0.95 | 0.96 | 0.96 | 0.94 | 0.95 | 0.95 |
| Depression | 0.79 | 0.75 | 0.77 | 0.89 | 0.90 | 0.89 |
| Suicidal | 0.69 | 0.75 | 0.72 | 0.85 | 0.88 | 0.86 |
| Anxiety | 0.91 | 0.84 | 0.88 | 0.90 | 0.88 | 0.89 |
| Stress | 0.70 | 0.80 | 0.75 | 0.84 | 0.85 | 0.84 |
| Bipolar | 0.91 | 0.86 | 0.88 | 0.88 | 0.87 | 0.87 |
| Personality Disorder | 0.69 | 0.76 | 0.72 | 0.83 | 0.82 | 0.82 |

as "Normal" there is a slight drop in precision (0.95 to 0.94), recall (0.96 to 0.95), and F1 score (0.96 to 0.95). Despite this, the overall performance gains for minority classes contribute to a more balanced and reliable classification system.

## V. Conclusion

This study introduced a BERT-based multiclass classification approach for detecting multiple mental disorders from social media text. To tackle the issue of class imbalance, a hybrid balancing approach was adopted, integrating Random Oversampling (ROS) with Random Undersampling (RUS). Experimental results demonstrated that balancing the dataset substantially improved model performance, increasing accuracy from 83% to 90%, while also reducing misclassification among minority classes. Comparative evaluations further highlighted the superior performance of the proposed method over traditional machine learning models, including Logistic Regression, Random Forest, XGBoost, and Naive Bayes, as well as LSTM-based deep learning networks. These findings underscore the effectiveness of integrating transformer-based architectures with robust data balancing strategies to achieve accurate and reliable multiclass mental health detection. From the Industry 5.0 perspective, this work supports the development of human-centric and sustainable AI solutions that enhance mental well-being and contribute to sustainable digital healthcare. Future work will explore larger and more diverse social-media datasets, advanced balancing and augmentation techniques, and multimodal data to further enhance model generalization. It will also include comparisons with other transformer models such as RoBERTa, DistilBERT, and XLNet to validate the robustness and adaptability of the proposed approach.

## References


[1] W. H. Organization, *World mental health report: Transforming mental health for all*. World Health Organization, 2022.

[2] M. L. Wainberg, P. Scorza, J. M. Shultz, L. Helpman, J. J. Mootz, K. A. Johnson, Y. Neria, J.-M. E. Bradford, M. A. Oquendo, and M. R. Arbuckle, "Challenges and opportunities in global mental health: a research-to-practice perspective," *Current psychiatry reports*, vol. 19, no. 5, p. 28, 2017.

[3] Y. Cai, H. Wang, H. Ye, Y. Jin, and W. Gao, "Depression detection on online social network with multivariate time series feature of user depressive symptoms," *Expert Systems with Applications*, vol. 217, p. 119538, 2023.

[4] A.-S. Uban, B. Chulvi, and P. Rosso, "An emotion and cognitive based analysis of mental health disorders from social media data," *Future Generation Computer Systems*, vol. 124, pp. 480–494, 2021.

[5] A. Ahmed, S. Aziz, C. T. Toro, M. Alzubaidi, S. Irshaidat, H. A. Serhan, A. A. Abd-Alrazaq, and M. Househ, "Machine learning models to detect anxiety and depression through social media: A scoping review," *Computer methods and programs in biomedicine update*, vol. 2, p. 100066, 2022.

[6] Z. N. Vasha, B. Sharma, I. J. Esha, J. Al Nahian, and J. A. Polin, "Depression detection in social media comments data using machine learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 987–996, 2023.

[7] M. M. Islam, S. Hassan, S. Akter, F. A. Jibon, and M. Sahidullah, "A comprehensive review of predictive analytics models for mental illness using machine learning algorithms," *Healthcare Analytics*, vol. 6, p. 100350, 2024.

[8] S. S. Sara, M. A. Rahman, R. Rahman, and A. Talukder, "Prediction of suicidal ideation with associated risk factors among university students in the southern part of bangladesh: Machine learning approach," *Journal of affective disorders*, vol. 349, pp. 502–508, 2024.

[9] S. Tutun, M. E. Johnson, A. Ahmed, A. Albizri, S. Irgil, I. Yesilkaya, E. N. Ucar, T. Sengun, and A. Harfouche, "An ai-based decision support system for predicting mental health disorders," *Information Systems Frontiers*, vol. 25, no. 3, pp. 1261–1276, 2023.

[10] J. P. Thekkekara, S. Yongchareon, and V. Liesaputra, "An attention-based cnn-bilstm model for depression detection on social media text," *Expert systems with applications*, vol. 249, p. 123834, 2024.

[11] A. Amanat, M. Rizwan, A. R. Javed, M. Abdelhaq, R. Alsaqour, S. Pandya, and M. Uddin, "Deep learning for depression detection from textual data," *Electronics*, vol. 11, no. 5, p. 676, 2022.

[12] R. Dash, S. Udgata, R. K. Mohapatra, V. Dash, and A. Das, "A deep learning approach to unveil types of mental illness by analyzing social media posts," *Mathematical and Computational Applications*, vol. 30, no. 3, p. 49, 2025.

[13] B. Ogunleye, H. Sharma, and O. Shobayo, "Sentiment informed sentence bert-ensemble algorithm for depression detection," *Big Data and Cognitive Computing*, vol. 8, no. 9, p. 112, 2024.

[14] K. Noor, M. Rehman, M. Anjum, A. Hussain, and R. Saleem, "Sentiment analysis for depression detection: A stacking ensemble-based deep learning approach," *International Journal of Information Management Data Insights*, vol. 5, no. 2, p. 100358, 2025.

[15] A. Helmy, R. Nassar, and N. Ramdan, "Depression detection for twitter users using sentiment analysis in english and arabic tweets," *Artificial intelligence in medicine*, vol. 147, p. 102716, 2024.

[16] N. Oryngozha, P. Shamoi, and A. Igali, "Detection and analysis of stress-related posts in redditâs acamedic communities," *IEEE Access*, vol. 12, pp. 14 932–14 948, 2024.

[17] M. Z. Uddin, K. K. Dysthe, A. Følstad, and P. B. Brandtzaeg, "Deep learning for prediction of depressive symptoms in a large textual dataset," *Neural Computing and Applications*, vol. 34, no. 1, pp. 721–744, 2022.

[18] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich cnn and bi-directional lstm," *Multimedia Tools and Applications*, vol. 81, no. 17, pp. 23 649–23 685, 2022.

[19] C. S. Ang and R. Venkatachala, "Generalizability of machine learning to categorize various mental illness using social media activity patterns," *Societies*, vol. 13, no. 5, p. 117, 2023.

[20] J. P. Thekkekara, S. Yongchareon, and V. Liesaputra, "An attention-based cnn-bilstm model for depression detection on social media text," *Expert systems with applications*, vol. 249, p. 123834, 2024.

[21] A. Karamat, M. Imran, M. U. Yaseen, R. Bukhsh, S. Aslam, and N. Ashraf, "A hybrid transformer architecture for multiclass mental illness prediction using social media text," *IEEE Access*, 2024.

[22] S. Inamdar, R. Chapekar, S. Gite, and B. Pradhan, "Machine learning driven mental stress detection on reddit posts using natural language processing," *Human-Centric Intelligent Systems*, vol. 3, no. 2, pp. 80–91, 2023.

[23] J. Zhu, Z. Zhang, Z. Guo, and Z. Li, "Sentiment classification of anxiety-related texts in social media via fuzing linguistic and semantic features," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 5, pp. 6819–6829, 2024.

[24] M. A. Abbas, K. Munir, A. Raza, N. A. Samee, M. M. Jamjoom, and Z. Ullah, "Novel transformer based contextualized embedding and probabilistic features for depression detection from social media," *IEEE Access*, vol. 12, pp. 54 087–54 100, 2024.

[25] S. Sarkar, "Sentiment analysis for mental health," Kaggle dataset, 2025, accessed: 2025-08-13. [Online]. Available: https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health

[26] Q. Wang, "A hybrid sampling svm approach to imbalanced data classification," in *Abstract and applied analysis*, vol. 2014, no. 1. Wiley Online Library, 2014, p. 972786.