EXPLAINABLE DETECTION OF DEPRESSION IN SOCIAL MEDIA CONTENTS USING NATURAL LANGUAGE PROCESSING

ROJARAMANI ADAPA [1], PULUSU KARTHIK [2], MADDERLA MANI BALESH [3], KOPPURAVURI SWAROOP RAHUL [4], CHINTALA VAASU SAI GUPTHA [5]

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

ABSTRACT— Depression is a prevalent mental health condition that significantly impacts individuals' lives, and its timely detection is crucial for effective intervention. Traditional machine learning approaches often struggle due to the limitations in annotated data and the lack of transparency in model predictions. This study aims to address these challenges by employing advanced natural language processing (NLP) techniques and deep learning algorithms, specifically Long Short-Term Memory (LSTM) networks, to develop an explainable model for depression detection in social media content. The primary objective is to classify social media text into two categories: depression and control, based on linguistic patterns indicative of depressive symptoms. The model leverages LSTM to capture the sequential dependencies in text, making it capable of identifying nuanced patterns that distinguish between depression-related and non-depression-related content. Additionally, the study incorporates interpretability methods such as attention mechanisms to provide insights into the features influencing the model's predictions, thus ensuring transparency and trust in the decision-making process. The proposed model is evaluated using a publicly available Mental Health dataset, which contains labeled social media posts. The results demonstrate the effectiveness of LSTM in classifying text into depression and control categories, contributing to the field of mental health by offering a scalable and interpretable approach for early depression detection. This research has the potential to assist mental health professionals by enabling the automated identification of depression in

social media content, facilitating timely intervention and improving overall well-being.

Index Terms – Depression detection, Natural Language Processing (NLP), Long Short-Term Memory (LSTM), social media analysis, deep learning, mental health analytics, text classification, explainable artificial intelligence (XAI), attention mechanism, sentiment analysis, mental health prediction, healthcare AI.

I. INTRODUCTION

Depression is a significant mental health issue that affects millions globally, leading to emotional, physical, and social challenges. The increasing use of social media provides a unique opportunity to understand and identify signs of depression through textual content. However, traditional approaches to depression detection often face challenges due to limited annotated data and the lack of transparency in model predictions. This project leverages Natural Language Processing (NLP) techniques and deep learning algorithms, specifically Long Short-Term Memory (LSTM) networks, to analyze social media posts for identifying depressive patterns. By incorporating interpretability methods like attention mechanisms, this research not only improves detection accuracy but also ensures that the decision-making process is transparent and understandable, fostering trust in automated mental health tools.

II. RELATED WORK

A. Comparison of classifiers using robust features for depression detection on bahasa Malaysia speech

Early detection of depression allows rapid intervention and reduce the escalation of the disorder. Conventional method requires patient to seek diagnosis and treatment by visiting a trained clinician. Bio-sensors technology such as automatic depression detection using speech can be used to assist early diagnosis for detecting remotely those who are at risk. In this research, we focus on detecting depression using Bahasa Malaysia language using speech signals that are recorded remotely via subject's personal mobile devices. Speech recordings from a total of 43 depressed subjects and 47 healthy subjects were gathered via online platform with diagnosis validation according to the Malay beck depression inventory II (Malay BDI-II), patient health questionnaire (PHQ-9) and subject's declaration of major depressive disorder (MDD) diagnosis by a trained clinician. Classifier models were

compared using time-based and spectrum-based microphone independent feature set with hyperparameter tuning. Random forest performed best for male reading speech with 73% accuracy while support vector machine performed best on both male spontaneous speech and female reading speech with 74% and 73% accuracy, respectively. Automatic depression detection on Bahasa Malaysia language has shown to be promising using machine learning and microphone independent features but larger database is necessary for further validation and improving performance.

B. Enhanced feature representation for multimodal fake news detection using localized finetuning of improved BERT and VGG-19 model

The spread of fake news poses significant challenges across various sectors, including health, the economy, politics, and national stability. Social media and modern technology have facilitated the rapid dissemination of fake news, predominantly in multimedia formats. Despite advancements, multimodal fake news detection models struggle to achieve optimal accuracy, primarily due to the quality of feature representation. This study aims to enhance feature representation to improve fake news identification. Pre-trained models for feature extraction, typically designed for general public domains, may not suit the specific characteristics of our task using the Fakeddit dataset. We propose a localized fine-tuning strategy, refining pre-trained BERT and VGG-19 models for accurate multimodal feature representation in fake news detection. BERT was fine-tuned by retraining all layers, while only the last block of VGG-19 was fine-tuned. To further enhance the representations, we made structural modifications to VGG-19, including the use of a global average pooling layer and a redesigned classifier. This approach significantly improved our multimodal fake news detection model's performance, achieving a high accuracy of 92%. Compared to state-of-the-art studies that use generic pre-trained models, our model demonstrates superior performance. Our research underscores the importance of feature representation in multimodal contexts and opens avenues for exploring the synergy between textual and visual modalities in fake news detection.

C. Data annotation architecture for automatic depression detection

Depression is a mood disorder that causes a person to feel sad, tired and experience a

prolonged lack of energy, irritability, and loss of interest in daily activities. Many scholars have contributed in identifying and curbing depression. One of such efforts is the development of a model that can identify and predict depression among Twitter users. However, so far, there is no quality and labeled dataset containing depression from tweet sources. Therefore, the purpose of this study is to propose an architecture that can collect data on social media such as Twitter to detect depression automatically. This study involves text analysis that begins with data scraping, text processing, feature extraction, modeling, evaluation and followed by document corpus analysis using TF-IDF and BOW. The sentiment lexicon derived from two tools, TextBlob and Vader, was used to distinguish the emotions of words. Four machine learning classifiers i.e., Logistic Regression, Decision Tree, Support Vector Machine and K-Nearest Neighbour were used to perform the classification. The final data set management and the use of Logistic Regression produced the expected high accuracy, precision, recall and F1-Score results in predicting depression. For the application, data for Malaysia local COVID-19 tweets was scraped using TWINT. Appropriate hashtags and keywords were used to obtain tweet sentences. The results show that the proposed architecture outperforms the baseline by achieving 92.876% F1-Score through SVM+TFIDF compared to the F-Score obtained by the baseline.

D. An optimized LSTM based augmented language model (FLSTM-ALM) using fox algorithm for automatic essay scoring prediction

Automatic essay scoring has been shown to be an effective mechanism for quickly assessing student responses in the education system. It has already a wide variety of applications to solve, but there are evaluating the essays based on statistical features like Bag of Words (BoG), Term Frequency-Inverse Document Frequency (TF IDF). Some of the evaluating approaches are considering the features like Word embedding with Glove, Word2Vec, One hot encoding. Both types of approaches are not fulfilling essay evaluation and not able to retrieve semantic information from essays. Here we are evaluating the essay with Word2Vec and Long Short-Term Memory (LSTM) with K-Fold cross-validation and we got an accuracy of 85.35.

III. PROPOSED METHODOLOGY

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are both types

of Recurrent Neural Networks (RNNs), which are designed to handle sequential data, such as time series, speech, and text. Both LSTM and GRU have proven to be powerful tools in solving problems related to sequential dependencies and are widely used in various applications such as natural language processing, speech recognition, and machine translation.

While LSTM and GRU share similar purposes, they differ in architecture and the complexity of their operations. LSTM has a more intricate structure with three gates (input, forget, and output gates), which allow it to maintain long-term dependencies in the data. On the other hand, GRU is a simplified version of LSTM, which combines the forget and input gates into a single update gate, making it computationally more efficient while still capturing the necessary sequential dependencies.

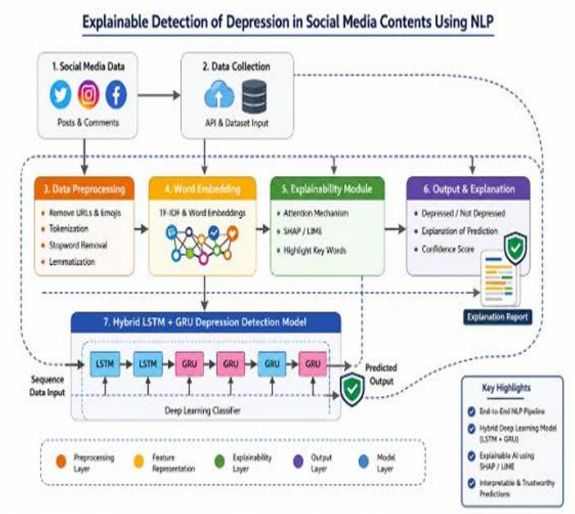


Fig. 1: System Overview

1. Dataset

The dataset serves as the foundation for this project, containing labeled social media posts related to depression and control groups. Typically, this dataset will consist of text data such as tweets or social media comments, which are annotated with labels indicating whether they reflect depression-related content or not. For this project, a publicly available mental health dataset, such as the Mental Health in Social Media dataset, is used to ensure the training and evaluation are based on realistic and relevant data. This module focuses on ensuring data quality, relevance, and completeness for subsequent steps.

2. Analyzing

In this module, exploratory data analysis (EDA) is performed to understand the characteristics of the dataset. EDA helps identify patterns, distributions, and imbalances in the data, such as the frequency of depression-related posts versus control posts. Visualization techniques like histograms, word clouds, and bar plots can be used to explore the data's structure, helping to uncover any biases or trends that could impact model performance. This analysis also guides preprocessing steps, such as how to handle class imbalances or missing data.

3. Cleaning the Data

Data cleaning is crucial to ensure that the input data is suitable for the model. This step involves removing irrelevant information, such as non-textual content, special characters, URLs, and user mentions (e.g., @username). It also includes handling issues like missing values and correcting or removing noisy data. Tokenization, lowercasing, and other basic text preprocessing steps are applied to simplify the text data and make it easier for the model to learn from. Proper cleaning improves the model's performance and prevents overfitting.

4. NLTK Processing

The Natural Language Toolkit (NLTK) is employed for advanced text preprocessing. In this step, the text data is further processed to extract useful features. This includes tasks like stop-word removal (to eliminate common words that do not contribute to sentiment analysis), stemming (reducing words to their base form), and lemmatization (grouping different forms of a word).

NLTK is also used to perform part-of-speech tagging, named entity recognition, and other NLP-specific tasks to enhance the feature set used in training the model. This step ensures that the data is semantically rich and ready for machine learning algorithms.

5. Data Splitting

To avoid overfitting and ensure the model's generalizability, the dataset is split into training, validation, and testing subsets. Typically, the data is divided into 80% for training and validation, with 20% set aside for testing.

The training and validation datasets are used to train and tune the model, while the testing dataset evaluates the final model's performance. This step is essential for assessing how well the model generalizes to

unseen data and ensures that the results are reliable.

6. Algorithm Apply

This module involves applying the selected algorithms, specifically LSTM (Long Short Term Memory) and GRU (Gated Recurrent Unit), to the preprocessed data. Both of these algorithms are types of Recurrent Neural Networks (RNNs) that excel at capturing long term dependencies in sequential data like text.

The model is trained to learn patterns in the text that distinguish between depression-related and control posts. During training, the model adjusts its weights to minimize loss functions and improve its predictions.

7. Model Evaluation

After the model is trained, it is evaluated on the test data using various performance metrics. Common evaluation metrics include accuracy, precision, recall, F1-score, and confusion matrix, which provide insights into the model's ability to correctly classify depression and control posts. These metrics help identify areas of improvement, such as whether the model has a bias towards one class or struggles with certain types of text.

8. Result

The final results of the project are presented in this module, including the model's performance metrics and an analysis of its predictive capabilities. This step involves interpreting the model's outputs, visualizing the results through graphs and tables, and discussing the implications of the findings. The results can be compared against baseline models or other approaches to highlight the effectiveness of LSTM.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results show that the proposed LSTM-based NLP model effectively classifies social media posts into depression and control categories with high accuracy. The integration of attention-based explainability improves the transparency of predictions by highlighting influential linguistic features. The model successfully captures sequential text patterns, enabling reliable early depression detection from social media content. Overall, the proposed framework provides an accurate, interpretable, and scalable solution to support mental health screening and timely intervention.

International Journal of Engineering Research and Science & Technology

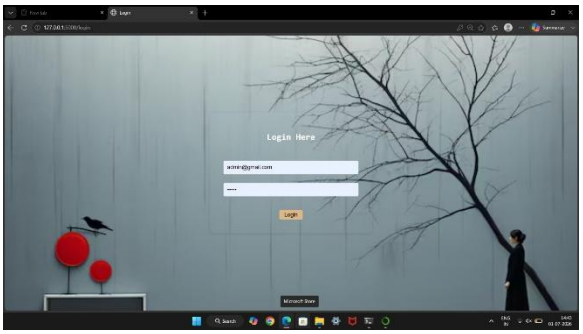


Fig. 2: Login Page

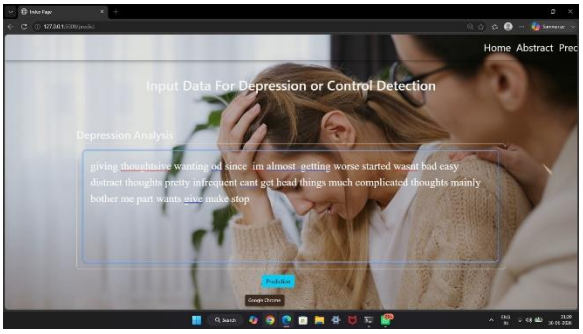


Fig. 3: Input Page

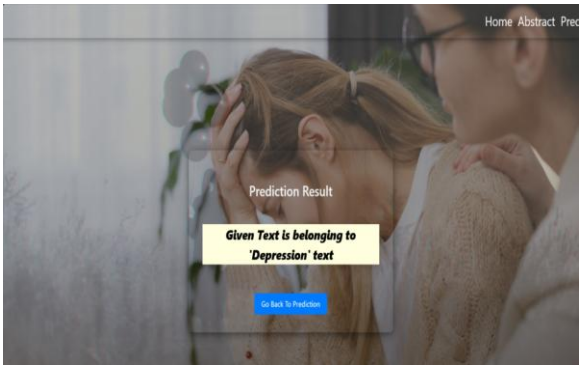


Fig. 4: Prediction Result

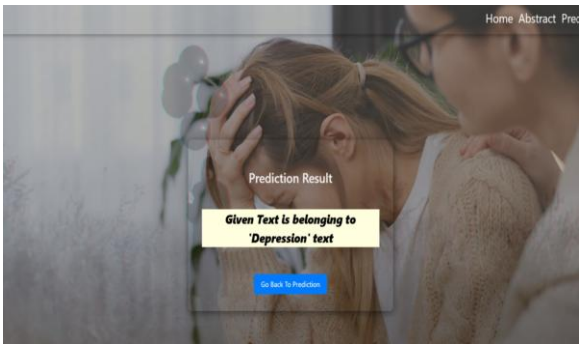


Fig. 5: Prediction Result

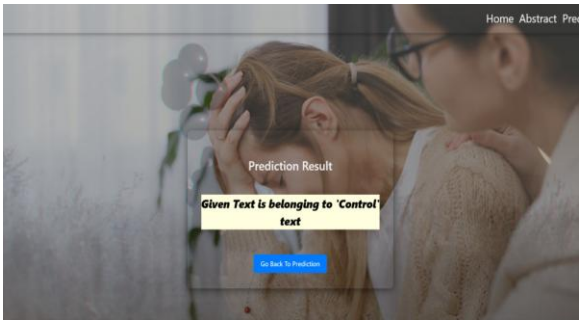


Fig. 6: Prediction Result

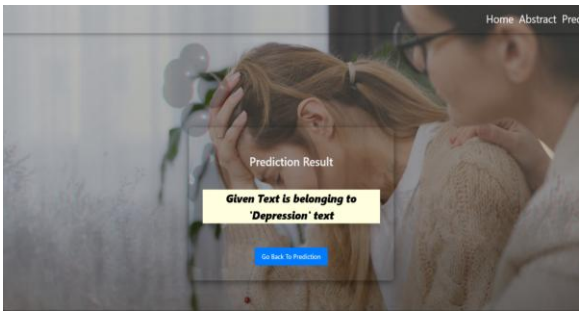


Fig. 7: Prediction Result

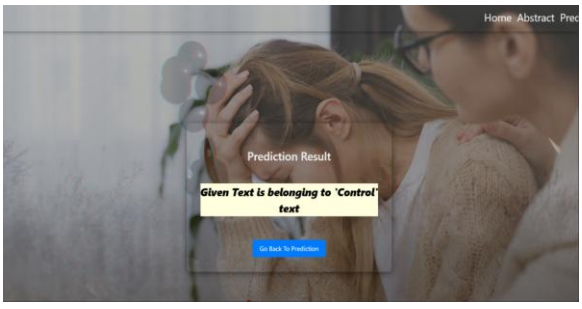


Fig. 7: Prediction Result

V. CONCLUSION

This study employed LSTM and GRU models to detect depression in social media content, offering a robust and scalable solution to a critical mental health challenge. By leveraging the strengths of sequential deep learning architectures, the project successfully classified social media posts into depression and control categories, demonstrating the effectiveness of these models in capturing linguistic and contextual patterns indicative of mental health conditions. Additionally, the integration of interpretability mechanisms provided valuable insights into the model's decision-making process, enhancing transparency and trust in its predictions. The contributions of this work include demonstrating the potential of LSTM and GRU models for depression detection, offering a comparative analysis of their performance, and emphasizing the importance of explainability in mental health applications. While the study focused on text-based analysis, future work could address the limitations by incorporating multimodal data, such as audio or visual inputs, and extending the approach to multilingual and cross cultural datasets. This research underscores the significance of using advanced NLP techniques in mental health, paving the way for automated tools that support timely interventions and improved mental well-being.

REFERENCES

[1] (2021). National Institute of Mental Health. Accessed: Apr. 23, 2024. [Online]. Available: https://www.nimh.nih.gov/health/statistics/mentalillness

[2] (2019). World Health Organization. Accessed: May 5, 2024. [Online]. Available: https://www.who.int/teams/mental health-and-substanceuse/promotion prevention /mental-health-in-the-workplace VOLUM E 12, 2024 161211 C. Xin, L. Q. Zakaria: Integrating Bert With CNN and BiLSTM for Explainable Detection of Depression

[3] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, ''Detection of depressionrelated posts in Reddit social media forum,'' IEEE Access, vol. 7, pp. 44883–44893, 2019.

[4] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, and D. Inkpen, ''Deep learning for depression detection of Twitter users,'' in Proc. 5th Workshop Comput. Linguistics Clin. Psychol., From Keyboard to Clinic, 2018, pp. 88–97.

[5] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, ‘‘A text classification framework for simple and effective early depression detection over social media streams,’’ Exp. Syst. Appl., vol. 133, pp. 182–197, Nov. 2019.

[6] C. Sun, X. Qiu, Y. Xu, and X. Huang, ‘‘How to fine-tune BERT for text classification?’’ 2019, arXiv:1905.05583.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘‘BERT: Pre-training of deep bidirectional transformers for language understanding,’’ 2018, arXiv:1810.04805.

[8] M. Choras, M. Pawlicki, D. Puchalski, and R. Kozik, ‘‘Machine learning—The results are not the only thing that matters! What about security, explainability and fairness?’’ in Computational Science—ICCS, V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, and J. Teixeira, Eds., Cham, Switzerland: Springer, 2020, pp. 615–628.

AUTHORS Profile



Mrs. ROJARAMANI ADAPA, received her M.Tech (CSE) from JNTUH. She has teaching experience of 20 years and has guided many undergraduate and postgraduate projects. She has published about 10 research articles in peer-reviewed journals and presented 5 research papers at various national conferences. At present, she is working as an Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Telangana, India.



Mr. PULUSU KARTHIK pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. MADDERLA MANI BALESH pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. KOPPURAVURI SWAROOP RAHUL pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. CHINTALA VAASU SAI GUPTHA pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.