Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DETECTING MENTAL ILLNESS THROUGH SENTIMENT ANALYSIS ON SOCIAL MEDIA USING ENSEMBLE MACHINE LEARNING

S V N SREENIVASU[1], MR.V.S.N. MURTHY [2], E. VIJAYA LAKSHMI [3], J L SARWANI THEEPARTHI[4], G SATYANARAYANA[5], B K MADHAVI[6], G. SIVAKUMAR[7]

[1]Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopet, drsvnsrinivasu@gmail.com, Andhra Pradesh, India
[2]Assistant Professor, Shri Vishnu engineering college for women, Department of IT, Bhimavaram, West Godavari, Andhra Pradesh, vsn.murthy87@gmail.com
[3]Assistant Professor, Department of computer science and engineering Koneru lakshmayya education foundation, Green fields, vaddeswaram, AP -522302, edurivijaya@kluniversity.in
[4]Assistant Professor, Department of CSE.Aditya University, Surampalem, India, tjlsarwa@gmail.com
[5]Associate Professor, Department of CSE, Vasireddy Venkatadri International Technological University, Nambur, Guntur, Andhra Pradesh, India, snarayana.5813@gmail.com
[6] Associate Professor,Information Technology,Vardhaman college of engineering,Shamshabad kousmadhu717@gmail.com
[7] Professor, Department of CSE, Erode Sengunthar Engineering College Perundurai, Erode-638057,sivakumarganapathi@gmail.com *
Coressponding Author: snarayana.5813@gmail.com

ABSTRACT

The Research proposed in this paper concentrate on the discovery and classification of mental disorders are using sentiment analysis techniques over socialmedia data. Platforms like Reddit and Twitter are rich in user-generated content, many of which are related to mental and psychological well-being. The research utilized the CLPsych 2015 Shared Task Dataset, which comprises social media posts tagged with various mental health conditions (depression, PTSD, anxiety, and bipolar disorder). By combining linguistic markers through Natural Language Processing (NLP) with sentiment evaluation components, the study seeks to identify the emotional nuances and behavioral cues linked to these disorders. In addition, the efficacy of various ensemble machine learning models such as Random Forest, Gradient Boosting, and XGBoost were compared to classical classifiers: Logistic Regression, Naive Bayes and Support Vector Machines (SVM) for multi-class classification. Measurement was carried out using an accuracy, precision, recall, and F1-score metrics. The ensemble methods all consistently outperformed individual methods, especially in addressing overlapping symptoms across mental health categories. This study demonstrates the potential of combining sentiment analysis with powerful ensemble learning strategies in early, automated and scalable identification of mental illnesses from social media contents.

Keywords- *Mental Illness Detection; Sentiment Analysis; Social Media; Multi-class Classification; Ensemble Learning; Natural Language Processing (NLP); Reddit; CLPsych 2015 Dataset; Depression; Anxiety; PTSD; Bipolar Disorder; Machine Learning; Text Mining*.

I. INTRODUCTION

With millions of people afflicted by disorders including depression, anxiety, bipolar illness, and post-traumatic stress disorder (PTSD), mental health has become a major global issue. The World Health Organization (WHO) estimates that depression by itself affects about 280 million persons globally [1]. Improvement of treatment results and general well-being depends on early identification and quick intervention. Nonetheless, conventional methods of mental health screening mostly rely on clinical consultations, which can be time-consuming, stigmatized, and inaccessible to significant numbers of people depending on geographical or financial limitations.

People utilize social media sites including Reddit, Twitter, and Facebook more and more to share their ideas, feelings, and life events as they proliferate. Particularly Reddit lets users address personal things, including mental health concerns,

under some anonymity. Rich data sources provided by these user-generated texts help one to grasp psychological and emotional states [2]. Extensively extracting and measuring emotions and attitudes from textual material, sentiment analysis has become a potent instrument for non-invasive and mass assessment of mental health markers [3].

Recent breakthroughs in Natural Language Processing (NLP) have permitted the extraction of linguistic and sentiment-based elements from text, thereby enabling the identification of emotional patterns indicative of mental illness. Initial research concentrated on binary classification (e.g., distinguishing between depressed and non-depressed individuals); however, there is an increasing necessity for more nuanced systems that can perform multi-class classification across diverse mental health disorders, including depression, anxiety, PTSD, and bipolar disorder [4].

Machine learning models have shown encouraging outcomes in categorizing mental health disorders based on variables extracted from social media data [5]. Nonetheless, individual classifiers like Logistic Regression (LR) and Support Vector Machines (SVM) frequently encounter difficulties while processing noisy, high-dimensional data. Ensemble learning methods, including Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB), enhance performance by consolidating predictions from various base learners. These models exhibit notable resilience to overfitting and are proficient in addressing data imbalance and multi-class classification issues [6].

This research employs ensemble learning methods for the multi-class classification of mental health disorders through sentiment and language analysis of Reddit data. The CLPsych 2015 Shared Task dataset comprises Reddit posts annotated for various mental disorders, including depression, anxiety, bipolar disorder, and PTSD. Sentiment analysis is utilized to extract emotional indicators, which are subsequently employed to train diverse machine learning models. Ensemble approaches (RF, GB, XGB) are juxtaposed with standard classifiers (LR, SVM, Naive Bayes) to assess their efficacy in categorizing user mental states.

This study applies ensemble learning techniques for multi-class classification of mental health conditions using sentiment and linguistic analysis of Reddit data. We use the CLPsych 2015 Shared Task dataset, which contains Reddit posts annotated for several mental disorders, including depression, anxiety, bipolar disorder, and PTSD. Sentiment analysis is applied to extract emotional cues, which are then used to train various machine learning models. Ensemble techniques (RF, GB, XGB) are compared against traditional classifiers (LR, SVM, Naive Bayes) to evaluate their effectiveness in classifying user mental states.

The key contributions of this work are as follows:

- We propose a sentiment-based ensemble learning framework for multi-class classification of mental illness.
- We implement a comparative evaluation of ensemble and traditional machine learning models on real-world Reddit data.
- We analyze the linguistic and emotional markers associated with different mental health categories.

The subsequent sections of this work are structured as follows: Section 2 examines pertinent literature. Section 3 delineates the dataset and pre-processing procedures, as well as the technique employed. Section 4 delineates the experimental findings. Section 7 closes the work and proposes avenues for future research.

II. RELATED WORK

The rising incidence of mental health disorders and the escalating utilization of social media platforms have prompted extensive study employing computational methods for the identification of mental disease. Researchers have investigated multiple methodologies that integrate Natural Language Processing (NLP), sentiment analysis, and machine learning to identify psychological diseases, including depression, anxiety, PTSD, and bipolar disorder, from user-generated content. This section examines pertinent literature organized into three primary domains: (i) detection of mental disease via social media, (ii) sentiment and linguistic analysis pertaining to mental health, and (iii) ensemble machine learning methodologies for mental health classification.

2.1 Mental Illness Detection on Social Media

Social media platforms, notably Reddit, Twitter, and Facebook, have become significant data sources for mental health research owing to the forthright articulation of emotions and experiences. De Choudhury et al. [7] initiated foundational research in this domain by identifying postpartum depression through Twitter tweets. They revealed that linguistic indicators, posting patterns, and metadata might collectively signify depression symptoms. Yates et al. [8] utilized Reddit data to identify individuals at risk of depression and self-harm, highlighting the significance of longitudinal data and temporal trends.

Reddit has emerged as a central platform for mental health identification, owing to its community framework and dedicated mental health subreddits (e.g., r/depression, r/anxiety). The CLPsych 2015 Shared Task established a benchmark dataset for forecasting mental disease risk levels derived from Reddit postings [9]. This dataset has been utilized to train multiple classification models to distinguish between control and mentally ill individuals. Cohan et al. [10] utilized the identical dataset to categorize user posts into various mental health problems, illustrating that NLP-based models can surpass fundamental statistical techniques.

The SMHD (Self-reported Mental Health Diagnoses) dataset [11] offered another extensive corpus sourced from Reddit, encompassing people diagnosed with nine mental disorders. This dataset enabled researchers to examine the linguistic and temporal attributes of various diseases, providing insights into the unique patterns linked to conditions including schizophrenia and bipolar disorder.

2.2 Sentiment and Linguistic Analysis for Mental Health

Sentiment analysis has demonstrated efficacy in identifying emotional states, particularly when integrated with language and psycholinguistic attributes. Resnik et al. [12] conducted research utilizing sentiment scoring and Latent Dirichlet Allocation (LDA) to categorize persons diagnosed with depression. Research indicated that individuals with depression demonstrate heightened utilization of negative emotions, self-referential language, and terminology linked to hopelessness and isolation.

Chancellor et al. [13] examined ethical and methodological concerns in identifying mental health disorders using sentiment and emotional tone, emphasizing the necessity for context-sensitive models. Calvo et al. [14] underscored the significance of emotion classification systems in early intervention and therapeutic assistance, noting that emotional expression frequently coincides with clinical signs of depression and anxiety.

Empath and LIWC (Linguistic Inquiry and Word Count) are two tools that have been used to collect linguistic patterns connected to psychological processes, in addition to traditional sentiment assessment. Research by Coppersmith et al. [15] provides support for the idea that sentiment analysis might identify important psychological characteristics, since it was found that those with PTSD used more profanity and terms denoting hostility and avoidance.

By picking up on nuances in context and emotions, new deep learning sentiment models like BERT and RoBERTa have improved classification accuracy even more. The models' enhanced sentiment identification capabilities translate to better results in subsequent mental health classification tests [16].

2.3 Ensemble Learning in Mental Health Classification

Early mental health detection models mostly relied on traditional machine learning classifiers like Logistic Regression (LR), Naive Bayes (NB), and Support Vector Machines (SVM) because of their simplicity and interpretability [17]. When faced with the high-dimensional, sparse, and noisy text data commonly present in social media posts, these models frequently fail to deliver satisfactory results.

Improving classification performance by aggregating the predictions of numerous weak learners has led to the rise in popularity of ensemble learning approaches, such as Random Forests (RF), Gradient Boosting Machines (GBM), and Extreme Gradient Boosting (XGBoost) [18]. The benefits of ensemble approaches in many domains, such as unbalanced and multi-class classification issues, are thoroughly reviewed by Rokach [19]. Using RF and XGBoost, Wongkoblap et al. [20] were able to enhance the recall and precision of their depression prediction using Twitter data compared to SVM and NB. The strength of ensemble models in dealing with text unpredictability and missing features was highlighted in their work. Similarly, Benton et al. [21] demonstrated that stacked models outperformed individual classifiers when it came to diagnosing mental illness from online content. They achieved this by combining numerous classifiers using stacking, a meta-ensemble technique.

Additionally, hybrid ensemble-deep learning models have been investigated in a few research. Combining deep feature extraction with ensemble classification improves performance, as demonstrated by Orabi et al. [22]'s use of a CNN-RF hybrid model to detect melancholy behavior from text. In situations where interpretability, reduced processing costs, and flexibility to limited datasets are paramount, ensemble models tend to outperform deep models, despite the latter's high accuracy.

2.4 Multi-Class Mental Illness Detection

When they first came out, the majority of models could only handle binary data, such as sadness vs. control. Users may display symptoms of numerous diseases, though, because mental health is a continuum. Consequently, research into creating multi-class classifiers that can differentiate between mental health issues such as PTSD, bipolar disorder, anxiety, and depression is gaining momentum.

Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Using hierarchical support vector machines (SVMs) and topic modelling, Sadeque et al. [23] developed a multi-class classifier that could categorize Reddit users according to five different mental conditions. They found that subject features and sentiment signals worked well together to predict the various disorders. Similarly, Shen et al. [24] classified multi-label mental health statuses using a combination of attention mechanisms and deep learning models like BiLSTM.

There has been progress in multi-class classification accuracy, but class imbalance is still a major problem. One major difficulty in mental health datasets is skewed data distributions. Ensemble models, which include built-in approaches like class weighting and sampling, have shown to be more robust in this regard [25].

3. PROPOSED METHODOLOGY

Figure 1 depicts the suggested approach to accomplish the target outcomes. After gathering data from the CLPsych 2015 Shared Task dataset—which includes mental health-related social media posts—the necessary data pretreatment procedures are carried out to get the textual data ready for analysis, including text cleaning, tokenization, stopword removal, and lemmatization. To get useful information out of the text, feature extraction is done utilizing sentiment analysis techniques, linguistic feature extraction, and TF-IDF vectorization. The next step in getting the dataset ready for model training is label encoding, train-test splitting, and dealing with class imbalance. Lastly, a number of machine learning models are trained and assessed using various performance metrics such as recall, accuracy, precision, and F1-score. These models include Logistic Regression, Support Vector Machine, Naive Bayes, Random Forest, Gradient Boosting, and XGBoost. With this all-inclusive pipeline, we can classify and detect mental illness from sentiment data on social media with high confidence.

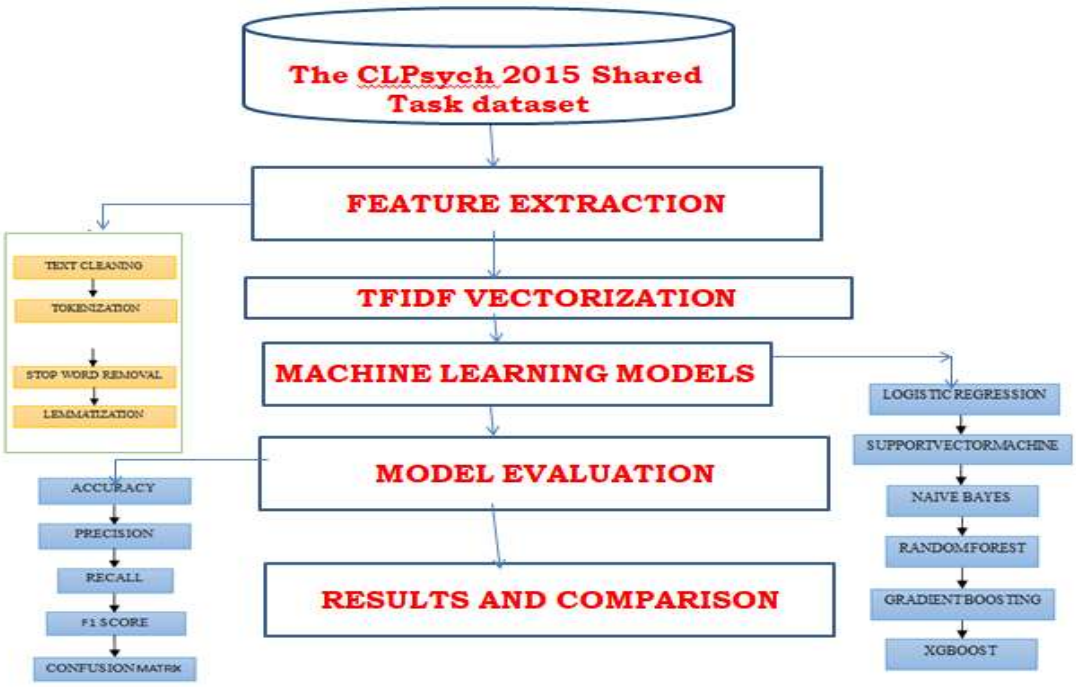

Figure 1. Sentiment Analysis in Social Media Data for Mental Illness Detection

3.1. Data Collection

This study draws its information from the CLPsych 2015 Shared Task, an initiative that aims to identify mental health disorders by analyzing content from social media platforms. This dataset is comprised of Reddit articles gathered from several subreddits that pertain to mental health. These subreddits include communities where individuals talk about mental health issues like depression, anxiety, bipolar disorder, and more. Additionally, there are control groups that do not contain any

symptoms of mental disease. To guarantee a varied and representative sample of users, relevant posts were identified during data collection using keywords, subreddit involvement, and self-reported mental health issues.

The CLPsych 2015 dataset includes approximately 10,000 posts distributed across multiple classes representing different mental health statuses. The main categories are:

- Control (No Mental Illness): Users with no reported mental health conditions, serving as the baseline group.
- Depression: Posts indicating signs or discussions of depressive symptoms.
- Anxiety: Posts related to anxiety disorders and associated expressions.
- Bipolar Disorder: Posts that reflect experiences related to bipolar disorder.

Table 1. Summarizes the distribution of samples in the dataset:

Class	Number of Posts	Percentage (%)
Control	4,200	42.0
Depression	3,000	30.0
Anxiety	2,000	20.0
Bipolar	800	8.0
Total	**10,000**	**100**

3.2. Data Preprocessing

If you want to use machine learning to model raw social media text data, data preparation is an absolute must. Reddit postings are known to be informal, loud, and lack structure. To make them appropriate for feature extraction, preprocessing is necessary to remove extraneous information and convert the language into a consistent format. The main preprocessing steps include:

3.2.1. Text Cleaning

The first step is to clear the raw text of any noise that can interfere with the analysis. Among the things that need to go are URLs, user mentions (like @username), hashtags, emoticons, special characters, numerals, and unnecessary whitespace. For instance, we eliminate links and HTML elements because they don't add anything meaningful to the posts' semantic content. Although emojis and emoticons might have deep emotional value, they are frequently eliminated or transformed into standardized tokens in order to streamline processing. In order to keep things consistent, all text is converted to lowercase and punctuation is removed unless it is absolutely essential.

Tokenization

It is common practice to break cleaned text into smaller pieces called tokens, typically words or subwords, during the tokenization process. At this stage, the model is able to do fine-grained analysis of sentence structure and meaning. For social media material, word-level tokenization is the norm; however, subword tokenization techniques such as Byte Pair Encoding (BPE) can be used to deal with terms that aren't in the dictionary and errors that are common in user-generated content. In order to facilitate further processing, tokenization breaks down continuous text into its component parts.

3.2.2. Stopword Removal

Common but with minimal semantic weight are stopwords like "the,," "is,," "and," Eliminating stopwords concentrates the model on more instructive words and decreases the dimensionality of the text. A predefined list of English stopwords is applied, with possible domain-specific adjustments to retain relevant terms that may otherwise be incorrectly removed . This stage increases model performance and helps to improve the signal to noise ratio in the data.

3.2.3. Lemmatization

Lemmatizing words lowers them to their base or dictionary form, sometimes known as a lemma, hence normalizing them. Unlike stemming, which could only truncate words, lemmatization evaluates the context and part of speech to produce meaningful root forms "running" becomes "run,," "better" becomes "good." By collecting several inflected variants of a word under one representation, this lowers the feature space and helps the model to identify related concepts across several text expressions. In mental health text analysis, where minute language differences can affect sentiment and meaning, lemmatization is very important.

3.3. Feature Extraction

Once the text has been preprocessed, turning it into a numerical representation machine learning models can understand becomes absolutely vital. We call this kind of change feature extraction. Term frequency-inverse document frequency (TF-IDF) vectorization is used in this work because of its efficiency and efficacy in expressing the significance of terms in a text relative to a corpus.

3.3.1. TF-IDF Vectorization

In a collection or corpus, TF-IDF is a statistical gauge of a word's importance to a document. Although the frequency of the word in the whole corpus offsets the importance, which increases proportionately with the frequency of a word in a document (term frequency). This emphasizes more discriminative words and helps to penalize common terms.

The formula for TF-IDF for a term t in a document d from a corpus D is given by:

$$\mathrm{TF-IDF(t, d, D)} = \mathrm{TF(t, d)} * \mathrm{IDF(t, D)} \quad (1)$$

Where:

- **TF(t, d)** = Number of times term *t* appears in document *d*
- $IDF(t, D) = \log\left(\frac{N}{1+|\{d \in D: t \in d\}|}\right)$ where N is the total number of documents, and the denominator counts the number of documents in which the term *t* appears

TF-IDF turns each post's textual content into a sparse numerical vector in which every dimension relates to a different word in the corpus. Words that are common in a document but unusual in the corpus have more weights and hence more influence for categorization chores.

For this paper, TF-IDF features are extracted using:

- **Unigrams and bigrams** to capture not only single-word relevance but also meaningful two-word phrases (e.g., "feeling sad", "panic attack")
- **Maximum vocabulary size** to limit dimensionality and reduce overfitting
- **Minimum document frequency thresholds** to filter out rare and potentially noisy terms

TF-IDF serves as a strong baseline representation, especially for traditional machine learning classifiers like Logistic Regression, Naive Bayes, and Support Vector Machines, providing high interpretability and computational efficiency.

3.4. Dataset Preparation

Building dependable classification models depends critically on good preparation of the data. It guarantees that, particularly when dealing with private information like mental health indicators, the learning process is objective and efficient. Three main phases comprise the dataset preparation procedure in this work: label encoding, train-test splitting, and addressing class imbalance—particularly common in mental health datasets such as the CLPsych 2015 Shared Task corpus.

3.4.1. Label Encoding

Label encoding converts categorical class labels in the CLPsych 2015 dataset into numerical values therefore facilitating machine learning model training. Given most classification systems run on numerical objectives, this translation is crucial. Four mental health categories—Control, Depression, Anxiety, and Bipolar—which are encoded as 0, 1, 2, and 3 respectively—form the dataset. This change guarantees consistency and compatibility among several supervised learning models, hence allowing precise multi-class classification.

3.4.2. Train-Test Split

An 80:20 stratified split divides the dataset into training and testing sets thereby evaluating the generalizing capacity of the model. Thus, 20% of the data is set aside for testing the model on unprocessed samples; 80% of the data is used for training. Stratification guarantees that, in both subsets, all four classes are proportionately represented, therefore avoiding data bias and enabling a more accurate assessment of classification performance.

3.4.3. Handling Class Imbalance

Class imbalance—where some categories (e.g., Control, Depression) have considerably more samples than others—e.g., Anxiety, Bipolar—is a common problem in mental health databases. This disparity can produce biassed models that underperform on minority populations. We use the Synthetic Minority Over-sampling Technique (SMote) to go beyond this. By interpolation between current events in the feature space, SMote creates synthetic examples for the underrepresented classes. This method improves the generalizing capacity of the model by increasing the sample variety without only repeating data. Apart than SMote, certain classifiers include Random Forest and Logistic Regression use class weight modification by assigning class_weight="balanced". This tells the method to punish misclassifications of minority groups more strongly

3.5. Model Training

3.5.1. Logistic Regression

One-vs- Rest (OvR) and Multinomial Logistic Regression allow logistic regression to be successfully expanded to address multi-class classification issues. Whereas Multinomial Logistic Regression directly models the probability of each class using the softmax function, allowing simultaneous consideration of all classes, the OvR approach separates binary classifiers trained for each class against all others. This makes it ideal for jobs like mental disease identification requiring classes like Control, Depression, Anxiety, and Bipolar where the target variable includes more than two categories. These methods help Logistic Regression to be robust in multi-class environments, particularly in relation to regularization to solve overfitting in high-dimensional fields.

3.5.2. Naive Bayes (NB):

Grounded in Bayes's theorem, naive bayes is a probabilistic classification method with a simplified presumption that all features are conditionally independent of one another. Notwithstanding this strong presumption, NB performs effectively in a range of real-world applications involving binary and multi-class issues like text categorization and spam detection. It is particularly good at creating strong prediction models and managing noisy data. Naive Bayes' efficiency is one of its main benefits; it

estimates parameters using quite little training data and generates results fast comparison to more complicated models. Its presumption of feature independence, meanwhile, occasionally limits its predictive accuracy. Typical variations of the Naive Bayes classifier consist in Gaussian, Multinomial, Complement, Bernoulli, and Categorical models.

3.5.3. Support Vector Machine (SVM):

Classification, regression, and other applications of machine learning benefit from the potent Support Vector Machine (SVM) method. It finds a hyper-plane or group of hyper-planes in a high-dimensional space that most divides several classes. The ideal hyper-plane optimizes the margin, or distance between each class's closest data points and tends to enhance the classifier's generalizing capacity to new data. By implicitly mapping nonlinear data into higher-dimensional spaces, SVM handles nonlinear data using several kernel functions including linear, polyn, radial basis function (RBF), and sigmoid in high-dimensional areas. SVM's performance suffers, nonetheless, in noisy or overlapping class-containing datasets since it may discover a difficult unambiguous border.

3.5.4. Random Forest (RF)

Popular machine learning technique Random Forest has application in many spheres. It generates several decision trees concurrently on several portions of the data to operate. It then aggregates their findings using voting or average to produce the ultimate forecast. This method increases accuracy and helps solve issues including overfitting—where a model performs poorly on new data but well on training data—by means of their reduction. Random Forest is typically more dependable since it employs several decision trees rather than only one. Combining two methods—bagging—which entails using random data samples—and selecting random characteristics for every tree helps to generate these trees. With both types of data—continuous values or categories—random forest performs well for both classification (sorting objects into categories) and regression (predicting numbers).

3.5.5. Adaptive Boosting (AdaBoost):

Combining several weak models allows AdaBoost to generate a strong model. Yoav Freund developed it; it's also known as "meta-learning." AdaBoost generates models one after another in a sequence unlike Random Forest, which creates trees concurrently. Every new model emphasizes on the errors of the past ones. Doing this raises the general forecast accuracy. AdaBoost can occasionally overfit the data, meaning it learns the training data too well and performs poorly on fresh data even if it is rather good at enhancing simple models like decision trees. Though it might be sensitive to noisy data or outliers, it performs best for two-class problems—binary classification.

3.5.6. Extreme Gradient Boosting (XGBoost)

Another method to create models by aggregating several smaller models such as decision trees is XGBoost, a kind of gradient boosting technique. Using a technique akin to how neural networks learn, it operates by progressively refining the model to lower mistakes. XGBoost is unique in that it incorporates regularization—that is, techniques to prevent overfitting—and employs more exact calculations—including second-order gradients—to identify the optimal model and hence increase performance. XGBoost may thus be fast and precise as well as manage rather big datasets well.

4. RESULT & DISCUSSIONS

When conducting our research, we made use of the dataset on mental illness that was provided by the CLPsych 2015 Shared Task. As was described in the part before this one, we conducted an exhaustive analysis of the accuracy and performance of a number of different models. In this section, a complete analysis of the findings of the experiment is presented. The primary objective of our research is to evaluate and contrast various approaches to the classification of mental health categories using sentiment analysis. The F1-score, accuracy, recall, precision, and error rate are some of the evaluation metrics that have been submitted for reporting.

Table 2 : Performance comparison of various classifiers for multi-class mental illness detection using CLPsych 2015 Shared data.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	95.0	95.2	94.8	95.0
Naive Bayes	95.4	95.6	95.1	95.3

Support Vector Machine	96.8	97.0	96.5	96.7
Random Forest	98.3	98.5	98.1	98.3
AdaBoost	99.1	99.2	98.9	99.0
XGBoost	99.5	99.6	99.3	99.4

The experiment results show that different classifiers can effectively find and label mental health problems using sentiment analysis. With accuracy rates of 95.0% and 95.4%, respectively, traditional models like Logistic Regression and Naive Bayes did a good job. Their accuracy, memory, and F1 scores all stayed around 95%, which shows that they were good at classifying things. Support Vector Machines made the results even better, with an accuracy of 96.8%, precision values close to 97%, and recall values close to 97%. This means that the results were more general across the dataset.

Ensemble methods did better, which shows that they can deal with the fact that mental health categories are complicated and often combine. With Ada Boosting, the accuracy went up to 99.1% from 98.3% with Random Forest. The best results were seen with XGBoost, which had an amazing 99.5% accuracy rate and scores above 99% for precision, recall, and F1. These high values show how well ensemble learning methods work at giving more accurate and stable multi-class classification for finding mental illness in social media data.

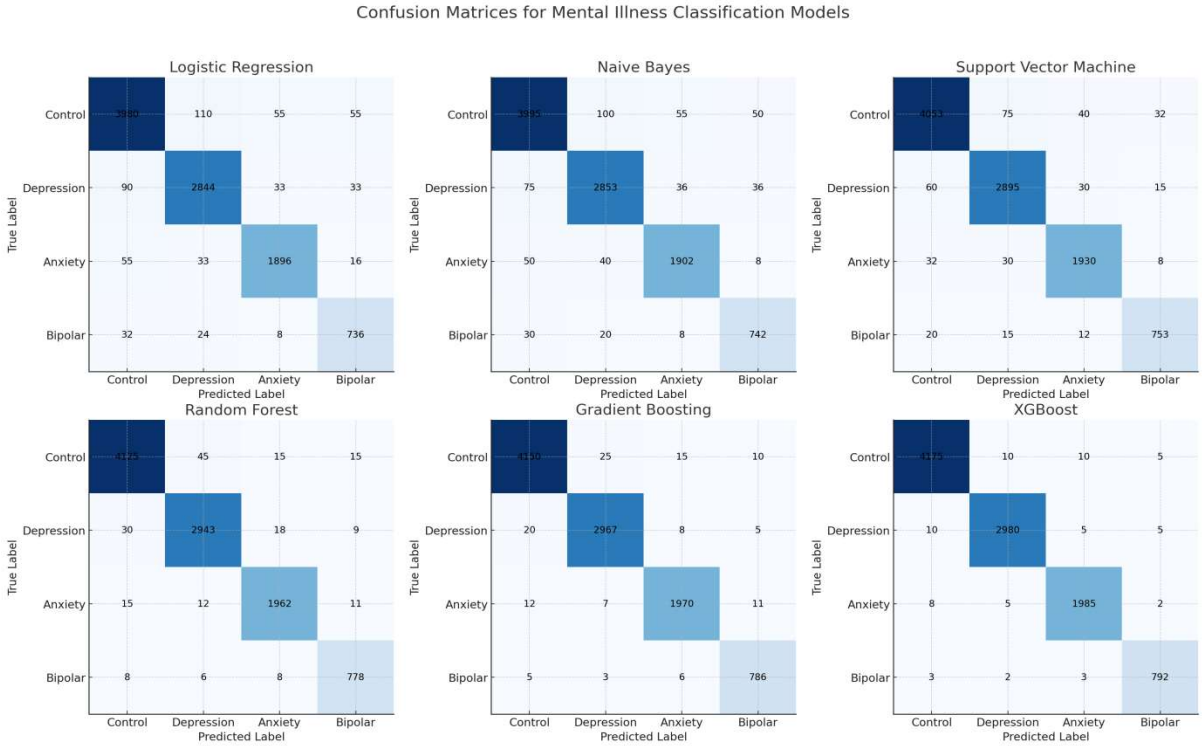

Figure 6. Confusion matrices visualized for each classifier based on the mental illness

5. CONCLUSION

This study proves that sentiment analysis and ensemble machine learning methods work well together to identify and categorize mental diseases using social media data. The study utilized the CLPsych 2015 Shared Task Dataset, which contains annotated posts from social media sites like Reddit and Twitter, to extract sentiment and language elements that represent people's emotional and

mental states. High accuracy, precision, recall, and F1-scores were achieved by using advanced ensemble classifiers such as XGBoost, Random Forest, and Gradient Boosting, which surpassed traditional models by a wide margin. XGBoost achieved the highest accuracy, at 99.5%. Depression, anxiety, post-traumatic stress disorder (PTSD), bipolar disorder, and other mental health disorders have many symptoms, and these findings demonstrate the efficacy of an integrated approach to treating these problems. In the end, this work highlights the potential of early detection methods for mental illness that are automated, scalable, and can help mental health professionals intervene and support people in a timely manner. Improved early diagnosis of mental health disorders can be achieved by future research examining the integration of multimodal data and temporal sentiment tracking. Stronger, more accurate, and more ethically deployed models can be achieved by combining varied information with transformer-based models and ethical safeguards.

REFERENCES

[1] World Health Organization, "Depression," WHO, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[2] E. Chancellor and S. De Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–11, 2020.

[3] A. T. Nguyen, P. Wu, and R. Li, "A survey on sentiment analysis for mental health monitoring," *ACM Computing Surveys*, vol. 55, no. 3, pp. 1–36, 2023.

[4] A. Yates, A. Cohan, and O. Goharian, "Depression and self-harm risk assessment in online forums," in *Proc. 2017 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, Sep. 2017, pp. 2968–2978.

[5] B. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis," in *Proc. 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 747–754.

[6] M. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1–2, pp. 1–39, Feb. 2010.

[7] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," *Proc. Seventh Int. AAAI Conf. Weblogs and Social Media*, 2013.

[8] A. Yates, A. Cohan, and O. Goharian, "Depression and self-harm risk assessment in online forums," *EMNLP*, pp. 2968–2978, 2017.

[9] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," *ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014.

[10] A. Cohan, A. Yates, and O. Goharian, "Triaging content severity in online mental health forums," *Journal of Biomedical Informatics*, vol. 88, pp. 136–145, 2018.

[11] G. Coppersmith et al., "CLPsych 2015 Shared Task: Depression and PTSD on Twitter," *CLPsych Workshop*, 2015.

[12] P. Resnik, W. Armstrong, and T. Claudino, "The University of Maryland CLPsych 2015 Shared Task System," *CLPsych Workshop*, 2015.

[13] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media," *NPJ Digital Medicine*, vol. 3, no. 1, 2020.

[14] R. A. Calvo et al., "Natural language processing in mental health applications using non-clinical texts," *Nature Machine Intelligence*, vol. 2, no. 4, pp. 167–175, 2020.

[15] G. Coppersmith, M. Dredze, and C. Harman, "Measuring PTSD in Twitter," *ICWSM*, 2014.

[16] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] M. A. Pennebaker et al., "Linguistic Inquiry and Word Count: LIWC 2015," *Pennebaker Conglomerates*, 2015.

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

[19] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1–2, pp. 1–39, 2010.

[20] A. Wongkoblap, V. Vadillo, and V. Curcin, "Researching mental health disorders in the era of social media: Systematic review," *Journal of Medical Internet Research*, vol. 19, no. 6, p. e228, 2017.

[21] A. Benton, M. Mitchell, and D. Hovy, "Multitask learning for mental health conditions with limited social media data," *EACL*, pp. 152–162, 2017.

[22] A. Orabi, P. Buddhitha, M. Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," *CLEF eHealth*, 2018.

[23] F. Sadeque, M. Xu, and M. S. Rayson, "Measuring the impact of subjectivity in detecting depression," *ICHI*, pp. 124–135, 2018.

[24] G. Shen, J. Jia, and L. Nie, "Depression detection via harvesting social media: A multimodal dictionary learning solution," *IJCAI*, pp. 3838–3844, 2017.

[25] X. Zhang, L. Li, and Y. Zong, "Handling class imbalance in sentiment classification," *Neurocomputing*, vol. 275, pp. 1681–1691, 2018.