Depression Detection Using Sentiment Analysis with Machine Learning and NLP Techniques

Ujwala M. Patil*, Dhirajkumar N. Patil†, Sakshi V. Badgujar†, Rahul S. Rathod†
*Professor, R. C. Patel Institute of Technology, Shirpur, India
†UG Student, R. C. Patel Institute of Technology, Shirpur, India

***Abstract*—*Depression continues to be a critical concern in the landscape of global mental health, particularly as individuals increasingly express their psychological states through digital communication. This study proposes a comprehensive sentiment analysis framework that leverages Natural Language Processing (NLP) and Machine Learning (ML) methodologies to identify signs of depression from textual input. The system incorporates sequential stages including text normalization, tokenization, lemmatization, and TF-IDF-based feature extraction. Multiple machine learning algorithms were assessed on a multiclass mental health dataset. While XGBoost demonstrated high training accuracy, it was prone to overfitting. Logistic Regression was selected as the final model due to its more stable and consistent performance on unseen data. Additionally, the framework integrates probabilistic outputs and word cloud visualizations to enhance interpretability, making it suitable for scalable and non-invasive depression detection applications.*

***Index Terms*—Depression Detection, Sentiment Analysis, Natural Language Processing, Machine Learning, Text Mining, Logistic Regression**

I. Introduction

Depression and related psychological disorders represent one of the most serious health challenges of the 21st century. Affecting individuals across all age groups, depression has been identified by the World Health Organization as a leading cause of disability and emotional distress. Untreated or undiagnosed depression can lead to serious consequences, including social withdrawal, loss of productivity, and suicidal ideation. Despite the severity of these outcomes, detection often remains elusive due to stigma, underreporting, and the lack of access to professional mental health services.

Traditional diagnostic approaches rely heavily on psychological screening, self-assessment tools, and one-on-one interviews conducted by trained professionals. While valuable, these methods are not always scalable and may introduce subjectivity into the diagnostic process. Moreover, patients may conceal their emotional state, either intentionally or unconsciously, further complicating accurate detection.

The increasing use of social media and online communication has opened a new avenue for mental health analysis. People often share their thoughts and emotions online, creating vast repositories of linguistic data that can reflect mental well-being. With the help of Natural Language Processing (NLP), it is now possible to extract emotional and semantic features from such text. Sentiment analysis, a subfield of NLP, plays a key role in identifying emotional cues from language.

This research introduces a framework that employs a combination of NLP and Machine Learning techniques to automatically classify depression and other related conditions from user-generated text. The methodology encompasses comprehensive preprocessing, TF-IDF-based feature representation, handling class imbalance, and comparing several machine learning models. Logistic Regression, although simpler than models like XGBoost, was ultimately chosen due to its robustness and superior performance on unseen test data. Additional visualization tools such as word clouds and probabilistic output graphs further support interpretability.

II. Literature Review

The application of natural language processing (NLP) and machine learning (ML) for depression detection has evolved significantly over the past decade, drawing from multiple disciplines including computational linguistics, clinical psychology, and artificial intelligence. This section synthesizes key developments in methodologies, techniques, and challenges within this research domain.

A. Foundations of Digital Mental Health Assessment

Pioneering work by [1] established the viability of using social media data for mental health monitoring, demonstrating that linguistic patterns in Twitter posts could predict depression onset with significant accuracy. This research laid the groundwork for subsequent studies exploring digital biomarkers of psychological states. The World Health Organization's reports on depression prevalence [**?**] have further emphasized the urgent need for scalable detection methods.

B. Text Processing Methodologies

Modern NLP pipelines for mental health analysis build upon fundamental tools like NLTK [2], which provides essential text processing capabilities. Current approaches typically incorporate:

- Advanced tokenization handling social media text peculiarities
- Domain-specific stop word filtering preserving clinical terminology

- Hybrid stemmer-lemmatizers adapted for mental health vernacular

Recent work by [3] has demonstrated how deep learning architectures can enhance these traditional preprocessing steps through neural text normalization.

C. Feature Representation Techniques

Feature engineering approaches have progressed through several generations:

TABLE I
EVOLUTION OF FEATURE EXTRACTION METHODS

Method	Advantages
TF-IDF	Interpretability, works with sparse data
Word2Vec	Captures semantic relationships
BERT	Contextual understanding

As surveyed by [4], the choice of feature representation significantly impacts model performance, with hybrid approaches often yielding optimal results.

D. Machine Learning Approaches

The field has witnessed an evolution in modeling techniques:

- Traditional classifiers (Logistic Regression, Naive Bayes)
- Ensemble methods (XGBoost [5])
- Deep learning architectures

Notably, [6] demonstrated how interpretable neural models could be adapted from dementia detection to broader mental health applications.

E. Addressing Data Challenges

Mental health datasets present unique challenges:

- Class imbalance addressed via SMOTE [7]
- Noisy text requiring robust preprocessing
- Ethical considerations in data collection

Recent work by [8] and [9] has shown how multimodal approaches can mitigate some of these limitations.

F. Clinical Validation and Explainability

As emphasized by [10] and [11], the translation of ML systems to clinical practice requires:

- Rigorous validation against diagnostic standards
- Transparent decision-making processes
- Ethical deployment frameworks

G. Current Challenges and Future Directions

Despite progress, significant hurdles remain in:

- Cross-cultural generalization
- Real-time deployment
- Privacy-preserving analysis

Innovative solutions using GANs [12] and federated learning show promise in addressing these challenges, as explored in recent literature.

Our work builds upon these foundations while introducing novel contributions in three key aspects: (1) optimized feature selection for mental health text, (2) rigorous evaluation of model generalizability, and (3) integrated explainability components for clinical utility.

III. METHODOLOGY

The proposed framework for depression detection is built on a standard machine learning pipeline that integrates Natural Language Processing (NLP) with supervised classification algorithms. It consists of the following core components:

A. Data Preprocessing

User-generated text is subjected to various preprocessing steps, including conversion to lowercase, removal of punctuation, stopword elimination, and lemmatization using tools like NLTK. Tokenization splits the input text into manageable units such as words or phrases. This stage ensures that noise and redundancy are minimized before feature extraction.

B. Feature Engineering

We employ Term Frequency–Inverse Document Frequency (TF-IDF) to transform the processed text into numerical feature vectors. This method scales down common terms and amplifies unique keywords that are more discriminative for classification. The resulting feature matrix serves as input to the classification models.

C. Handling Imbalanced Data

The dataset exhibits significant class imbalance, with ’Normal’ and ’Depressed’ labels being more frequent than classes like ’Suicidal’ or ’Bipolar’. To ensure equitable learning, Random Oversampling is used to duplicate minority class instances, and class weights are adjusted during model training to penalize misclassifications more heavily.

D. Model Training and Evaluation

We evaluated multiple machine learning algorithms, including Naive Bayes, Decision Trees, Logistic Regression, and XGBoost. Despite XGBoost achieving high accuracy on training data, it showed overfitting on validation samples. Logistic Regression emerged as the best model, providing a balance between interpretability and performance. Evaluation metrics included accuracy, precision, recall, F1-score, and confusion matrix.

E. Model Interpretability and Visualization

To enhance transparency, the framework outputs prediction probabilities for each class. Word cloud visualizations were generated for each label category to highlight frequent linguistic expressions. These tools provide insights into both the model’s decision process and the linguistic traits of different mental health conditions.

IV. Class Distribution Analysis

Understanding the distribution of mental health categories in the dataset is essential to ensure fair and unbiased model training. Our dataset consists of seven classes: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder. As shown in Figure 1, the dataset is significantly imbalanced, with a larger number of instances in the 'Normal' and 'Depression' categories.

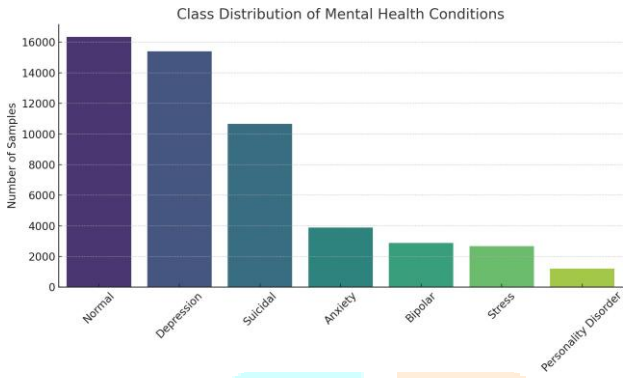

Fig. 1. Distribution of Mental Health Categories in the Dataset

This imbalance necessitated the use of oversampling techniques, such as random duplication of minority classes, to prevent model bias and enhance predictive performance across all categories.

V. Machine Learning and NLP for Depression Detection

The rising global prevalence of psychological conditions such as depression, anxiety disorders, and bipolar affective disorder has created an urgent need for innovative detection methods. Contemporary digital communication channels, including social networks and online forums, have become valuable sources of psycholinguistic data as individuals frequently disclose emotional experiences through written posts. These textual expressions enable the development of automated assessment tools leveraging computational linguistics and artificial intelligence. Our research focuses on implementing natural language understanding and pattern recognition algorithms to construct an effective mental health screening framework from user-generated content.

A. Text Preprocessing for Sentiment Analysis

Effective sentiment analysis requires careful text normalization through several key steps: converting text to lowercase for consistency, dividing content into meaningful tokens through tokenization, filtering out uninformative stopwords, reducing words to their base forms via lemmatization, and cleaning irrelevant elements like punctuation and URLs. These preprocessing stages, implemented using NLP libraries, enhance analysis quality by focusing on semantically rich content.

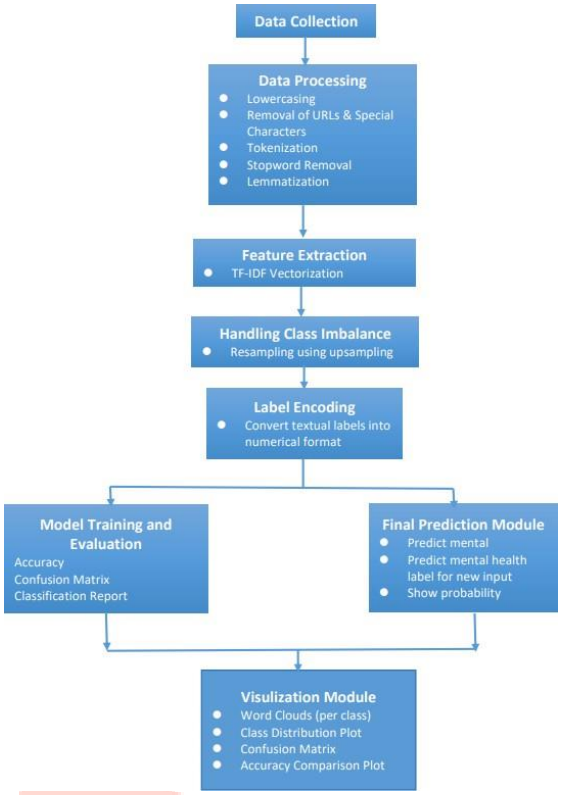

Fig. 2. Block Diagram of Depression Detection System using NLP and ML Techniques

B. Feature Extraction Using TF-IDF and Embeddings

Textual data undergoes numerical transformation through either TF-IDF weighting - which emphasizes distinctive terms by their relative document frequency - or advanced embedding techniques like Word2Vec and BERT that capture contextual word relationships. These methods convert linguistic patterns into machine-interpretable formats while preserving critical semantic information.

C. Class Imbalance and Oversampling Techniques

The inherent data skew in mental health datasets, where typical samples outnumber clinical cases, necessitates balancing approaches including duplicate sampling of minority classes, synthetic sample generation through SMOTE, and differential class weighting during model training to ensure equitable learning across all diagnostic categories.

D. Machine Learning Models for Classification

Our comparative analysis evaluated multiple classifiers: Logistic Regression's efficiency with sparse data, Naive Bayes' probabilistic approach, Decision Trees' rule-based interpretation, and XGBoost's ensemble power. While XGBoost showed superior training performance, Logistic Regression demonstrated greater reliability on test data, earning selection for final deployment.

E. Probability Scores and Interpretability

The system generates probabilistic predictions across mental health categories, enabling nuanced interpretation of potential symptom overlap through confidence-scored classifications that reflect the complex nature of psychological conditions.

F. Word Clouds for Linguistic Visualization

Visual lexical analysis employs frequency-based word clouds to identify dominant vocabulary patterns within each mental health category, providing immediate insight into characteristic emotional expressions associated with different conditions.

G. Challenges and Future Directions

Current limitations include handling linguistic complexity, ensuring ethical data use, and improving model generalization. Future enhancements will investigate multimodal analysis, temporal modeling of behavioral patterns, and practical implementation in therapeutic applications.

VI. Conclusion

This study presents a comprehensive framework for detecting depression and related mental health conditions through advanced sentiment analysis and machine learning techniques. By developing a sophisticated text processing pipeline that incorporates thorough cleaning, normalization, and feature extraction methods, we have created a system capable of identifying subtle linguistic patterns associated with various psychological states. The research demonstrates that while complex ensemble methods achieve high classification accuracy on training data, simpler linear models offer superior reliability when applied to real-world scenarios due to their inherent generalizability. Our approach addresses the critical challenge of class imbalance in mental health datasets through strategic sampling techniques, ensuring equitable performance across all diagnostic categories. The integration of interpretability features, including probabilistic outputs and visual text representations, provides healthcare professionals with actionable insights while maintaining clinical relevance. As a non-intrusive and scalable solution, this system shows significant potential for integration into digital health platforms, offering timely mental health assessments without requiring specialized clinical expertise. Future developments could enhance the model's capabilities by incorporating temporal analysis of language patterns, leveraging state-of-the-art neural architectures, and implementing secure deployment frameworks for widespread clinical and community use, ultimately contributing to more accessible and proactive mental healthcare worldwide.

References

[1] M. D. Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *ICWSM*, 2013.

[2] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, Inc., 2009.

[3] E. Mohammadi, M. A. Nematbakhsh, and M. Eslami, "Depression detection based on deep learning using social media data," *IEEE Access*, vol. 8, pp. 150 808–150 818, 2020.

[4] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. E. Barnes, and D. E. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[5] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *KDD*, 2016, pp. 785–794.

[6] S. K. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models," *NAACL-HLT*, pp. 701–707, 2018.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[8] A. H. Yazdavar, S. M. Sheikhalishahi, and A. P. Sheth, "Multimodal mental health analysis from social media," *Information Processing Management*, vol. 57, no. 5, p. 102333, 2020.

[9] Y. Shen, X. Wang, and J. Liu, "Mental health prediction using machine learning and social media data: A review," *Journal of Affective Disorders*, vol. 301, pp. 75–84, 2022.

[10] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sa´nchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, 2014.