MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
A Constituent Unit of MAHE, Manipal

EchoMind: A Window to Mental Health

Project Phase-4 Report Submitted

to

MANIPAL ACADEMY OF HIGHER EDUCATION

For Partial Fulfillment of the Requirement for the
Award of the Degree
Of

Bachelor of Technology

in

Data Science & Engineering

by

Neha Bandlamudi, Himanjali Ganapa, Puja Rakshit,

Reg. No. 230968001, Reg. No. 230968038, Reg. No. 230968048

Under the guidance of

Dr. Gangothri S.

Assistant Professor

School of Computer Engineering

Manipal Institute of Technology

MAHE, Manipal, Karnataka, India

Contents

Problem Statement:

Depression ranks among the major causes of disability globally, making it one of the most urgent public health issues. Due to its extensive repercussions, which include decreased social functioning and suicidal thoughts, early and correct detection is critically needed. Even though prompt action can have a big impact, social stigma and accessibility issues are two common obstacles to standard diagnostic techniques. Consequently, AI-powered methods that examine readily available data, such as text or audio, have become potential tools for the early detection of depression.

Literature Review:

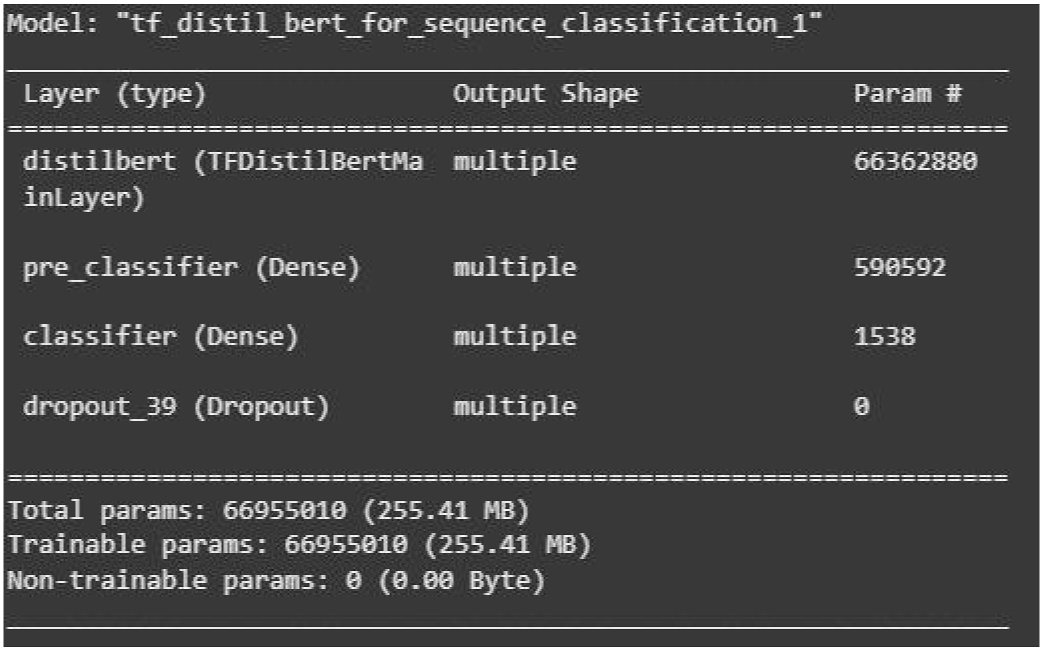
Paper Name	**Methodology**	**Conclusion**
Ensembles of BERT for Depression Classification	Uses transcripts from the DAIC-WOZ clinical interview corpus; three individual BERT variants + four different ensemble strategies of BERT variants; classification of depressed vs non-depressed from interview response transcripts.	Ensembles of BERT models improve robustness and mean F1 scores over individual BERT variants in transcript-based depression classification.
Diagnosis of Depression Based on the Four-Stream Model of Bi-LSTM and CNN from Audio and Text Information	Proposes a four-stream model combining Bidirectional LSTM and CNN on audio and text information (multi-modal) for depression diagnosis; trained/tested on clinical interview datasets.	The multi-stream approach combining audio and text features yields high accuracy for depression diagnosis, showing the benefit of multi-modality rather than unimodal text or audio alone.

Depression Detection and Analysis using Large Language Models on Textual and Audio-Visual Modalities	Uses transcripts (extracted via a Whisper model) + audio-visual modalities; textual network (LLMs such as GPT-4, RoBERTa) + audio/visual BiLSTM; regression of PHQ-8 scores + classification.	Their multimodal framework achieves strong performance in both classification and severity estimation of depression (PHQ-8), underscoring the value of combining modalities and leveraging large language models.

Viability Of Our Model Chosen:

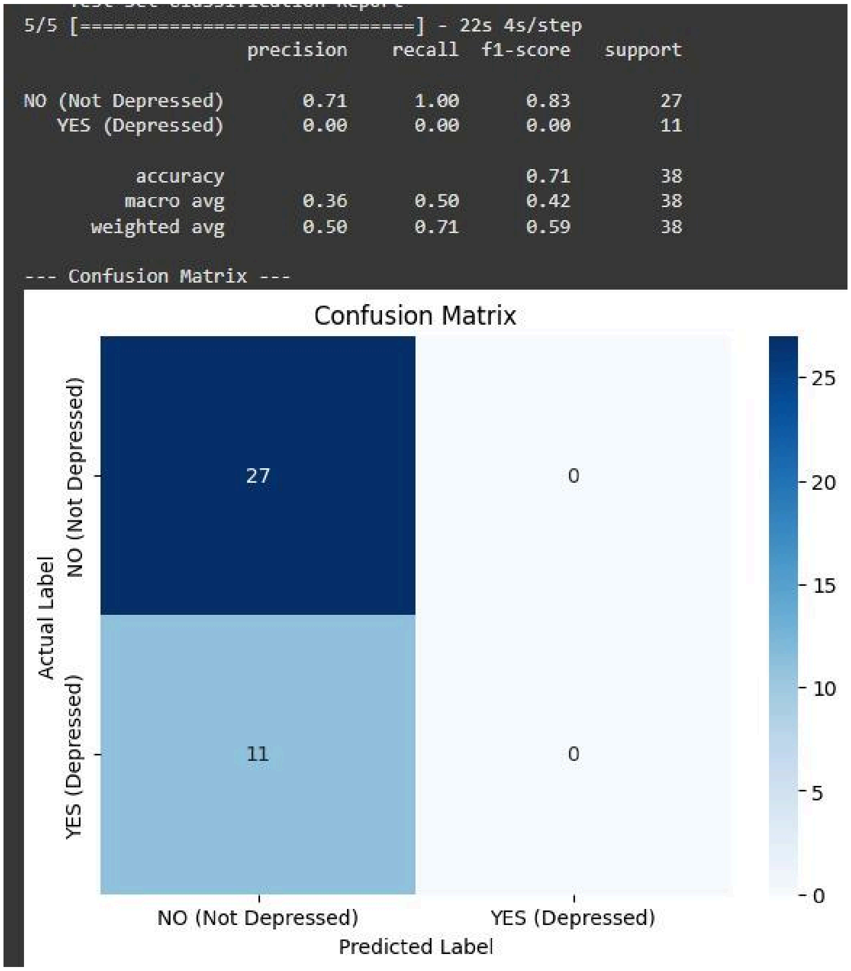
For our model, we used DistilBERT, which is a lighter and faster version of the original BERT model. Even though it's smaller, it still understands language really well, which is important because we want to pick up subtle signs of depression in what people say. One big advantage is that it's much faster and doesn't need as much computer power, so training the model and testing it on interviews becomes easier and more practical. It is more efficient for our dataset as we included an auto-tokenizer with it, which boosted our test accuracy. It also works nicely with audio features, meaning we can combine what people say with how they say it without making the model too heavy or slow. Overall, DistilBERT helps us get strong results while keeping the system efficient and manageable, making it a smart choice for our project.

Model Summary:



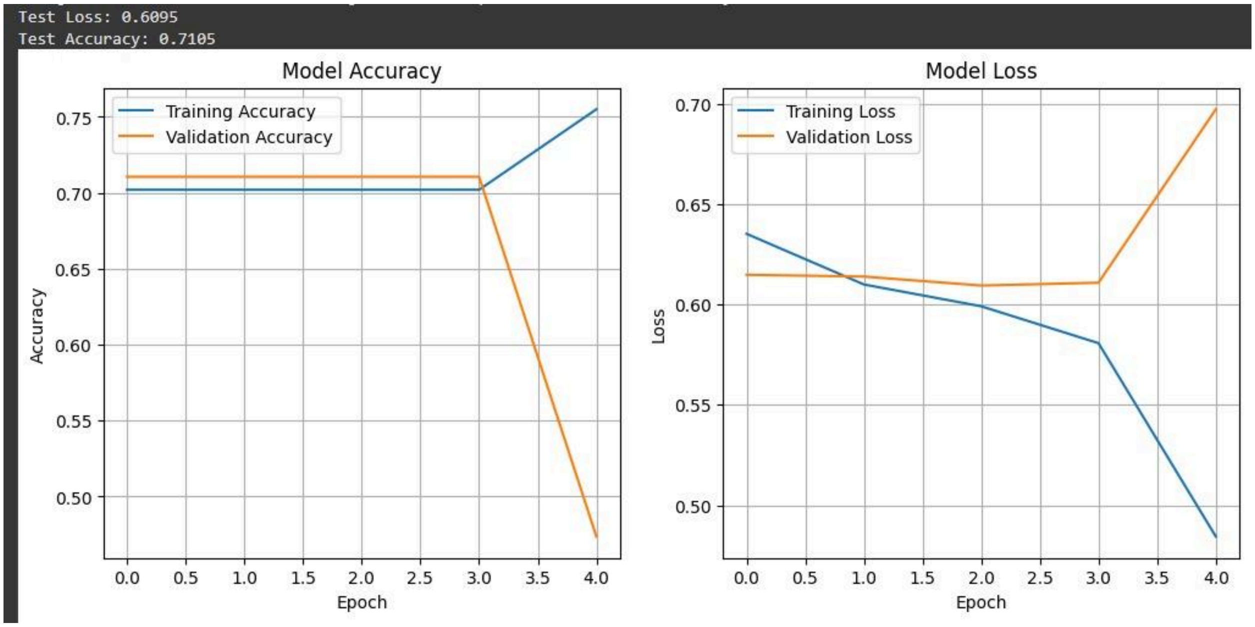
```
Model: "tf_distil_bert_for_sequence_classification_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 distilbert (TFDistilBertMa  multiple                  66362880
 inLayer)

 pre_classifier (Dense)      multiple                  590592

 classifier (Dense)          multiple                  1538

 dropout_39 (Dropout)        multiple                  0

=================================================================
Total params: 66955010 (255.41 MB)
Trainable params: 66955010 (255.41 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Metrics Result:



```
5/5 [==============================] - 22s 4s/step
                    precision    recall  f1-score   support

NO (Not Depressed)       0.71      1.00      0.83        27
   YES (Depressed)       0.00      0.00      0.00        11

          accuracy                           0.71        38
         macro avg       0.36      0.50      0.42        38
      weighted avg       0.50      0.71      0.59        38

--- Confusion Matrix ---
```

Loss and Accuracy:



```
Test Loss: 0.6095
Test Accuracy: 0.7105
```

Models Not Chosen and Why:

CNN:

CNNs have been widely applied in speech-based emotion and depression detection due to their ability to extract local spectral features from audio inputs such as Mel-spectrograms or MFCCs. However, CNNs primarily focus on local spatial patterns and struggle to capture the long-term temporal dependencies present in human speech, which are critical for identifying depression-related cues like slowed speech rate, hesitation, and prolonged silence. Moreover, CNNs often require large training datasets to generalize effectively, but depression datasets such as DAIC-WOZ are typically small and imbalanced, leading to overfitting. As a result, while CNNs can learn surface-level acoustic representations, they fail to capture the subtle and context-dependent patterns necessary for accurate depression detection.

LSTM:

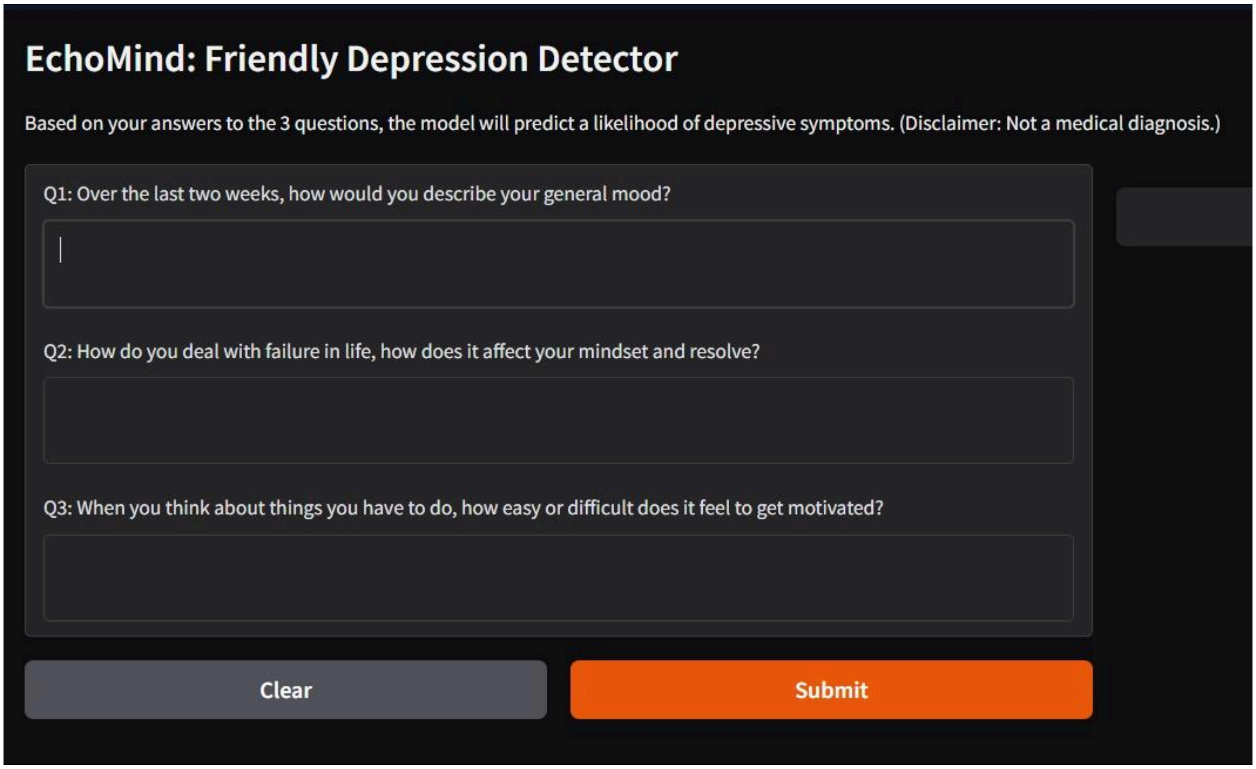
LSTM networks are designed to handle sequential data and are theoretically capable of modeling temporal dependencies in audio or text. However, in practice, their performance in depression detection is limited by data scarcity and high variability in human expression. LSTMs tend to forget long-range dependencies or bias toward recent context, which can hinder their ability to model emotional consistency throughout an interview. Additionally, training LSTMs on long sequences such as full interview transcripts or extended audio features can be computationally expensive and prone to vanishing or exploding gradients. Without large-scale pretraining, these models often struggle to extract generalized representations, resulting in reduced robustness across diverse speakers or recording conditions.

BERT transformer model:

When we tried using a full BERT transformer for our transcripts, we ran into some issues. BERT splits text into very small tokens, which works well for general text, but our transcripts were long and had complex sentences. Because of this, a lot of important context got broken up, and the model couldn’t really understand the flow of the conversation. This made it less reliable for spotting subtle signs of depression. On top of that, when we checked the results, metrics like accuracy were low, showing that the model wasn’t capturing the patterns we needed. It made us realize that smaller, more flexible models might work better for our data.

Information on our Application:

For deploying our project, we are using Hugging Face along with an API called Gradio, which makes it easy to create a simple and interactive interface. In our application, we ask the user three questions, and their answers are sent directly to the model. The model then analyzes these responses and predicts whether the person might be experiencing depression. Using TensorFlow in the backend helps the model run smoothly and efficiently, while Gradio allows anyone to interact with it easily through a web interface. This setup makes our project practical and user-friendly, letting people get quick insights without any complicated setup.



EchoMind: Friendly Depression Detector

Based on your answers to the 3 questions, the model will predict a likelihood of depressive symptoms. (Disclaimer: Not a medical diagnosis.)

Q1: Over the last two weeks, how would you describe your general mood?

Q2: How do you deal with failure in life, how does it affect your mindset and resolve?

Q3: When you think about things you have to do, how easy or difficult does it feel to get motivated?

Future Scope:

For the future, our project can be expanded to provide more meaningful insights rather than just a simple “depressed” or “not depressed” classification. One idea is to give users a score from 0 to 10, which would indicate the severity of depression. This could allow us to classify people in a hierarchical order, making it easier to identify those who might need immediate help. With this kind of scoring, the system could also be extended to detect more serious risks, such as suicidal thoughts, which could be life-saving. By providing a more nuanced output, the tool could become not just a screening system but a guide for intervention and support.

Another important future direction is to improve the model's accuracy and robustness. We could explore predicting depression using only audio, which would allow the system to work even when users are unable or unwilling to provide textual responses. Additionally, collecting and using a larger, more diverse dataset would help the model generalize better to different age groups, accents, and speaking styles. Overall, these improvements could make the project more practical, inclusive, and reliable, helping reach a wider range of people and providing better support for mental health monitoring.

References:

1. Main Research Paper: https://arxiv.org/abs/2409.08483 **.A BERT-Based Summarization approach for depression detection.**
2. https://git.unicaen.fr/kirill.milintsevich/hierarchical-depression-symptom-classifier
3. HuggingFace Link:https://huggingface.co/spaces/himanjali-g/EchoMind
4. DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-P. (2014). "SimSensei kiosk: A virtual human interviewer for healthcare decision support". In Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'14), Paris. The Distress Analysis Interview Corpus of human and computer interviews. InLREC 2014 May (pp. 3123-3128) https://schererstefan.net/assets/files/papers/508_Paper.pdf