Automated Depression Detection from Social Media Using Deep Learning

Pagadala Srinivasu[1], Bankuru Yashwanth[2], Chatla Tejesh Reddy[3],
Billa Ganesh Venkateswara Sai[4], Gollu Ramalakshmi[5]
[1]*Assistant Professor, Department of Computer Science and Engineering (Data Science), Raghu Engineering College (Autonomous), Dakamarri, Visakhapatnam*
[2,3,4]*B. Tech (Final Year Student), Department of Computer Science and Engineering (Data Science), Raghu Engineering College (Autonomous), Dakamarri, Visakhapatnam*

***Abstract*—Mental health conditions, particularly clinical depression, have emerged as a significant global health crisis, affecting millions of individuals across diverse demographics. In the digital era, social media platforms have become primary venues for self-expression, offering a unique "digital phenotype" of a user's psychological state. This research focuses on identifying depression through the automated analysis of social media text data using state-of-the-art deep learning architectures. We implement and evaluate four hybrid models: BERT combined with Convolutional Neural Networks (CNN), DistilBERT with Bidirectional Long Short-Term Memory (BiLSTM), DeBERTa with BiLSTM, and DistilGPT2 paired with BiLSTM. The methodology encompasses rigorous data cleaning, tokenization using model-specific vocabularies, and the extraction of high-dimensional contextual embeddings. Our experimental results, analysed through detailed confusion matrices, demonstrate that these models significantly outperform traditional machine learning techniques by capturing subtle semantic and temporal patterns in text. The final system is deployed via a Flask-based web application, providing an accessible and scalable tool for early mental health screening and monitoring.**

***Index Terms*—Mental Health, Depression Detection, Social Media Analytics, Deep Learning, Transformers, BERT, BiLSTM, Natural Language Processing (NLP), Flask.**

I. INTRODUCTION

1.1 Context and Motivation

Mental health is a fundamental pillar of overall well-being, yet depression remains one of the most prevalent and under-treated conditions worldwide. Traditional diagnostic methods often rely on self-reporting or clinical interviews, which can be hindered by social stigma, cost, and a lack of access to specialists. The widespread adoption of social media has provided a new frontier for mental health research. Users often share their internal thoughts, daily struggles, and emotional shifts in real-time. These text-based expressions contain linguistic markers such as the frequent use of first-person pronouns, increased negative sentiment, and diminished lexical variety that can serve as early indicators of depressive episodes. The motivation for this project lies in leveraging these digital footprints to create a proactive, automated screening tool that can assist individuals and clinicians in identifying early warning signs.

1.2 Problem Statement

Early detection of depression is critical for effective intervention, yet manual analysis of vast amounts of social media data is unfeasible. Existing automated systems often utilize traditional machine learning models (e.g., SVM, Naïve Bayes) that rely on shallow feature extraction like Bag-of-Words or TF-IDF. These methods fail to understand the "context" and "sentiment nuance" of informal social media language, which is often laden with slang, abbreviations, and complex emotional subtext. There is a clear need for a deep learning framework that can interpret these high-level semantic dependencies accurately.

1.3 Objectives

The primary objective of this project is to develop and compare multiple deep learning architectures to determine the most effective method for depression classification. Specific goals include:

- Developing a robust preprocessing pipeline for noisy social media text.
- Implementing Transformer-based models (BERT, DeBERTa, GPT-2) to capture deep contextual meanings.
- Integrating hybrid layers (CNN and BiLSTM) to analyze spatial and sequential patterns.
- Building a user-friendly web interface for real-time mental health prediction.

II. LITERATURE REVIEW

The field of automated depression detection has evolved rapidly with the advent of Large Language Models (LLMs).

- Sridharan et al. (2023): Explored deep learning-based detection, emphasizing that neural networks can learn hidden emotional cues better than manual feature engineering. Their research highlights the importance of balancing datasets to avoid bias toward "Healthy" classifications.
- Kim et al. (2025): Focused on "Explainable AI" (XAI) in mental health. By using LLM-derived embeddings, they demonstrated that models can not only classify a user as depressed but also highlight specific linguistic tokens that contributed to the decision, building trust between the AI and clinicians.
- Bucur et al. (2023): Introduced the concept of "Time-Enriched" Transformers. They argued that the frequency and interval of posts are as important as the content itself. Their multimodal approach showed that users entering a depressive state often change their posting frequency alongside their vocabulary.
- Qin et al. (2023): Proposed an interactive system where the AI "Read, Diagnose, and Chat." This emphasizes a human-centric design where the detection system serves as an interactive support tool rather than a black-box classifier.

III. METHODOLOGY

3.1 Data Acquisition and Preprocessing

We utilized a dataset consisting of labeled social media posts (e.g., from Reddit or Twitter) specifically curated for mental health research.

1. Noise Removal: Stripping HTML tags, URLs, emojis (or converting them to text), and special characters that do not contribute to sentiment.
2. Normalization: Standardizing slang and contractions (e.g., "don't" to "do not") to ensure the tokenizer recognizes the core intent.
3. Tokenization: We used model-specific tokenizers (e.g., WordPiece for BERT) to handle "Out-Of-Vocabulary" (OOV) words by breaking them into sub-word units.

3.2 Proposed Deep Learning Architectures

We developed four distinct pipelines to compare their efficacy:

1. BERT + CNN: BERT generates high-dimensional word embeddings. A 1D-CNN layer then acts as a feature extractor to identify local patterns (n-grams) that are indicative of depressive thoughts.
2. DistilBERT + BiLSTM: DistilBERT provides a computationally efficient contextual base. The BiLSTM layer processes these sequences from both directions (past-to-future and future-to-past), which is essential for understanding the emotional flow of a sentence.
3. DeBERTa + BiLSTM: DeBERTa uses a disentangled attention mechanism, allowing the model to understand the relative position and content of words more effectively than standard BERT.
4. DistilGPT2 + BiLSTM: Utilizing a generative pre-trained model as a feature extractor. Despite GPT-2's causal nature, its embeddings are highly sensitive to language patterns used in conversational social media text.

3.3 System Architecture

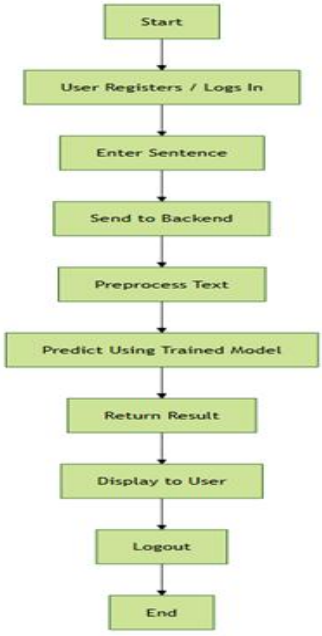


Figure 1: High-level System Architecture illustrating Data Flow from Input to Classification

IV. SYSTEM DESIGN AND IMPLEMENTATION

4.1 System Modules

- Data Import: Responsible for batch loading and label verification.
- Preprocessing: Implements the cleaning and tokenization logic.
- Training & Evaluation: Fine-tunes the Transformer models on the mental health dataset and calculates metrics.
- Backend (Flask): The bridge between the Python models and the web frontend.
- User Module: Manages secure registration, login sessions, and the prediction interface.

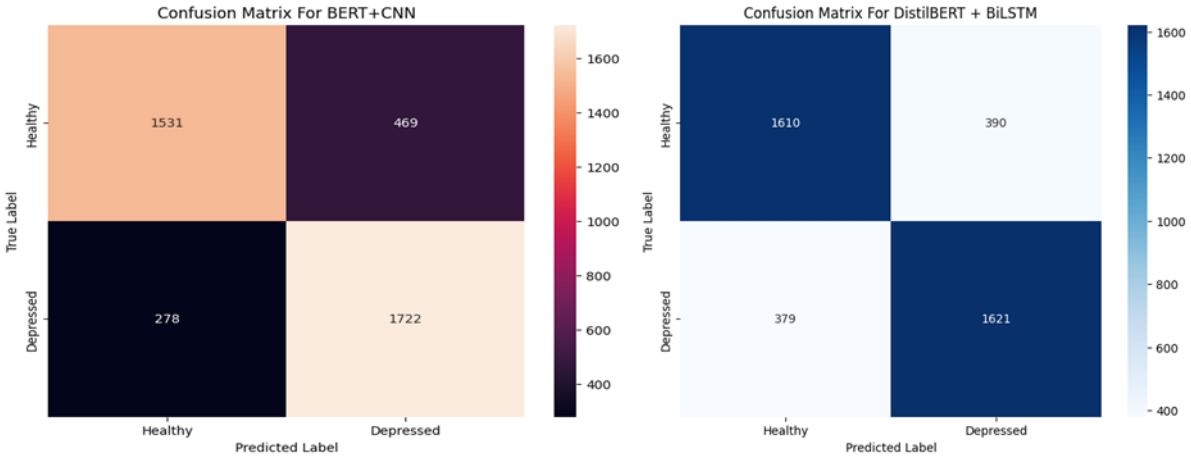
V. RESULTS AND DISCUSSION

5.1 Comparative Analysis

The models were evaluated on a test set to determine their accuracy in binary classification ("Healthy" vs. "Depressed").

Architecture	Correct Healthy (TN)	Correct Depressed (TP)	False Positive (FP)	False Negative (FN)
BERT + CNN	1531	1722	469	278
DistilBERT + BiLSTM	1610	1621	390	379
DeBERTa+ BiLSTM	1658	1585	342	415
DistilGPT2 + BiLSTM	1526	1689	474	311

5.2 Performance Visualization



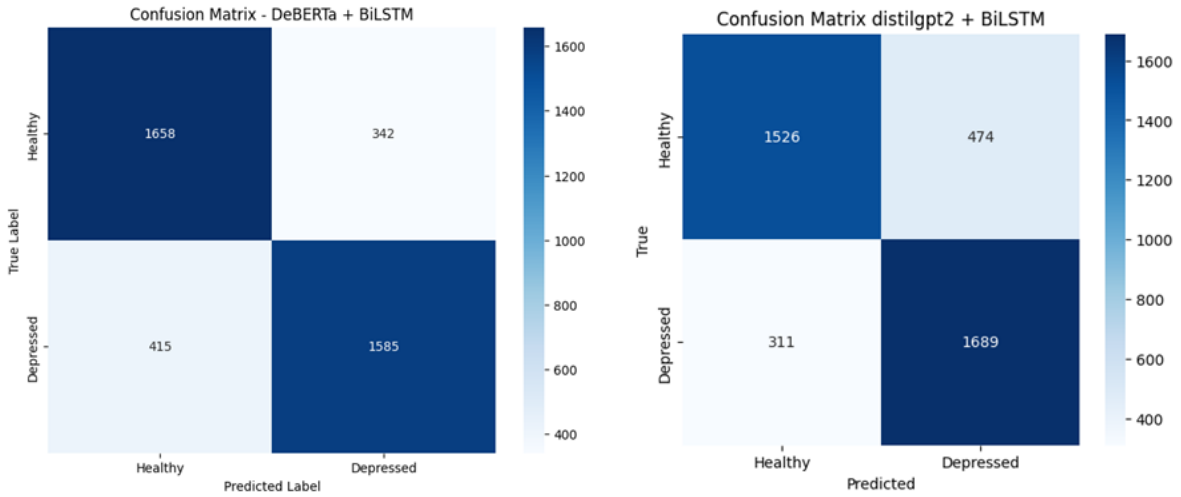


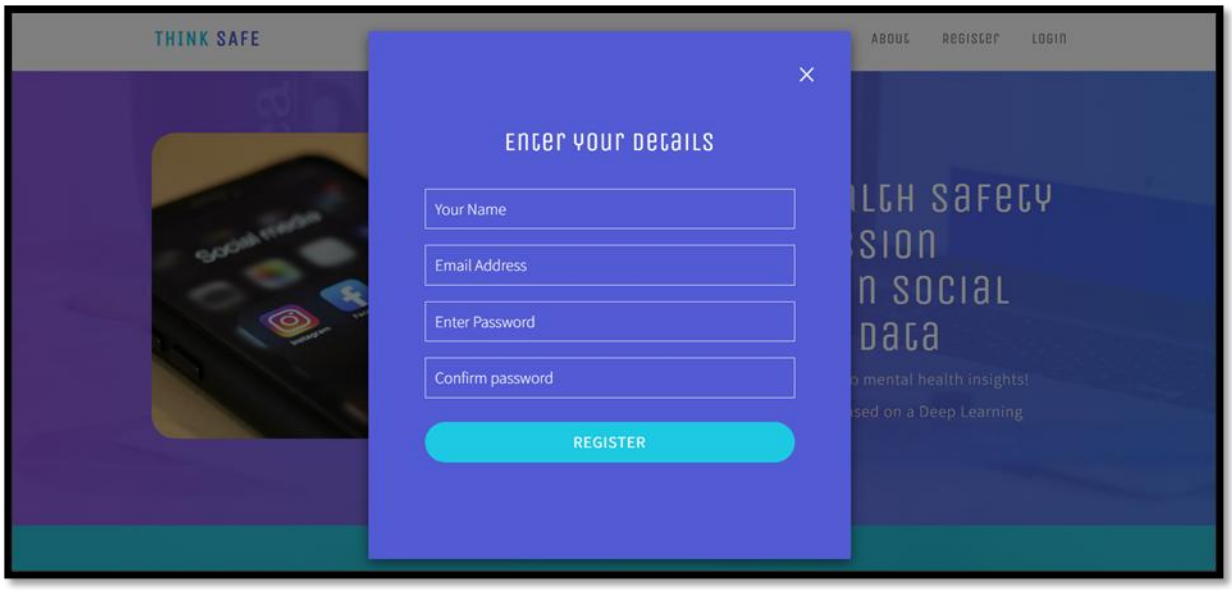
Figure 2: Grid of Confusion Matrices for the four tested architectures

Analysis:

- DeBERTa + BiLSTM proved to be the most "Specific" model, correctly identifying 1658 healthy cases. This minimizes the risk of "False Alarms."
- BERT + CNN achieved the highest "Recall" for depression (1722), meaning it is the most effective at catching every possible case of depression, though it may occasionally flag healthy users.
- DistilBERT offered the best balance between speed and accuracy, making it a strong candidate for real-time mobile applications.

5.3 Web Interface Integration

The final model was integrated into the "THINK SAFE" web application. The interface allows users to paste social media excerpts and receive an immediate classification with a confidence score.



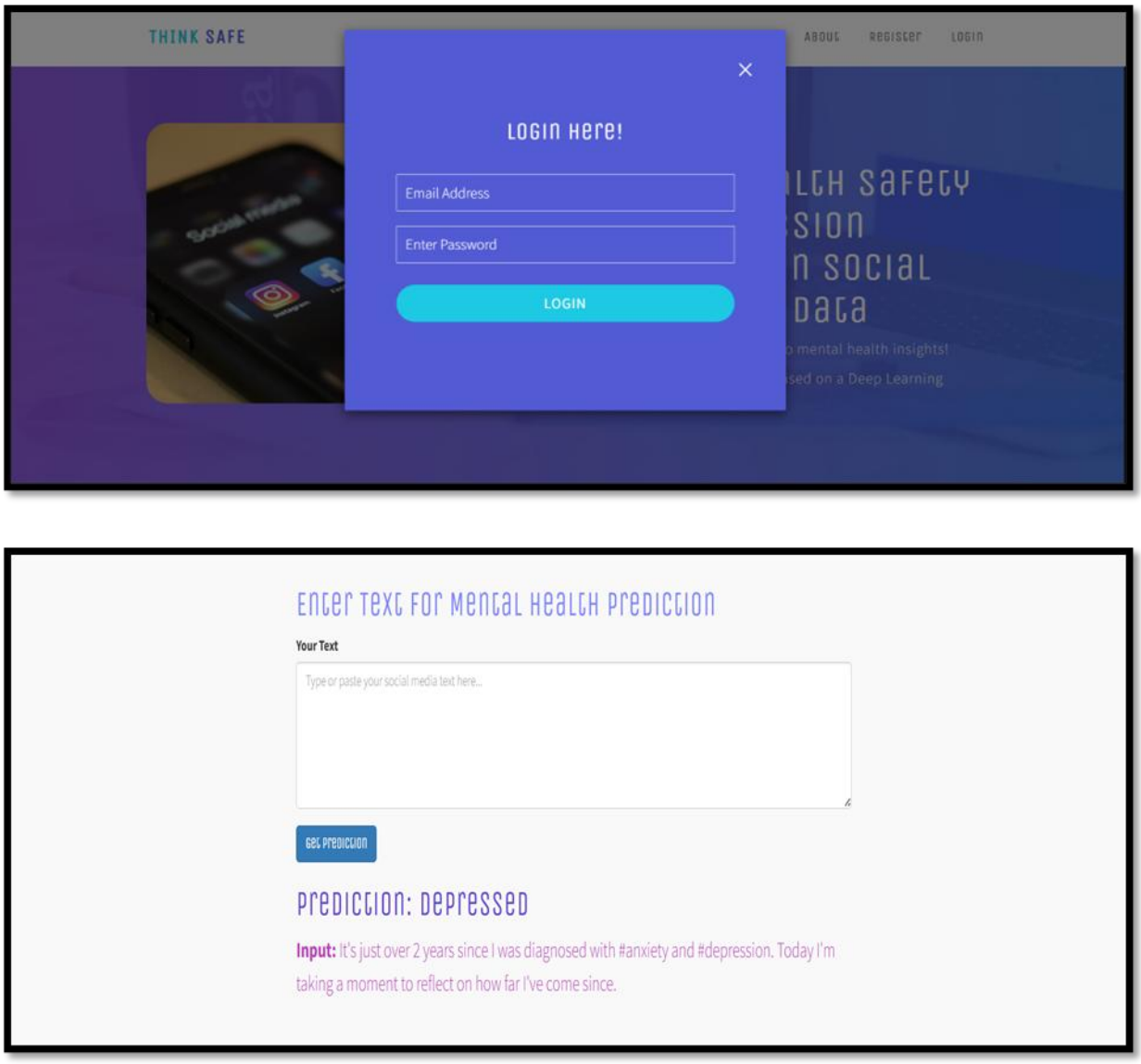


Figure 3: Screenshots of the Web Application including Registration, Login, and Prediction Results

VI. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This research successfully demonstrates that deep learning architectures, particularly hybrid Transformer models, can accurately identify depression from social media text. By combining the contextual understanding of BERT and DeBERTa with the sequential memory of BiLSTM, we created a system capable of interpreting complex emotional nuances. Our comparative study shows that while different models have different strengths (Precision vs. Recall), they all significantly outperform traditional linguistic analysis tools.

6.2 Future Work

- Multilingual Support: Expanding the dataset to include regional Indian languages to increase impact.
- Multimodal Analysis: Integrating image analysis (e.g., profile pictures or shared images) alongside text.
- Privacy-Preserving AI: Implementing Federated Learning to allow users to screen their data locally without uploading it to a server.
- Explainable Outputs: Adding a "heat map" feature to the UI to show users which words led to the prediction.

REFERENCES

[1] S. Sridharan, S. Aishwarya, S. P. Aishwarya, and S. Santhiya, "Depression Detection in Social Media Users Using Deep Learning," Am. J. Psychiatr. Rehabil., 2023.

[2] F. A. Sakib, A. A. Choudhury, and O. Uzuner, "MASON-NLP at eRisk 2023: Deep Learning-Based Detection of Depression Symptoms from Social Media Texts," in Proc. Working Notes of CLEF, 2023.

[3] F. K. Karim et al., "Deep Learning for Depression Detection Using Twitter Data," Intell. Autom. Soft Comput., vol. 37, no. 1, pp. 101–115, 2023.

[4] S. Banerjee and N. F. Shaikh, "Mental Health Care Using Social Media Data with Deep Learning Framework," in Proc. 3rd Int. Conf. Intell. Comput. Inf. Control Syst. (ICICICS), Springer, 2022, pp. 245–258.

[5] M. N. Hoque and U. Salma, "Detecting Level of Depression from Social Media Posts for the Low-resource Bengali Language," J. Eng. Adv., vol. 4, no. 1, pp. 12–20, 2023.

[6] T. S. Kumar, "A Deep Learning Framework with a Hybrid Model for Automatic Depression Detection in Social Media Posts," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 2, pp. 450–462, 2024.

[7] S. Kim, O. Imieye, and Y. Yin, "Interpretable Depression Detection from Social Media Text Using LLM-Derived Embeddings," arXiv preprint arXiv: 2501.XXXXX, 2025.

[8] A.-M. Bucur, A. Cosma, P. Rosso, and L. P. Dinu, "It's Just a Matter of Time: Detecting Depression with Time-Enriched Multimodal Transformers," in Proc. 45th Eur. Conf. Inf. Retrieval (ECIR), 2023, pp. 140–155.

[9] W. Qin et al., "Read, Diagnose and Chat: Towards Explainable and Interactive LLMs-Augmented Depression Detection in Social Media," in Proc. 37th Conf. Neural Inf. Process. Syst. (NeurIPS), 2023.

[10] N. Ali et al., "Diverse Perspectives, Divergent Models: Cross-Cultural Evaluation of Depression Detection on Twitter," in Proc. Int. Conf. Comput. Linguist. (COLING), 2024, pp. 89–102.

[11] F. Ahsan, A. A. Choudhury, and O. Uzuner, "Detecting Depression Symptoms in Social Media Texts," Bullet Papers, 2023.