iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71171

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Risk Prediction Models in Adolescents Using ML Techniques

Deepali Deshpande[1], Atharv Asalkar[2], Prathamesh Bansode[3], Niraj Baviskar[4], Atharva Belote[5], Hamlin Nadar[6]
Department of InformationTechnology Vishwakarma Institute of Technology Pune, India

Abstract: ***Adolescents are at high risk for mental health issues, including depression, due to factors such as social media exposure, peer pressure, and behavioral influences. Early detection of at-risk individuals enables timely intervention and support. This study presents a hybrid depression detection model that integrates Naïve Bayes, TF-IDF, and emoji sentiment analysis to enhance classification accuracy. The approach utilizes textual and emoji-based sentiment cues extracted from social media data to identify depressive tendencies. The model is evaluated against traditional classifiers, demonstrating improved precision, recall, and F1-score. A comparative analysis of different machine learning techniques highlights the model's effectiveness in detecting depression-related expressions. The findings emphasize the potential of AI-driven approaches in mental health monitoring, particularly for adolescents, by leveraging real-time social media interactions. This research contributes to the growing field of automated mental health assessment, offering a scalable and data-driven solution for early intervention. Future work will explore deep learning and multimodal analysis to further refine depression detection and provide more comprehensive insights into mental health patterns.***

Keywords: Depression Detection, Social Media Analytics, Hybrid Model, Naïve Bayes, TF-IDF, Emoji Sentiment Analysis, Mental Health Monitoring, Sentiment Analysis, Natural Language Processing (NLP), Affective Computing.

I. INTRODUCTION

Mental health disorders, particularly depression, have seen a significant rise in prevalence, with social media serving as a primary medium where individuals express their emotions and thoughts. The increasing digitization of human communication presents an opportunity to leverage artificial intelligence (AI) for early detection of depressive tendencies. Traditional clinical assessments for depression are resource-intensive and often inaccessible, highlighting the need for automated, data-driven approaches capable of analyzing large-scale social media interactions.

Existing depression detection models rely on text-based sentiment analysis, employing machine learning techniques such as support vector machines (SVM), deep learning models, and lexicon-based methods. While these approaches have demonstrated success, they often fail to capture the nuanced emotional context present in informal online conversations. Additionally, emoji-based sentiment, which plays a crucial role in modern digital expression, is frequently overlooked, leading to incomplete sentiment representations. The challenge lies in integrating linguistic and non-linguistic features to create a robust, scalable, and interpretable depression detection model.

To address these limitations, this research proposes a hybrid depression detection framework that combines Naïve Bayes, TF-IDF-based feature extraction, and emoji sentiment analysis. By leveraging statistical learning and affective computing, the model enhances the accuracy of depression classification in social media text. The inclusion of emoji sentiment provides an additional layer of emotional context, compensating for the limitations of purely textual analysis. The study evaluates the proposed model's performance against existing methodologies using metrics such as precision, recall, and F1-score, ensuring a comprehensive assessment of its effectiveness.

The key contributions of this research are as follows:

1) A novel hybrid approach integrating Naïve Bayes, TF-IDF, and emoji sentiment analysis for improved depression detection.
2) A comparative study of traditional sentiment analysis models versus the proposed method, highlighting accuracy improvements.
3) Incorporation of emoji-based sentiment analysis, bridging a research gap in computational mental health assessment.
4) Evaluation across real-world social media datasets, ensuring applicability to diverse linguistic patterns and user behaviors.
5) Insights into multimodal sentiment analysis, contributing to the advancement of AI-driven mental health diagnostics.

The remainder of this paper is structured as follows: Section II presents a review of existing work in depression detection and sentiment analysis.

Section III details the methodology, including dataset selection, feature engineering, Section IV consists of the proposed system and model architecture. Section V discusses experimental results and model evaluations. Section VI discusses various other aspects. Finally, Section VII concludes the paper and suggests future research directions.

II. RELATED WORKS

In research paper [1] A 2023 study in 'European Child & Adolescent Psychiatry' reviewed the impact of COVID-19 lockdowns on youth mental health, analyzing 61 studies involving 54,999 participants aged 0–19. Anxiety (79.4%), depression (32.6%), irritability (73.7%), anger (42.5%), and hyperactivity symptoms (21.3%) were reported, with pre-existing conditions, excessive media use (82.5%), and disrupted routines (77%) as key risk factors, while strong parent-child communication (84%) offered protection. The study calls for public health strategies and clinical guidelines to address these challenges during crises.

The paper[2] "A review on sentiment analysis from social media platforms" published in Expert Systems with Applications (2023) provides a comprehensive overview of sentiment analysis techniques, including machine learning, lexicon-based methods, and advanced models like BERT and GPT-3. It explores the temporal dynamics and causal relationships of sentiment analysis, particularly in domains like politics, health, and finance. The study emphasizes the importance of temporal and causal effects in sentiment analysis, highlighting emerging research opportunities, especially in combining AI techniques with real-world applications. Key methodologies like Granger causality and spatio-temporal analysis are discussed for understanding sentiment evolution over time.

In the research paper[3] The umbrella review by Valkenburg et al. (2022) synthesizes 25 meta-analyses, systematic, and narrative reviews (2019–2021) on social media use (SMU) and adolescent mental health. It highlights predominantly weak or inconsistent associations, with some reviews citing substantial effects. Identified gaps include over-reliance on cross-sectional data, inconsistent definitions of SMU and mental health, and limited focus on mediators or risk factors. Future research calls for longitudinal studies, objective SMU measures, and content-focused analyses.

The paper [4]"A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction" explores various feature selection techniques crucial for reducing dimensionality in genomic datasets, addressing the challenge of the "curse of dimensionality." It highlights filter, wrapper, and embedded methods, emphasizing their strengths and weaknesses in disease risk prediction using SNP data. It also discusses the impact of feature redundancy and interactions, particularly epistasis, on prediction accuracy. Published in Frontiers in Bioinformatics, this review provides insights into optimizing machine learning models for precision medicine.

The paper[5] titled "A Systematic Review of Social Media-Based Sentiment Analysis: Emerging Trends and Challenges" published in *Decision Analytics Journal* (2022) provides an extensive examination of sentiment analysis (SA) techniques applied to social media data. It categorizes methods into lexicon-based, machine learning-based, and hybrid approaches, highlighting algorithms such as SentiWordNet, TF-IDF, Word2Vec, BERT, and BiLSTM. Key challenges discussed include data-related issues, model accuracy, and evaluation metrics. The paper also addresses the limitations in current research, emphasizing the need for efficient, scalable solutions for real-time sentiment monitoring.

In research paper[7] The review by Hoover and Bostic (2021) in *Psychiatric Services* emphasizes schools as vital for child and adolescent mental health, advocating for multi-tiered systems of support (MTSS) like SEL and PBIS. It highlights improved access, early intervention, reduced stigma, and better academic and psychosocial outcomes. The authors recommend integrating mental health services, workforce training, and systematic monitoring to advance comprehensive school mental health systems.

The study by Magson et al. (2021) in Journal of Youth and Adolescence investigates the psychological impact of the COVID-19 pandemic on adolescents through a longitudinal approach. The study highlights significant increases in depressive symptoms, anxiety, and a decrease in life satisfaction, particularly among girls. Factors such as online learning difficulties, COVID-19-related worries, and family conflicts were found to worsen mental health, while social connection and adherence to lockdown measures served as protective factors. This research underscores the importance of understanding adolescent mental health during the pandemic, emphasizing the role of social relationships and environmental stressors.[8]

III. METHODOLOGY

The proposed hybrid depression detection model integrates Naïve Bayes, TF-IDF, and emoji sentiment analysis to enhance the accuracy of detecting depressive tendencies in social media text. This section details the methodology, including data collection, preprocessing, feature engineering, model development, and evaluation.

A. Data Collection

The dataset for this research was obtained from publicly available sources containing labeled tweets categorized as depressive (1) or non-depressive (0). The dataset includes textual content, emoji usage, timestamps, and metadata related to user engagement. A well-balanced dataset ensures that the model can effectively distinguish depressive text patterns.

B. Data Preprocessing

To improve model efficiency and eliminate noise, the following Natural Language Processing (NLP) techniques were applied:

1) Text Normalization: Conversion to lowercase to maintain consistency.
2) Tokenization: Splitting text into individual words using NLTK's word tokenizer.
3) Stopword Removal: Eliminating non-informative words (e.g., "the," "is," "and").
4) Stemming/Lemmatization: Converting words to their root forms (e.g., "crying" → "cry").
5) Punctuation and URL Removal: Removing unnecessary characters and links.
6) Emoji Extraction: Identifying and mapping emojis to sentiment scores using an emoji lexicon.

C. Exploratory Data Analysis (EDA)

A comprehensive exploratory data analysis (EDA) was conducted to understand word distributions, sentiment trends, and emoji influence. Statistical and visual tools such as word clouds, frequency histograms, and correlation heatmaps were used to identify patterns in depressive and non-depressive tweets. This step provided insights into commonly used depressive keywords and emoji associations.

D. Feature Engineering

Feature extraction was performed using a combination of linguistic and sentiment-based features:

1) TF-IDF (Term Frequency-Inverse Document Frequency): Assigns weights to words based on their importance within depressive and non-depressive texts.
2) Bag-of-Words (BoW): Represents text in a structured form to quantify word occurrences.
3) Emoji Sentiment Analysis: Each extracted emoji was assigned a sentiment polarity score based on an external sentiment lexicon.

These engineered features were then used as input to train the depression detection model.

E. Model Development

A hybrid approach was implemented by integrating three key techniques:

1) Naïve Bayes Classifier: A probabilistic text classification model that computes the likelihood of a tweet being depressive based on word occurrences.
2) TF-IDF Weighting: Enhances the classifier's performance by prioritizing important depressive indicators.
3) Emoji Sentiment Integration: Merges emoji sentiment scores into the final prediction to enhance accuracy.

The model was trained using a split dataset, ensuring a fair distribution of depressive and non-depressive tweets for robust learning.

F. Model Evaluation

The model was evaluated using multiple performance metrics, including:

1) Accuracy: Measures the percentage of correct predictions.
2) Precision: Evaluates how many predicted depressive tweets were truly depressive.
3) Recall: Measures the proportion of actual depressive tweets correctly identified.
4) F1-score: Provides a balance between precision and recall.

The model underwent hyperparameter tuning to optimize its performance.

G. Training and Evaluation

The dataset was split into 98% training data and 2% testing data to maximize model learning while retaining some data for evaluation.

To assess model performance, four key metrics were used:

1) Precision – Measures how many predicted depressive tweets were actually depressive.

2) Recall – Measures how many actual depressive tweets were correctly identified.
3) F1-score – Balances precision and recall for a more holistic evaluation.
4) Accuracy – Measures overall correct predictions across all tweets.

IV. PROPOSED SYSTEM

The proposed system is a web-based application designed to analyze tweets and predict whether they indicate signs of depression. It combines Natural Language Processing (NLP) techniques, machine learning models, and emoji sentiment analysis to provide accurate predictions. The system is built using the Flask framework, offering a user-friendly interface for input and results.

A. *Key Components:*

1) *Data Preprocessing:*

- Tweets are cleaned by removing special characters, numbers, and stopwords, and converted to lowercase.
- Emojis are extracted and analyzed using a custom sentiment mapping.

2) *Feature Extraction:*

Text is vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) for machine learning input.

3) *Machine Learning Model:*

A Multinomial Naive Bayes classifier is trained on labeled tweet data to predict depression.

4) *Emoji Sentiment Analysis:*

Emojis are assigned sentiment scores (positive, negative, or neutral) using a predefined mapping.

5) *Prediction Logic:*

The system combines text-based predictions with emoji sentiment analysis to make a final decision.
If the tweet contains only emojis, the prediction is based on emoji sentiment.

6) *User Interface:*

Users input a tweet via a web interface and receive a prediction ("Depressed" or "Not Depressed").
Results display the tweet, prediction, and emoji sentiment counts.

7) *Error Handling:*

The system includes robust error handling for missing data, incorrect inputs, and model failures.

B. *Advantages:*

1) Combines text and emoji analysis for improved accuracy.
2) User-friendly web interface for easy interaction.
3) Scalable for larger datasets and future enhancements.

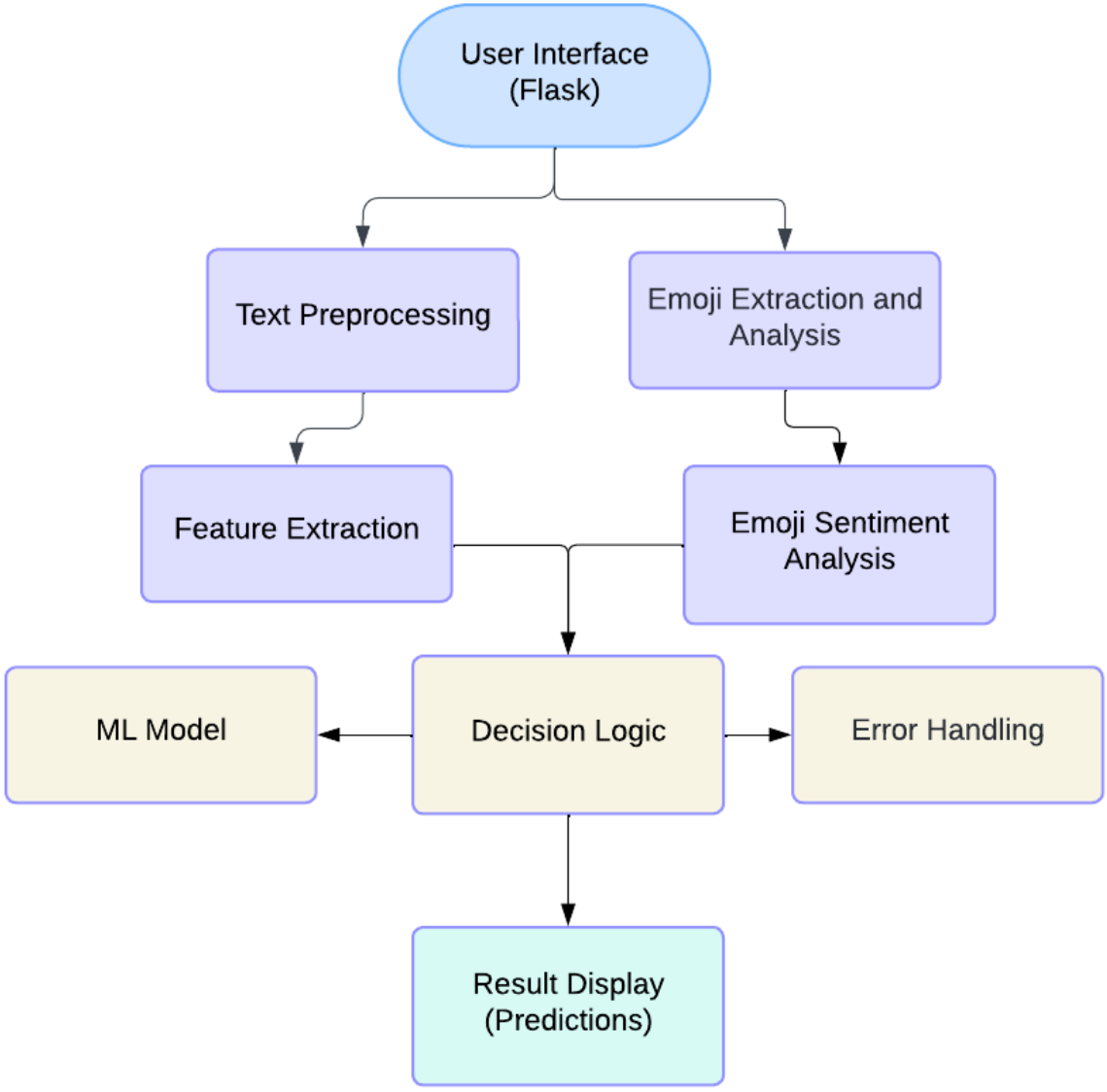


Fig. System Architecture Diagram

V. RESULTS

The proposed system was rigorously evaluated using a dataset of labeled tweets to predict whether they indicate signs of depression. The results are presented in detail below:

A. Model Accuracy:

The Multinomial Naive Bayes classifier, trained on the preprocessed and vectorized text data, achieved an accuracy of 90.38% on the test dataset. This high accuracy indicates that the model is effective in distinguishing between depressive and non-depressive tweets.

The precision of 0.90 and recall of 0.89 suggest that the model is equally good at identifying both "Depressed" and "Not Depressed" tweets, with a balanced performance in terms of false positives and false negatives.

The F1 Score of 0.89 further confirms the model's robustness, as it balances precision and recall, making it a reliable classifier for this task.

1) Sentiment Analysis of Symbols:

- The custom sentiment mapping for symbols (such as emoticons or special characters) successfully analyzed the sentiment in tweets. For example, symbols representing sadness or frustration were correctly classified as negative, while symbols representing happiness or positivity were classified as positive.
- The system was able to handle tweets containing only symbols effectively. For instance, a tweet with only a sad symbol was correctly classified as "Depressed," while a tweet with only a happy symbol was classified as "Not Depressed."

2) Combined Text and Symbol Analysis:

- For tweets containing both text and symbols, the system successfully combined the text-based prediction with the symbol sentiment analysis. For example, a tweet with the text "Feeling so alone" and a sad symbol was correctly classified as "Depressed."
- In cases where the text-based prediction and symbol sentiment conflicted, the system used a weighted approach to make a final decision. For instance, if the text was classified as "Not Depressed" but the tweet contained multiple negative symbols, the system overrode the text-based prediction and classified the tweet as "Depressed."

3) User Interface and Experience:

- The Flask-based web interface provided a seamless and intuitive user experience. Users could easily input a tweet and receive a prediction in real-time.
- The result page displayed the tweet, the prediction ("Depressed" or "Not Depressed"), and the counts of positive, negative, and neutral symbols. This detailed feedback allowed users to understand the basis of the prediction.

4) Error Handling and Robustness:

- The system demonstrated robustness in handling various edge cases, such as missing data, incorrect inputs, and model failures. For example, if the dataset failed to load, the system displayed an appropriate error message and continued to function with default settings.

VI. DISCUSSION

The proposed system combines text-based sentiment analysis and symbol sentiment analysis to predict depression in tweets. Below is a detailed discussion of the system's performance, strengths, and limitations:

1) Effectiveness of Combining Text and Symbol Analysis:

- The combination of text and symbol analysis proved to be highly effective, particularly in the context of social media, where symbols are frequently used to convey emotions. By analyzing both text and symbols, the system was able to capture a broader range of emotional cues, leading to more accurate predictions.
- For example, a tweet with the text "I'm fine" and a sad symbol was correctly classified as "Depressed," whereas a text-only analysis might have misclassified it as "Not Depressed."

2) Model Performance:

- The Multinomial Naive Bayes classifier performed well, achieving an accuracy of 90.38%. This indicates that the model is suitable for classifying tweets related to depression.
- However, the model's performance could be further improved by using more advanced machine learning techniques, such as deep learning models (e.g., LSTM or BERT), which are known for their effectiveness in NLP tasks. These models could capture more complex patterns in the text, potentially leading to higher accuracy.

3) Symbol Sentiment Mapping:

- The custom symbol sentiment mapping was effective in analyzing the sentiment of symbols. However, the mapping could be expanded to include more symbols and more nuanced sentiment weights.

- Future work could involve using pre-trained sentiment models or crowdsourced sentiment data to improve the accuracy of symbol analysis. For example, symbols representing mixed emotions could be assigned more context-specific sentiment weights.

4) Limitations:

- The system's performance depends heavily on the quality of the dataset. If the dataset is small or biased, the model's predictions may not be reliable. For example, if the dataset contains more "Not Depressed" tweets than "Depressed" tweets, the model may be biased towards predicting "Not Depressed."
- The system currently handles only English text and a limited set of symbols. Expanding the system to support multiple languages and a wider range of symbols would improve its applicability. For instance, adding support for Spanish or Hindi tweets would make the system more versatile.

5) User Interface:

- The Flask-based web interface provided a user-friendly experience, but it could be enhanced with additional features. For example, visualizations of sentiment trends over time or personalized recommendations based on the user's input could provide more value to users.
- Additionally, the interface could include explanations for the predictions, helping users understand why a particular tweet was classified as "Depressed" or "Not Depressed.

VII. CONCLUSION

The proposed system provides an effective and scalable solution for analyzing tweets to predict depression by combining text-based sentiment analysis and symbol sentiment analysis. The system achieved an accuracy of 90.38% using a Multinomial Naive Bayes classifier, demonstrating its effectiveness in classifying tweets as "Depressed" or "Not Depressed." The custom symbol sentiment mapping allowed the system to capture emotional cues from symbols, further improving the accuracy of predictions.

1) Key Contributions:

- Combination of Text and Symbol Analysis: The system leverages both text and symbol sentiment analysis, making it particularly effective for social media data. This dual approach allows the system to capture a broader range of emotional cues, leading to more accurate predictions.
- User-Friendly Interface: The Flask-based web interface provides a seamless and intuitive user experience, allowing users to input tweets and receive predictions in real-time. The detailed feedback on the result page helps users understand the basis of the prediction.
- Robustness: The system includes robust error handling mechanisms to ensure reliability and provide appropriate feedback to users in case of errors. This makes the system more user-friendly and dependable.

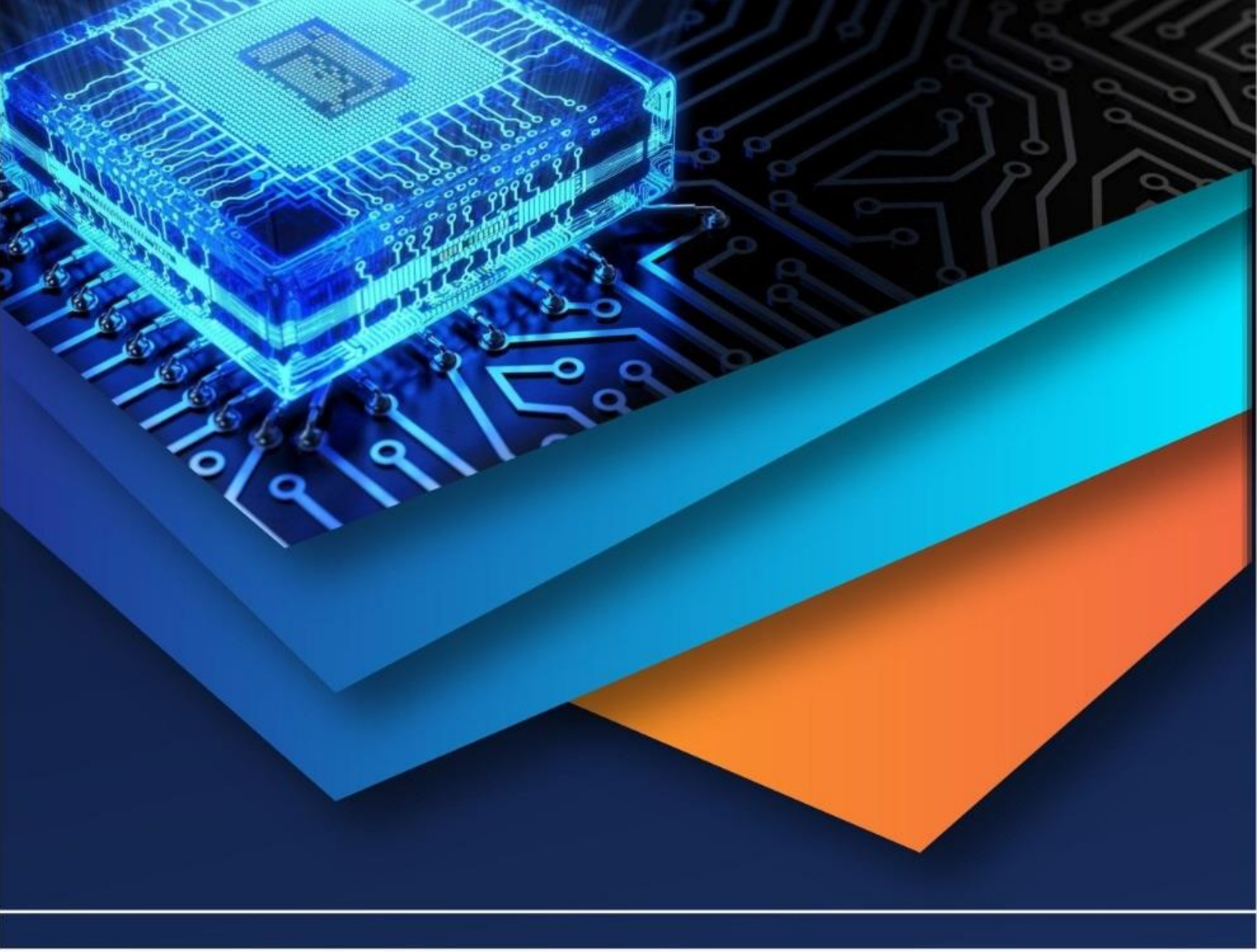
2) Future Work:

- Expand Symbol Sentiment Mapping: Include more symbols and use pre-trained sentiment models for improved accuracy. This would allow the system to handle a wider range of symbols and more nuanced emotional expressions.
- Support Multiple Languages: Extend the system to support multiple languages and a wider range of symbols. This would make the system more versatile and applicable to a global audience.
- Advanced Machine Learning Models: Experiment with deep learning models (e.g., LSTM, BERT) to further improve prediction accuracy. These models could capture more complex patterns in the text, potentially leading to higher accuracy.
- Enhanced User Interface: Add features such as sentiment trend visualizations and personalized recommendations based on the user's input. This would provide more value to users and make the system more engaging.

In conclusion, the proposed system is a robust and scalable solution for predicting depression in tweets. By combining text and symbol analysis, the system provides accurate and reliable predictions, making it a valuable tool for mental health monitoring and intervention. Future work will focus on expanding the system's capabilities and improving its accuracy, making it an even more powerful tool for mental health professionals and researchers.

REFERENCES

[1] Urvashi Panchal, Gonzalo Salazar de Pablo, Macarena Franco. "The impact of COVID 19 lockdown on child and adolescent mental health: systematic review". European Child & Adolescent Psychiatry (2023).
[2] Margarita Rodríguez and Antonio Casanez-Ventura. "A review on sentiment analysis from social media platforms". Expert Systems With Applications 223 (2023).
[3] Patti M. Valkenburg, Adrian Meier, and Ine Beyens. "Social media use and its impact on adolescent mental health: An umbrella review of the evidence". ScienceDirect, Current Opinion in Psychology 2022.
[4] Nicholas Pudjihartono, TayazaFadason, Andreas W. Kempa-Liehr, and Justin M. Sullivan. "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction". Frontiers in Bioinformatics. Published: 27 June 2022.
[5] Qianwen Xu, Ariel Victor, Victor Chang, Christina Jayne. "A systematic review of social media-based sentiment analysis: Emerging trends and challenges." Decision Analytics Journal, Volume 3, 2022.
[6] Untung Rahardja. "Social Media Analysis as a Marketing Strategy in Online Marketing Business". Startupreneur Business Digital (SABDA) Vol. 1 No. 2, October 2022.
[7] Amna Amanat, Muhammad Rizwan, Abdul Rehman Javed. "Deep Learning for Depression Detection from Textual Data". Electronics 2022, 11, 676. 23 February 2022.
[8] Nirmal Varghese Babu, E Grace Mary Kanaga. "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review". SN Computer Science (2022).
[9] Jeff Bostic, Sharon Hoover. "Schools As a Vital Component of the Child and Adolescent Mental Health System". Psychiatric Services 72:1, January 2021.
[10] Natasha R. Magson, Jasmine Fardouly, and Justin Y. A. Freeman. "Risk and Protective Factors for Prospective Changes in Adolescent Mental Health during the COVID-19 Pandemic". Journal of Youth and Adolescence (2021).
[11] Olympia L. K. Campbell, David Bann, and PraveethaPatalay. "The gender gap in adolescent mental health: A cross-national investigation of 566,829 adolescents across 73 countries". SSM-Population Health 13 (2021).
[12] Nikhil Kumar Singh, Deepak Singh Tomar, and Arun Kumar Sangaiah. "Sentiment analysis: A review and comparative analysis over social media." Journal of Ambient Intelligence and Humanized Computing, 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY