AN EXPLAINABLE TYPE OF DEPRESSION CLASSIFICATION USING DEEP LEARNING

Lu Yinuo
Dr.Lailatul Qadri Zakaria

Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor

ABSTRACT

Depression is a widespread mental health condition that requires early and accurate detection to support timely intervention. This project presents an AI-driven depression classification system that analyses user-submitted text using natural language processing (NLP) and deep learning techniques. The system is designed to detect depressive expressions and classify them into three major types: major depressive disorder (MDD), bipolar disorder, and postpartum depression (PPD). To address the challenge of interpretability and reliability in automated mental health tools, the project integrates two models: a binary classifier for depression detection and a multiclass classifier for subtype classification. The proposed solution employs the DistilBERT model and incorporates an explainable AI (XAI) module that highlights emotionally significant keywords contributing to the prediction. A user-friendly web interface built with Streamlit allows real-time interaction and visualization of results. The system development involved dataset preprocessing, model training and evaluation using standard metrics such as accuracy, precision, recall, and F1-score. Usability testing was conducted with non-technical participants to assess clarity and ease of use. Final results demonstrate strong classification performance (above 94% accuracy) and effective keyword-based explanation. Compared to MentalBERT, the DistilBERT model achieved better results in both binary and multiclass tasks. The project delivers a practical, interpretable, and efficient tool for depression screening, suitable for both individual users and potential clinical adaptation.

Keywords**:** DistilBERT, Depression

INTRODUCTION

Depression is one of the most prevalent and serious mental health disorders globally, affecting millions of individuals across different age groups and backgrounds. According to the World Health Organization (2023), major depressive disorder (MDD) affects approximately 3.8% of the global population, while bipolar disorder and postpartum depression (PPD) impact specific subpopulations with equally severe consequences. These disorders often manifest through subtle symptoms such as emotional instability, lack of motivation, and social withdrawal, making early diagnosis a significant challenge. With the rise of social media platforms such as Twitter and Reddit, users frequently express their thoughts and emotional states online, providing an untapped data source for identifying signs of depression in real time. This creates an opportunity to utilize textual data for automated mental health screening.

Traditional diagnostic methods such as clinical interviews and psychological questionnaires heavily rely on self-reported responses and professional interpretation, which can lead to delays in diagnosis, subjective bias, or underreporting. These limitations are particularly evident when attempting to distinguish among specific types of depression, as overlapping symptoms often make classification difficult. In light of these challenges, this study proposes the development of an AI-driven depression classification system that utilizes natural language processing (NLP) and deep learning technologies. The system is designed to detect the presence of depression in user-submitted text and further classify it into one of three common subtypes: major depressive disorder, bipolar disorder, or postpartum depression.

The primary objective of this project is to design and implement a web-based application that analyses user-generated textual input to identify and classify depressive expressions. By incorporating deep learning techniques such as DistilBERT and explainable AI (XAI), the system not only improves detection accuracy but also enhances interpretability by highlighting emotionally significant words contributing to the classification. This approach provides an accessible, privacy-preserving tool for early mental health assessment, especially for individuals who may hesitate to seek professional help directly.

LITERATURE REVIEW

Overview of Depression and Its Types

Depression is a major public health concern that affects individuals across all demographics. Among the many types of depression, this study focuses on three common forms: Major Depressive Disorder (MDD), Bipolar Disorder, and Postpartum Depression (PPD). Major Depressive Disorder is characterized by persistent sadness, fatigue, and loss of interest, severely impacting a person's ability to function. Bipolar Disorder is marked by alternating periods of depression and mania, which can include impulsive behavior and emotional extremes. Postpartum Depression typically arises after childbirth and may involve exhaustion, detachment, and feelings of worthlessness, often impairing mother-child bonding. These disorders require distinct treatment approaches, which underscores the importance of early and accurate classification.

Traditional vs AI-Based Detection Methods

Traditional diagnostic tools such as the Beck Depression Inventory (BDI) and PHQ-9 rely on self-reported responses, making them susceptible to subjectivity, response bias, and social desirability effects. Individuals may underreport symptoms due to stigma, lack of self-awareness, or cultural influences, leading to delayed or inaccurate diagnosis. Moreover, these tools primarily assess general depressive symptoms and often lack the specificity required to distinguish between subtypes such as Major Depressive Disorder, Bipolar Disorder, and Postpartum Depression, which differ significantly in treatment needs and clinical presentation.

Artificial intelligence (AI), particularly in natural language processing (NLP), has emerged as a promising alternative. Machine learning methods such as Support Vector Machines (SVM), Random Forests, and Logistic Regression have been used to classify depression based on linguistic and behavioural features. For example, Maheshwar et al. (2023) demonstrated that SVM combined with MFCC features effectively detects emotional variations in speech data. Haque et al. (2021) applied Random Forest with Boruta feature selection to achieve over 95% accuracy in detecting depression in children. Similarly, Deb et al. (2023) showed that logistic regression can be used to predict depressive symptoms based on demographic and social factors.

Deep learning models go a step further by automatically learning complex representations from raw text data. Bhuvaneswari and Prabha (2023) proposed a hybrid deep learning model combining CNN and LSTM with attention mechanisms to detect depressive tendencies from tweets, achieving impressive accuracy. Transformer-based models like BERT and DistilBERT have proven even more powerful in text classification tasks. DistilBERT, being a lighter and faster version of BERT, retains most of its performance while being more suitable for real-time applications (Temidayo Omoniyi, 2024).

Importance of Explainable AI (XAI)

While deep learning models offer high accuracy, their opaque nature poses a challenge in sensitive domains such as mental health. Explainable AI (XAI) addresses this issue by providing transparency into the model's decision-making process. Techniques such as attention weight visualization and SHAP (Shapley Additive Explanations) highlight which words or features contributed most to the prediction. Arya et al. (2020) emphasized that interpretability not only builds user trust but also ensures ethical deployment of AI in clinical settings. In this study, attention mechanisms were employed to visualize emotionally charged keywords such as “hopeless” or “postpartum,” giving users insight into how the system arrived at its classification.

Related Works and Research Gap

Many recent studies have explored the application of AI in mental health. For instance, Mansoor and Ansari (2024) used a multimodal deep learning model integrating BERT, LSTM, and time-series analysis to detect psychological crises from Facebook posts. Zafar et al. (2024) emphasized the importance of multimodal input—combining linguistic, behavioural, and physiological data—to improve depression detection. Shimada (2023) highlighted AI's role in personalized treatment and suicide prevention through emotion and behaviour tracking. Other works, such as Ballesteros et al. (2024), applied CNNs for facial emotion recognition, while Walschots et al. (2024) explored wearable devices for continuous monitoring of depressive symptoms.

Despite these advancements, most studies either focus on general depression detection or rely on black-box models without providing interpretability. Moreover, few systems classify depression into distinct subtypes. This study fills that gap by implementing a web-based system that detects and classifies three type of depression using DistilBERT, while incorporating explainable AI features for transparency and user trust.

METHODOLOGY

This section describes the complete process of data preparation, model development, evaluation, and system implementation for depression type classification based on social media text input. The methodology follows six major stages: data collection, preprocessing, feature extraction, model training, model evaluation, and system deployment with explainable AI integration.

Data Collection

Two publicly available datasets were used in this study: the Reddit Self-reported Depression Dataset and the Twitter Depression Dataset. The Reddit dataset was labeled for Major Depressive Disorder (MDD), Bipolar Disorder, and Postpartum Depression (PPD), while the Twitter dataset consisted of binary-labeled depression and non-depression posts. Combined, the datasets provided sufficient diversity for both binary and multiclass classification tasks.

Data Preprocessing

Text cleaning was performed separately for each dataset. This included removing URLs, punctuation, emojis, repetitive characters, and converting all text to lowercase. Reddit comments were filtered based on subreddit relevance, while Twitter posts were selected using keyword-based filtering and annotated for depression relevance. Preprocessing was essential to ensure clean input for deep learning models.

Feature Extraction

The tokenization and embedding were performed using DistilBERT, a lighter version of BERT from HuggingFace Transformers. Each post was tokenized into subword units using the DistilBertTokenizer, and the resulting embeddings were used as input features for model training. Maximum sequence length was capped at 128 tokens to optimize memory and speed.

Model Training

Two separate models were fine-tuned:

- Binary Classification Model: Trained to distinguish between depressed and non-depressed text.
- Multiclass Classification Model: Trained to classify among MDD, Bipolar Disorder, and PPD.

Both models were fine-tuned using the DistilBERTForSequenceClassification architecture in TensorFlow. The training utilized the Adam optimizer (learning rate = 5e-5) and cross-entropy loss. Early stopping was applied with a patience of 3 epochs to prevent overfitting. Figure 1 shows the architecture flowchart of the training pipeline, from tokenization to prediction.

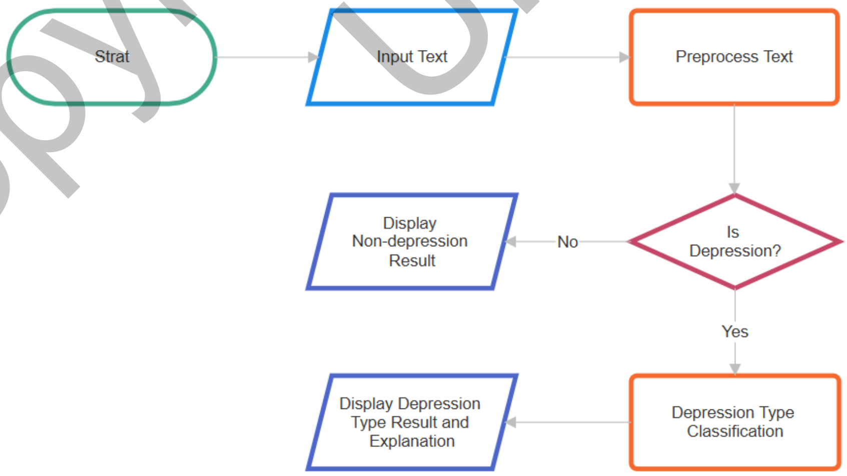


Figure 1: Model Training Pipeline Flowchart

Model Evaluation

The models were evaluated using four metrics: accuracy, precision, recall, and F1-score. Confusion matrices were generated for both binary and multiclass models to visualize classification performance.

System Implementation

A web application was developed using Streamlit to allow users to input free-form text and receive predictions on their depression status. The system included:

1. A search bar for symptom-based input.
2. A backend model that loads the trained DistilBERT weights.
3. Real-time output of the predicted depression class.
4. Integration of Explainable AI (XAI) components, where attention heatmaps highlight influential words contributing to the prediction.

Explainable AI Visualization

To improve transparency, the system visualizes attention weights extracted from the final Transformer layer. Words contributing most to the classification are highlighted in color (e.g., red for high influence). This helps users and clinicians interpret the system's decision and increases trustworthiness.

Tools and Frameworks

The following tools were used:

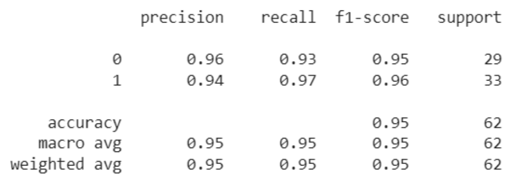
1. HuggingFace Transformers – for pretrained DistilBERT
2. TensorFlow – for model fine-tuning and training
3. Streamlit – for web application
4. scikit-learn – for evaluation metrics
5. Matplotlib / Seaborn – for confusion matrix visualization

RESULT

This section presents the performance of the binary and multiclass classification models, the effectiveness of the explainable AI integration, and the functionality of the deployed web application.

Binary Classification Results

The binary classification model was designed to distinguish between depressed and non-depressed text samples. As shown in Figure 2, the model achieved a high accuracy of 95%, with precision, recall, and F1-score all at 95%, indicating a well-balanced performance across both classes.



```
              precision    recall  f1-score   support

           0       0.96      0.93      0.95        29
           1       0.94      0.97      0.96        33

    accuracy                           0.95        62
   macro avg       0.95      0.95      0.95        62
weighted avg       0.95      0.95      0.95        62
```

Figure 2: Classification Report for Binary Class

The corresponding confusion matrix in Figure 3 further supports these results, where the model correctly identified 27 out of 29 non-depressed and 32 out of 33 depressed instances, with only 3 misclassifications in total.

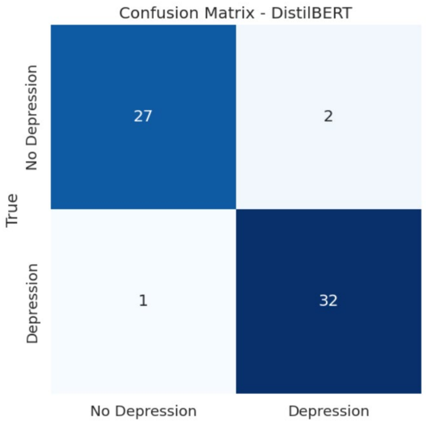
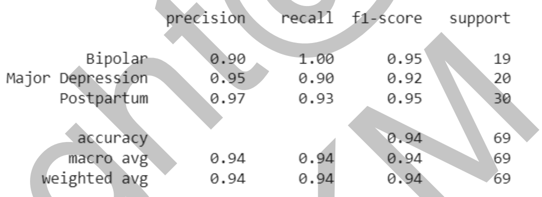


Figure 3: Confusion Matrix for Binary Class

Multiclass Classification Results

The multiclass model extended the classification task to distinguish among Bipolar Disorder, Major Depression, and Postpartum Depression. As shown in Figure 4, the model maintained a high overall accuracy of 94%, with F1-scores ranging from 0.92 (Major Depression) to 0.95 (Bipolar and Postpartum). The model achieved the highest precision on Postpartum Depression (0.97) and perfect recall for Bipolar Disorder (1.00).



```
                  precision    recall  f1-score   support

         Bipolar       0.90      1.00      0.95        19
Major Depression       0.95      0.90      0.92        20
      Postpartum       0.97      0.93      0.95        30

        accuracy                           0.94        69
       macro avg       0.94      0.94      0.94        69
    weighted avg       0.94      0.94      0.94        69
```

Figure 4: Classification Report for Multiclass

The confusion matrix in Figure 5 illustrates that most misclassifications occurred between Major Depression and the other two classes. Nonetheless, the model correctly predicted 28 out of 30 Postpartum samples and 19 out of 19 Bipolar cases, demonstrating its effectiveness in distinguishing between the three subtypes.

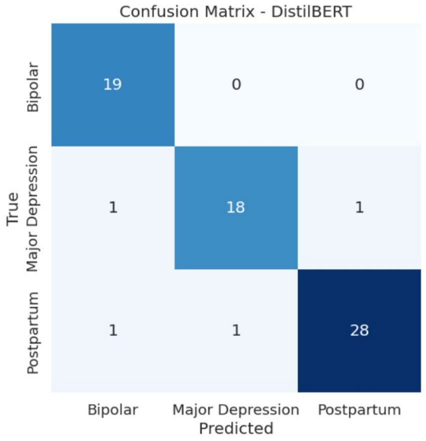
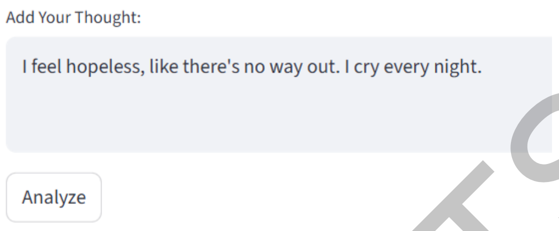


Figure 5: Confusion Matrix for Multiclass

Explainable AI Results

To enhance model transparency, an attention-based visualization mechanism was implemented to show the impact of each word on the classification decision. As demonstrated in Figures 6 and 7, when a user inputs the statement "I feel hopeless, like there's no way out. I cry every night.", the system identifies and highlights emotionally charged terms such as "hopeless," "way," and "cry" in shades of red, indicating a strong association with depressive sentiment.



Add Your Thought:

I feel hopeless, like there's no way out. I cry every night.

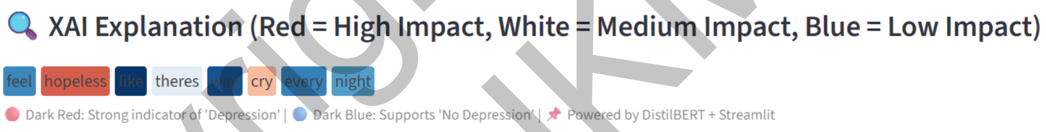
Analyze

Figure 6: Example User Input in Web Interface

The color-coded explanation helps users understand how the model reaches its prediction, where:

1. Dark Red indicates a strong signal contributing to a 'Depression' classification.
2. White represents moderate impact.
3. Blue (e.g., "feel", "every", "night") signals low contribution or even neutral/non-depressive context.

This explainability layer not only supports interpretability but also builds user trust, especially in sensitive mental health applications.



🔍 XAI Explanation (Red = High Impact, White = Medium Impact, Blue = Low Impact)

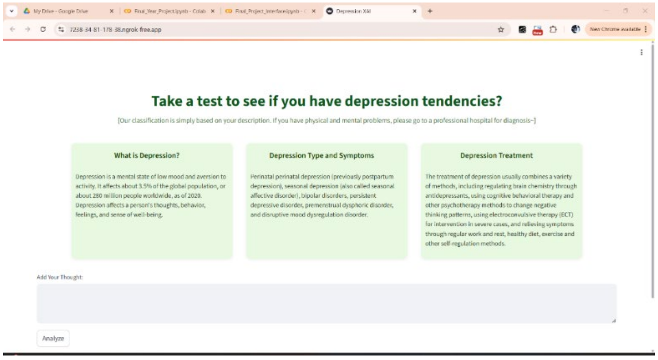
feel hopeless like theres way cry every night

● Dark Red: Strong indicator of 'Depression' | ● Dark Blue: Supports 'No Depression' | 📌 Powered by DistilBERT + Streamlit

Figure 7: Word-Level Attention Visualization with Color Impact Explanation

System Interface and Functionality

The web-based depression detection system was developed using Streamlit and provides a user-friendly interface for real-time interaction. As illustrated in Figure 8, the homepage introduces basic mental health concepts, depression types, and potential treatments. Users can enter a brief textual description of their emotional state in the input box and click the Analyze button to initiate prediction.



My Drive - Google Drive
Final_Year_Project.ipynb - Colab
Final_Project_Interface.ipynb - C
Depression XAI
7238-34-81-178-38.ngrok-free.app
New Chrome available

Take a test to see if you have depression tendencies?

[Our classification is simply based on your description. If you have physical and mental problems, please go to a professional hospital for diagnosis~]

What is Depression?

Depression is a mental state of low mood and aversion to activity. It affects about 3.5% of the global population, or about 280 million people worldwide, as of 2020. Depression affects a person's thoughts, behavior, feelings, and sense of well-being.

Depression Type and Symptoms

Perinatal perinatal depression (previously postpartum depression), seasonal depression (also called seasonal affective disorder), bipolar disorders, persistent depressive disorder, premenstrual dysphoric disorder, and disruptive mood dysregulation disorder.

Depression Treatment

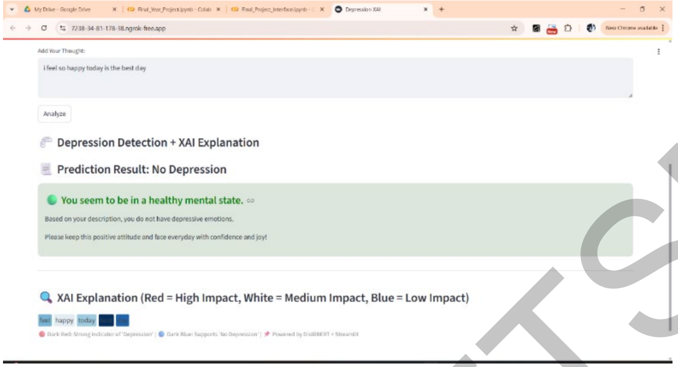
The treatment of depression usually combines a variety of methods, including regulating brain chemistry through antidepressants, using cognitive behavioral therapy and other psychotherapy methods to change negative thinking patterns, using electroconvulsive therapy (ECT) for intervention in severe cases, and relieving symptoms through regular work and rest, healthy diet, exercise and other self-regulation methods.

Add Your Thought:

Analyze

Figure 8: System Interface - Streamlit Web UI

Upon submission, the system immediately displays the analysis result. For users with neutral or positive expressions, the interface delivers a reassuring message indicating no signs of depression (Figure 9). This output is enhanced with an Explainable AI (XAI) visualization that highlights emotionally relevant words, color-coded by impact strength.


My Drive - Google Drive
Final_Year_Project.ipynb - Colab
Final_Project_Interface.ipynb
Depression XAI
7238-34-81-178-38.ngrok-free.app
New Chrome available
Analyze
feel
happy
today

Add Your Thought:

i feel so happy today is the best day

Depression Detection + XAI Explanation

Prediction Result: No Depression

● You seem to be in a healthy mental state.

Based on your description, you do not have depressive emotions.

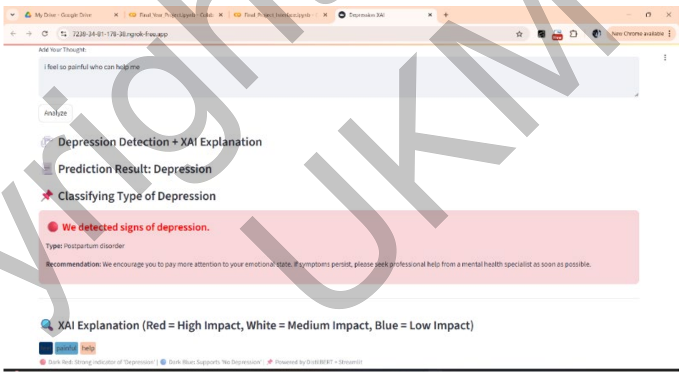
Please keep this positive attitude and face everyday with confidence and joy!

XAI Explanation (Red = High Impact, White = Medium Impact, Blue = Low Impact)

● Dark Red: Strong indicator of 'Depression' | ● Dark Blue: Supports 'No Depression' | 📌 Powered by DistilBERT + Streamlit

Figure 9: No Depressive Emotions Display Interface

If the input indicates depressive tendencies, the system not only flags the presence of depression but also attempts to classify it into one of the predefined categories: Major Depression, Bipolar Disorder, or Postpartum Depression. As seen in 10, a warning alert is shown along with a recommended course of action, accompanied by a heatmap visualization explaining the model's reasoning.


My Drive - Google Drive
Final_Year_Project.ipynb - Colab
Final_Project_Interface.ipynb
Depression XAI
7238-34-81-178-38.ngrok-free.app
New Chrome available
Analyze
painful
help

Add Your Thought:

i feel so painful who can help me

Depression Detection + XAI Explanation

Prediction Result: Depression

📌 Classifying Type of Depression

● We detected signs of depression.

Type: Postpartum disorder

Recommendation: We encourage you to pay more attention to your emotional state. If symptoms persist, please seek professional help from a mental health specialist as soon as possible.

XAI Explanation (Red = High Impact, White = Medium Impact, Blue = Low Impact)

● Dark Red: Strong indicator of 'Depression' | ● Dark Blue: Supports 'No Depression' | 📌 Powered by DistilBERT + Streamlit

Figure 10: Depressive Emotions Display Interface

This interactive and transparent feedback process empowers users with both a diagnosis and an explanation, enhancing interpretability and user trust.

Model Comparison

A comparative analysis was conducted between DistilBERT and MentalBERT on both binary and multiclass classification tasks. As shown in Table 1, DistilBERT consistently outperformed MentalBERT across all evaluation metrics in both settings. In binary classification, DistilBERT achieved 95% in accuracy, precision, recall, and F1-score, while MentalBERT achieved 92% for each metric. In multiclass classification, the performance gap became more pronounced, with DistilBERT maintaining a high accuracy of 94%, compared to only 66% for MentalBERT. Similar differences were observed in precision (94% vs. 68%), recall (94% vs. 66%), and F1-score (94% vs. 63%).

These results indicate that DistilBERT is more reliable and robust, even in complex classification scenarios involving multiple depressive subtypes. The performance degradation of MentalBERT in multiclass tasks may stem from overfitting or limited domain-specific generalizability, despite being pre-trained on mental health data.

Table 1: Performance Comparison Between DistilBERT and MentalBERT

Metric	DistilBERT (Binary Class)	MentalBERT (Binary Class)	DistilBERT (Multiclass)	MentalBERT (Multiclass)
Accuracy (%)	95	92	94	66
Precision (%)	95	92	94	68
Recall (%)	95	92	94	66
F1-Score (%)	95	92	94	63

CONCLUSION

This study presents a reliable approach to detecting and classifying depressive disorders using natural language processing and deep learning techniques. The proposed system effectively combines a fine-tuned DistilBERT model with an explainable AI (XAI) framework to identify and distinguish among Major Depressive Disorder (MDD), Bipolar Disorder (BP), and Postpartum Depression (PPD) based on user-submitted text. The binary classification model achieved an impressive accuracy of 95%, while the multiclass model demonstrated strong performance in identifying depressive subtypes, particularly Postpartum Depression.

The integration of attention-based visualization enhances interpretability by highlighting emotionally significant keywords, enabling users to understand the reasoning behind model predictions. The web-based interface, developed using Streamlit, offers a seamless and accessible platform for real-time interaction, with intuitive feedback for both depressive and non-depressive inputs. This system demonstrates practical potential for early screening, self-assessment, and mental health awareness, particularly among social media users or individuals reluctant to seek professional help.

Comparative analysis further reveals that while MentalBERT offers marginal advantages in recognizing PPD, DistilBERT outperforms in overall accuracy, efficiency, and inference speed, making it suitable for lightweight deployment scenarios.

Despite its promising results, the study acknowledges several limitations. The dataset primarily consists of English-language social media posts, which may not generalize well across cultures or clinical populations. Furthermore, the system's ability to detect atypical or co-morbid depressive symptoms remains limited.

Future improvements may focus on expanding the dataset to include multilingual and multicultural samples, integrating multimodal features such as voice tone and facial expressions, and deploying the system on mobile platforms for broader accessibility. Additionally, clinical validation through collaboration with mental health professionals could enhance the reliability and acceptance of the system in real-world healthcare settings.

In summary, this research contributes an effective, interpretable, and user-friendly solution to the growing demand for AI-assisted mental health screening tools. By bridging the gap between machine learning performance and human-centered design, the proposed system lays the groundwork for future advancements in personalized mental health support.

ACKNOWLEDGEMENT

I would like to express my sincere appreciation to my supervisor, Dr. Lailatul Qadri Zakaria, for the continuous guidance, constructive feedback, and encouragement throughout the development of this project. I am also grateful to all individuals who participated in testing the system—your input was invaluable in improving its performance and usability.
Lastly, I wish to thank my family and friends for their constant support, patience, and motivation during the entire research journey.

REFERENCES

Anubrata Deb, Bristi Samadder, Souroja Chowdhury, Mrs. Sulekha Das, Mrs. Shweta Banarjee. 2023. Measuring Mental Health Condition using Logistic regression. Ijetms. 7(2):327-338. https://doi.org/10.46647/ijetms.2023.v07i02.040.

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. 2022. AI Explainability 360: Impact and Design. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12651-12657. https://doi.org/10.1609/aaai.v36i11.21540.

Ballesteros JA, Ramírez V. GM, Moreira F, Solano A and Pelaez CA. 2024. Facial emotion recognition through artificial intelligence. *Front. Comput. Sci.* 6:1359471. https://doi.org/10.3389/fcomp.2024.1359471.

Bhuvaneswari, M., & Prabha, V.L. 2023. A deep learning approach for the depression detection of social media data with hybrid feature selection and attention mechanism. *Expert Systems 40*(9). https://doi.org/10.1111/exsy.13371.

Haque UM, Kabir E, Khanam R. 2021. Detection of child depression using machine learning methods. *PLoS One. 16*(12). https://doi.org/10.1371/journal.pone.0261131.

Mansoor MA, Ansari KH. 2024. Early Detection of Mental Health Crises through Artifical-Intelligence-Powered Social Media Analysis: A Prospective Observational Study. *J Pers Med. 14*(9):958. https://doi.org/10.3390/jpm14090958.

Shimada, K. 2023. The Role of Artificial Intelligence in Mental Health: A Review. *Science Insights*, *43*(5), 1119–1127. https://doi.org/10.15354/si.23.re820

Temidayo Omoniyi, 2024, DistilBERT for Multiclass Text Classification Using Transformers, Medium, https://medium.com/@kiddojazz/distilbert-for-multiclass-text-classification-using-transformers-d6374e6678ba.

V. Maheshwar., N. Venu Gopal., V. Naveen Kumar., D. Pranavi. and Y. Padma Sai. 2023. Development of an SVM-based Depression Detection Model using MFCC Feature Extraction. *International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 2023, pp. 808-814, https://doi.org/10.1109/ICSCSS57650.2023.10169770.

Vijay Arya; Rachel K. E. Bellamy; Pin-Yu Chen; Amit Dhurandhar; Michael Hind; Samuel C. Hoffman; Stephanie Houde; Q. Vera Liao; Ronny Luss; Aleksandra Mojsilović; Sami Mourad; Pablo Pedemonte; Ramya Raghavendra; John Richards, Prasanna Sattigeri; Karthikeyan Shanmugam; Moninder Singh, Kush R. Varshney; Dennis Wei; Yunfeng Zhang. 2020. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. https://doi.org/10.48550/arXiv.1909.03012.

Walschots, Quinty, Milan Zarchev, Maurits Unkel, and Astrid Kamperman. 2024. Using Wearable Technology to Detect, Monitor, and Predict Major Depressive Disorder—A Scoping Review and Introductory Text for Clinical Professionals. *Algorithms* 17, no. 9: 408. https://doi.org/10.3390/a17090408.

Zafar F, Fakhare Alam L, Vivas RR, Wang J, Whei SJ, Mehmood S, Sadeghzadegan A, Lakkimsetti M, Nazir Z. 2024. The Role of Artificial Intelligence in Identifying Depression and Anxiety: A Comprehensive Literature Review. *Cureus* 16(3):e56472. https://doi.org/10.7759/cureus.56472.

Lu Yinuo (A191411)
Dr.Lailatul Qadri Zakaria
Faculty of Information Technology & Science
National University of Malaysia