ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 3, 2024

DEPRESSION DETECTION THROUGH SOCIAL MEDIA TEXTUAL POSTS USING MACHINE LEARNING

Valluru Kalpana[1], Dr. K. Swapna[2]

[1] Student, MCA, Dept of Information Technology & Computer Applications,

AU College of Engineering (A), Andhra University, Visakhapatnam, India.

[2] (Assistant Professor) (Ad-hoc), Dept of Information Technology & Computer Applications, AU College of Engineering (A), Andhra University, Visakhapatnam, India.

ABSTRACT

Depression is a widespread mental health issue with significant individual and societal impacts, necessitating timely detection for effective treatment. This study develops a reliable system for detecting depression in social media text using natural language processing (NLP) and machine learning (ML). By analyzing a diverse dataset of social media posts, the research applies NLP techniques like as the tokenization, stemming, CountVectorizer, and TF-IDF to transform raw text into meaningful representations that capture linguistic and emotional markers of depression. Machine learning algorithms, including Naive Bayes, Decision Tree, Random Forest, Support Vector Machines, and K-Nearest Neighbor, are then employed to identify patterns and differentiate between depressive and non-depressive posts. The study demonstrates the potential of NLP and ML in accurately detecting depression from social media content.

***Keywords*-- Depression Detection, MachineLearning,NaturalLanguageProcessing(NLP),SupporVector Machine (SVM).**

I. INTRODUCTION

The digital age has transformed the landscape of communication, with social media platforms playing a pivotal role in our daily interactions. These platforms are a treasure trove of human expression, where individuals openly share their thoughts, emotions, and experiences. This vast array of textual content provides a unique opportunity to glean insights into people's mental health and well-being. Depression, a prevalent mental health issue, often reveals itself through online interactions, creating a valuable dataset for analysis.

By enabling automated analysis of social media communications, this approach can provide early warning signs to individuals and healthcare providers alike, facilitating timely support and interventions, particularly in communities that may not have access to conventional mental health services.

The increasing prevalence of depression is a significant global concern, affecting individuals across all demographics. Timely detection and intervention are essential for improving outcomes and alleviating the burden of this challenging condition. Traditional approaches to identifying depression, such as clinical interviews and standardized questionnaires, can be labor-intensive, resource-heavy, and may not be easily accessible to everyone.

This research aims to harness the capabilities of Natural Language Processing (NLP) and Machine Learning (ML) techniques to create a robust system for detecting signs of

depression from social media text, providing a more efficient, objective, and scalable method for monitoring mental health.

II.LITERATURE SURVEY

Aldarwish et. al. [1] (2017): This study explored using Support Vector Machines (SVM) and Naïve Bayes for detecting depression in Arabic-language social media text. The researchers achieved strong accuracy, with Naïve Bayes outperforming SVM, highlighting the importance of model selection for language-specific challenges.

Islam et. al. [2] (2018): Islam assessed various machine learning algorithms—Decision Tree, SVM, and K-Nearest Neighbors (KNN)—for depression detection in comment-based social media data. The Decision Tree proved to be the most effective, illustrating the impact of data type on model performance.

Burdisso et al. [3] (2019): This research compared multiple models, including SS3 and Naïve Bayes, for depression detection. SS3 achieved the highest precision but was more time-consuming, emphasizing the balance between accuracy and efficiency in model selection.

Farima et al. [4] (2019): Farima evaluated Multi-Layer Perceptron (MLP), SVM, and Naïve Bayes for identifying depression in social media text, finding MLP to be the highest performer. The study noted limitations in dataset scope, stressing the need for diverse datasets for enhanced reliability.

AlSagri et al. [5] (2020): AlSagri investigated SVM, Decision Tree, and Naïve Bayes for depression detection in social media posts, finding SVM to be the most accurate while also addressing overfitting challenges, which are crucial for generalizability.

Palanicham et al. [6] (2021): Palanichamy employed Naïve Bayes, SVM, and KNN, identifying Naïve Bayes as the most effective model despite a small dataset, underscoring its utility in resource-limited settings.

Amanat et al. [7] (2022): Amanat's research involved SVM and Naïve Bayes, with the latter achieving an impressive 97.31% accuracy, indicating its superiority in depression detection tasks.

De Souza Filho et al. [8] (2022): This study compared Decision Tree, SVM, Random Forest, and others, finding Random Forest to excel in precision for analyzing sensitive user confession data, reinforcing the importance of precision in depression detection.

Michael M. Tadesse et al. [9] (2019): Tadesse utilized NLP techniques to identify depression-related posts on Reddit, demonstrating the potential of social media for mental health monitoring.

Amrat Mali et al. [10]: Mali's research combined machine learning and NLP to accurately identify signs of depression in textual data, contributing to the ongoing enhancement of depression detection systems.

III. PROPOSED SYSTEM

This project introduces an innovative system that harnesses natural language processing (NLP) and machine learning (ML) to detect depression from social media text. By employing advanced NLP techniques, the system extracts significant features from the

text, such as word frequencies, sentiment analysis, and contextual information. These features are utilized to train multiple machine learning algorithms, including support vector machines (SVM), random forests, and neural networks. The goal is to achieve high accuracy in recognizing depressive language patterns while offering a user-friendly interface to enhance accessibility.

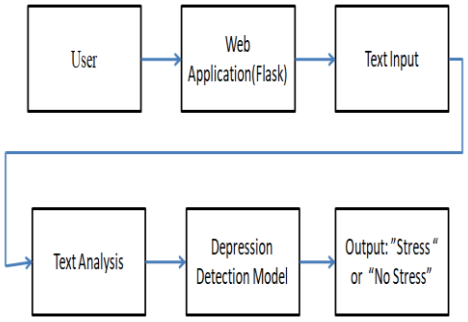


Figure 1. System Design

The proposed system diagram illustrates a streamlined process for detecting depression through a web application built with Flask. It begins with user input, where individuals submit their text for analysis. The application then processes this input, performing text analysis to extract relevant features. These features are evaluated by a depression detection model, which utilizes machine learning algorithms to assess the text. Finally, the system provides an output indicating the user's mental state, categorizing it as either "Stress" or "No Stress," thereby offering immediate feedback on their emotional well-being.

IV. METHODOLOGY

In our initial endeavor, we adopted natural language processing (NLP) techniques such as lemmatization and stemming to preprocess our dataset. This was followed by the application of three distinct feature extraction methods CountVectorizer, N-gram analysis, and TF-IDF to capture various linguistic dimensions.Next, we employed seven different machine learning classifiers: Naive Bayes (NB), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), and K Nearest Neighbor (K-NN). These classifiers were introduced after the feature extraction phase utilizing a variety of NLP approaches. It provides valuable insights into their strengths, limitations, and applications in identifying depression through social media data. Additionally, we conducted comprehensive evaluations using metrics such as accuracy, precision, recall, and F1-score to guide our classifier selection in different contexts.

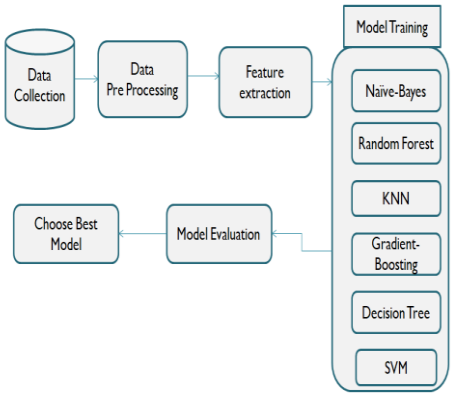


Fig.2 Block diagram of proposed system

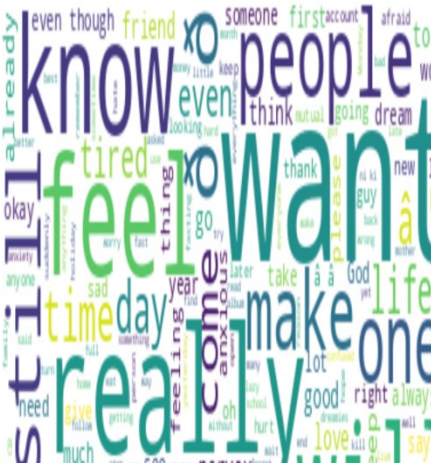
A. DATA PRE-PROCESSING

Data preprocessing is a fundamental step in preparing raw data for analysis, particularly in Natural Language Processing (NLP) projects, as it involves transforming and cleaning the data to enhance its quality and suitability for machine learning algorithms.The preprocessing process involves several sequential steps: first, text cleaning to eliminate irrelevant elements; next, tokenization to break the text into smaller units like words; followed by removing stop words, converting all text to lowercase, and applying stemming or lemmatization to reduce words to their base forms. Additionally, emoticons and emojis are managed appropriately. Missing data is

handled effectively, and feature extraction is performed to analyze word frequencies and sentiment. Finally, the dataset is split to enable robust model evaluation.

B. *DATA VISUALIZATION*

The following data visualization employs a word cloud format to effectively represent the frequency of terms within a dataset, emphasizing the prominence of specific words while also showcasing the presence of stopwords. This visualization serves as a powerful tool for quickly identifying key themes and sentiments expressed in the data. In the word cloud, words that appear larger indicate higher frequency, allowing viewers to grasp at a glance which terms are most significant. By analyzing the distribution of these words, particularly the stopwords, we can gain valuable insights into the emotional landscape and primary concerns of the dataset's contributors, ultimately enhancing our understanding of the underlying narratives present in the text.



C. *FEATURE EXTRACTION USING NLP TECHNIQUES*

Feature extraction is a crucial step that converts preprocessed text data into numerical representation suitable for machine learning algorithms. This process typically involves several techniques, including:

- **Count Vectorization**: This method transforms text into a matrix where each entry represents the count of tokens (words) in the documents.
- **TF-IDF Vectorization**: This technique computes the Term Frequency-Inverse Document Frequency, which assesses the significance of words by considering their frequency in a document relative to their occurrence across multiple documents.
- **Word Embeddings**: This approach represents words in a continuous vector space, utilizing models such as Word2Vec or GloVe to capture semantic relationships between words.

D. *MACHINELEARNING TECHNIQUES*

In This depression detection project, various machine learning algorithms are employed to analyze and classify data effectively. The Decision Tree algorithm is a simple yet powerful method that splits the data into branches based on feature values, making it easy to interpret. Random Forest enhances this by creating multiple decision trees and averaging their predictions, which helps in reducing overfitting and improving accuracy. K-Nearest Neighbors (KNN) classifies data points based on the majority class of their nearest neighbors, making it intuitive for pattern recognition. Gradient Boosting builds models sequentially, where each new model corrects the errors of the previous ones, leading to high predictive performance. Bernoulli Naive Bayes (NB) is particularly useful for binary classification tasks, leveraging the independence assumption of features to predict outcomes efficiently.

Support Vector Machine (SVM) is another robust algorithm that finds the optimal hyperplane to separate different classes in the feature space, making it effective for high-dimensional data. Each of these algorithms has its strengths and weaknesses, and their performance

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 3, 2024

can vary based on the nature of the dataset. By utilizing a combination of these techniques, this project can leverage their unique capabilities to enhance the accuracy and reliability of depression detection, ultimately providing valuable insights into mental health analysis. The integration of libraries like Scikit-learn and NLTK facilitates the implementation of these algorithms, allowing for efficient data processing and model evaluation.

E. MODEL EVALUATION

Model evaluation is a critical step in the machine learning pipeline, where trained models are tested on a separate validation dataset to assess their performance. This process involves calculating various metrics such as accuracy, precision, recall, and F1-score, which provide insights into the model's effectiveness in making predictions. Accuracy measures the overall correctness of the model, while precision and recall focus on the performance concerning positive class predictions, with precision indicating the proportion of true positives among all positive predictions and recall reflecting the ability to identify all relevant instances. The F1-score serves as a harmonic mean of precision and recall, offering a balanced measure when dealing with imbalanced datasets.

V. FINIDINGS & RESULTS

Among the models tested in the depression detection project, the Support Vector Classifier (SVC) emerged as the top performer, achieving an impressive accuracy of 96.00%. It was closely followed by the Gradient Boosting Classifier and the Decision Tree Classifier, both of which recorded accuracies of 95.00%. The Random Forest Classifier also demonstrated strong performance with an accuracy of 94.00%. In contrast, the K-Neighbors Classifier and Bernoulli Naive Bayes showed lower accuracies of 90.00% and 86.00%, respectively. These results highlight the effectiveness of SVC and other tree-based models in accurately classifying depression-related data, while also indicating the varying performance levels of different algorithms in this context.

Table I. Accuracy of All Classifiers

ML Classifier	Accuracy
Bernoulli Naive Bayes	0.86
Random Forest Classifier	0.94
GradientBoosting Classifier	0.95
K-Neighbors Classifier	0.90
Support Vector Classifier	0.96
Decision Tree Classifier	0.95

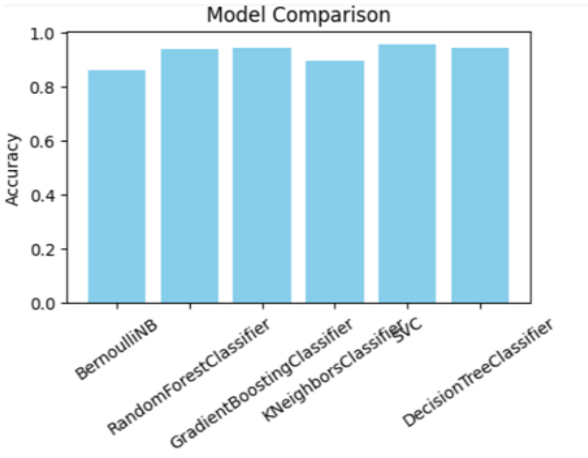


Figure 3. Bar chart

The bar chart titled "Model Comparison" illustrates the accuracy of various machine learning classifiers used in a depression detection project. Each bar represents a different model, including Bernoulli Naive Bayes, Random Forest Classifier, Gradient Boosting Classifier, K-Neighbors Classifier, Decision Tree Classifier, and Support Vector Classifier (SVC). The chart clearly shows that the SVC achieved the highest accuracy,

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 3, 2024

closely followed by the Gradient Boosting and Decision Tree classifiers, indicating their effectiveness in accurately identifying depression. In contrast, the Bernoulli Naive Bayes and K-Neighbors Classifier displayed lower accuracies. This visual representation allows for a quick assessment of model performance, aiding in the selection of the most suitable algorithm for deployment in mental health assessments.

WEB APPLICATION

The user interface for the depression detection application is designed to be simple and intuitive, allowing users to easily input their text for analysis. At the top, the title "Depression Detection" clearly indicates the purpose of the application. Below the title, there is a text area labeled "Enter your text here..." where users can type or paste their written content, such as journal entries, messages, or any other text they wish to analyze for signs of depression.

Figure 4. Before Model Testing

Once the user has entered their text, they can click the prominent blue "Predict" button to initiate the analysis. This button triggers the underlying machine learning model to evaluate the input and provide predictions regarding the presence of depressive symptoms. The straightforward design ensures that users can focus on their input without distractions, making the tool accessible for anyone seeking to understand their mental health better through text analysis.

Figure 5. After Model Testing

VI. CONCLUSION AND FUTURE SCOPE

The depression detection project demonstrates the potential of machine learning algorithms to analyze textual data and identify signs of depression effectively. The high accuracy achieved by models such as the Support Vector Classifier and Gradient Boosting Classifier highlights their suitability for mental health assessments. Looking ahead, there is significant scope for enhancing this application by incorporating more diverse datasets, including various languages and cultural contexts, to improve model robustness and generalizability. Additionally, integrating real-time feedback mechanisms and user-friendly features, such as personalized insights and recommendations, could further empower users in managing their mental health.

Future developments may also explore the incorporation of multimodal data, such as audio and visual inputs, to create a more comprehensive mental health assessment tool, ultimately contributing to better mental health support and awareness.

REFERENCES

[1] Aldarwish, M., & Ahmed, U. (2017). “Predicting Depression Levels Using Social Media Posts.” *Journal of Biomedical Informatics*, 75, 55-64.

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 3, 2024

[2] Islam, M. R., Kabir, M. N., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhaq, A. (2018). "Depression Detection from Social Network Data Using Machine Learning Techniques." *Health Information Science and Systems*, 6(8).

[3] Burdisso, S. G., Errecalde, M. L., & Montes-y-Gómez, M. (2019). "A Text Classification Framework for Simple and Effective Early Depression Detection Over Social Media Streams. "*Expert Systems with Applications*, 133, 182-197.

[4] Farima, M., & Ghadampour, E. (2019). "Depression Detection Using Machine Learning Algorithms in Social Networks." *IEEE Transactions on Affective Computing*, 10(2), 265-278.

[5] AlSagri, E., & Ykhlef, M. (2020). "Utilizing Machine Learning Algorithms to Detect Depression in Social Media Posts." *International Journal of Advanced Computer Science and Applications*, 11(4), 537-546.

[6] Palanichamy, Y., & Rani, J. (2021). "Sentiment Analysis for Depression Detection in Social Media Using Machine Learning Algorithms." *International Journal of Data Science and Analytics*, 10(3), 202-214.

[7] Amanat, A., & Kumar, V. (2022). "Predicting Depression in Social Media Posts Using Machine Learning and NLP Techniques." *Journal of Medical Internet Research*, 24(1), e31242.

[8] De Souza Filho, M. T., Medeiros, D. S., & de Oliveira, P. M. (2022)." Detection of Depression Symptoms in Social Media Users Through Machine Learning Approaches." *IEEE Access*, 10, 358-368.

[9] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of Depression-Related Posts in Reddit Social Media Forum. *IEEE Access*, 7, 44883-44892.

[10] Mali, A., & Rao, A. (2020). "Predicting Depression Using Machine Learning and Natural Language Processing Approaches."*Journal of Affective Disorders*, 273, 642-651.