Multimodal LLM for Mental Health Detection and Analysis

Chayan Tank
MT23030
chayan23030@iiitd.ac.in

Sarthak Pol
MT23082
sarthak23082@iiitd.ac.in

Sonik Sandip Sarungale
MT23098
sonik23098@iiitd.ac.in

Mahisha Ramesh
MT23121
mahisha23121@iiitd.ac.in

Shaina Mehta
MT23139
shaina23139@iiitd.ac.in

Vinayak Katoch
MT23105
vinayak23105@iiitd.ac.in

Github: https://github.com/shaina-12/Multimodal-LLM-for-Mental-Health-Detection-and-Analysis.git
(Baseline Branch)

1. Problem Formulation

In the 21st century, a large number of people are suffering from various mental health disorders such as depression, bipolar disorder, autism spectrum disorder, anxiety disorder, etc., which impact not only their cognition but also their behaviors, emotions, physical health, and their normal life [1]. It is mainly caused by working conditions, the lifestyle of an individual, family issues, abuse, etc [2]. It causes harm to an individual's physical, mental, and social well-being but also leads to suicide in the worst cases. Traditionally, several questionnaires/scales, such as the PHQ 8 questionnaire, YMRS scale, etc., have been proposed for the recognition of mental issues like depression and bipolar disorder, etc., but they are highly subjective in nature [3]. Nowadays, several researchers and organizations are working on AI-based solutions for mental health recognition. Most of the works in the literature are based on deep learning and machine learning-based solutions.

1.1 Project Objectives

The objective of this project is as follows:

1. To develop a multi-modal LLM on audio and textual modality for mental health recognition.
2. Later on, we will try to incorporate the visual modality into the proposed multi-modal LLM

1.2 Objective of the Baseline Report

The objective of this report is to show the inference results of the modified LTU-AS architecture i.e our Baseline proposed by [4] on the depression detection task on our custom dataset.

2. Dataset Description

We have collected a custom dataset named Depression Video Dataset which consists of 974 video segments (30 seconds each).

Then we extracted multiple features, which are Audio, video, and textual features, for further analysis.

For extracting Audio from Video we have used Moviepy library.
For extracting Text (Transcripts) from Audio we have used the whisper-large-v3 model.

3. About LTU-AS

LTU-AS [4] is intended to address a wide range of audio and speech tasks, and understand the audio and speech jointly.

The LTU-AS model is designed with the incorporation of full audio and speech understanding. It uses both perception and reasoning properties. The architecture comprises two major parts: Whisper and LLaMA. Whisper is deployed as the audio encoder-decoder and performs encoding for both speech and nonspeech sounds, obtaining the spoken text transcripts and additional paralinguistic information like tone, emotion, and background sounds. In addition to this, the Time and Layer-Wise Transformer (TLTR), which is pre-trained on AudioSet, captures more audio events.

The outputs of Whisper, including audio features and spoken text, will be fed to LLaMA, which is a very strong language model that works on reasoning. With the incorporation of both spoken text and audio events, LLaMA will generate highly sensible responses to natural language questions. This collaborative architecture allows LTU-AS to conduct operations like audio event recognition, speech context comprehension, and open-ended question answering by exploiting speech and non-speech cues.

LTU-AS takes the text prompt and the audio file as input. From that audio file, the transcript is taken out by the Open AI Whisper decoder, and the audio features are taken out by the Whisper encoder having a context window of 10 seconds. After that, audio features are passed through a Time and Layer-Wise Transformer (TLTR) transformer, pre-trained on AudioSet, which captures more audio-based features and then it is passed through the projection layer. Then the processed audio features, embeddings of the transcript, and the input prompt are passed through the LLAMA model, which is best suited for generation tasks, question-answering tasks, etc. The final output is returned in the form of text.

4. Methodology

We have modified LTU-AS for a classification task on our mental health objective using out custom dataset. In implementing the LTU-AS architecture for depression classification, we focus on gathering diverse audio samples to ensure a comprehensive representation of emotional states. We utilize the framework to transcribe and analyze these samples, aiming to detect subtle emotional cues indicative of depression. Our project aspires to enhance the accuracy of mental health assessments, providing valuable insights that can inform timely interventions.

Steps:

1. Dataset collection.
2. Get transcripts from audio and classify class using the **Meta-Llama-3.1-405B-Instruct-Turbo** model for ground truth.

Prompt used for ground Truth Generation:

"You are a psychologist tasked with assessing whether a person is experiencing symptoms of depression based on the transcript of an interview. The conversation typically includes discussions around mood, energy levels, sleep patterns, eating habits, concentration, interest in activities, and relationships. Use this information to classify if the person is likely depressed or not."

3. We modified LTU-AS architecture and prompts for processing of audio length from the default 10 seconds to 30 seconds (updated) and getting the proper outputs, respectively.

Prompt used for Baseline output Generation:

"Is the person in this audio depressed or not? Give me the answer in one word."

4. Get the outputs (classes - **Depressed** / **Not Depressed**) from the extracted audio files and create a pipeline for inference instead of the default using gradio client-server and using the frontend GUI to CLI script for sequential inference and saving the responses in a CSV file for evaluation.

LTU architecture, refer to the original research paper: (Joint Audio and Speech Understanding).

6. Evaluation and Results:

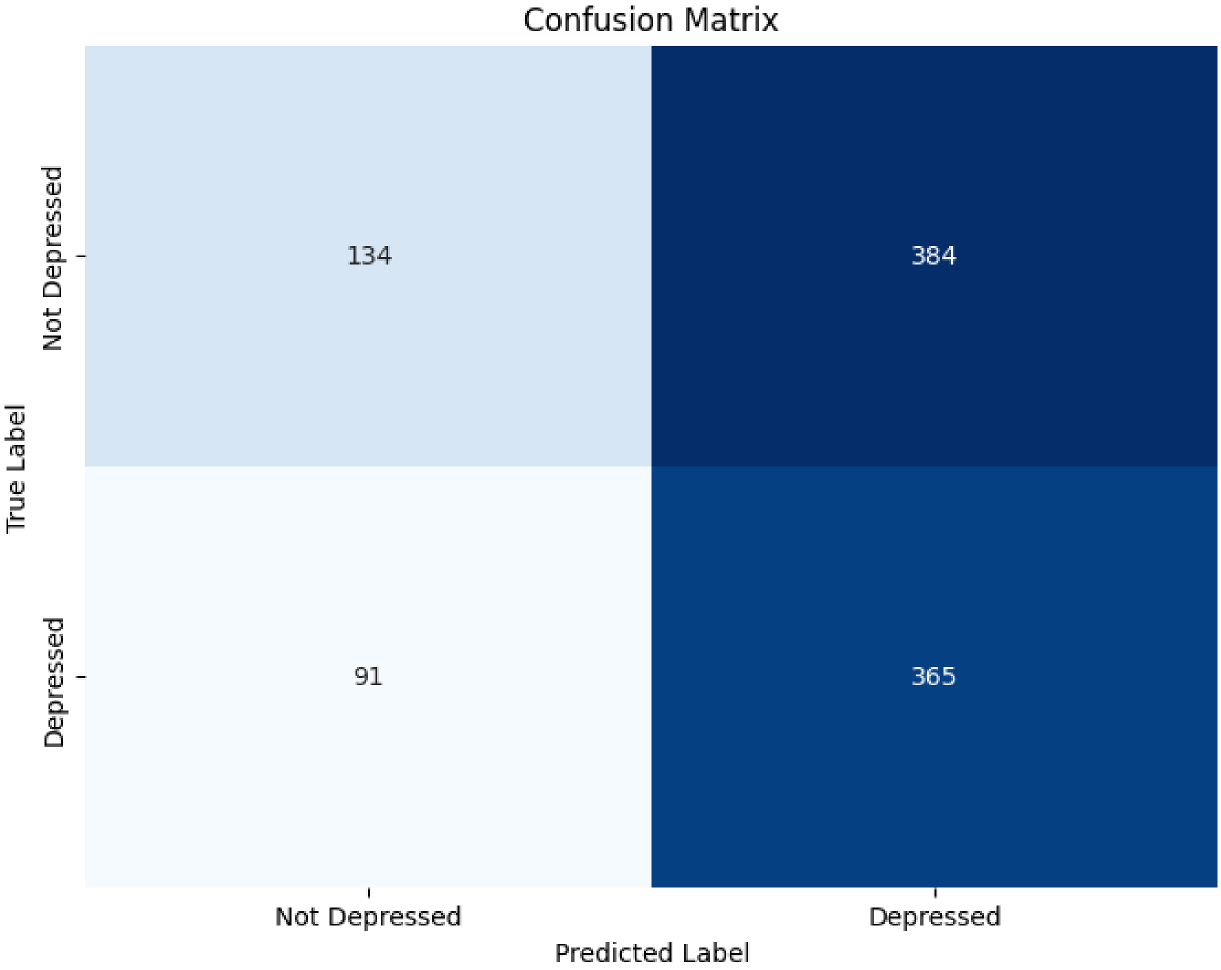
We have annotated the dataset using **'Meta-Llama-3.1-405B-Instruct-Turbo'** to generate the ground truth i.e. true labels for the samples. (in the future, we will get this dataset annotated by trained clinicians). For this, we extracted the Textual transcripts of every sample using the **whisper-large-v3 model**. These transcripts are then sent to Llama-3b-405B to get them classified into Two predefined classes: **Depressed and Not depressed.**

These generated labels are treated as the true labels. Now we have generated the outputs from our Baseline by performing inference on modified LTU-AS architecture for our dataset. These generated outputs from the baseline are then evaluated against the true labels.

Results:

Accuracy: 0.5123
Classification Report:

	precision	recall	f1-score	support
0	**0.60**	**0.26**	**0.36**	**518**
1	**0.49**	**0.80**	**0.61**	**456**
accuracy			**0.51**	**974**
macro avg	**0.54**	**0.53**	**0.48**	**974**
weighted avg	**0.54**	**0.51**	**0.48**	**974**



7. Conclusion and Future Work:

The majority of not-depressed class samples are classified as depressed. It can be concluded that the architecture is not good at understanding the markers for a non depressed individual. The Architecture, however, performed decently while predicting the depressed samples.

In the future, We will aim to modify the architecture such that the non-depressed class also contributes properly towards a better mental health classification. We also propose to improve this architecture by including video modality along with audio and text in the baseline. We will also explore modified training of video, text, and audio modality in 3-tuple-like contrastive learning for a better understanding of bifurcating the two classes and the contribution of each modality in its classification. We will use our custom dataset to benchmark these modifications. We will also get our dataset annotated by mental health clinicians for reliable ground truth.

References

[1] Mental disorders (no date). Available at: https://www.who.int/news-room/fact-sheets/detail/mental-disorders (Accessed: 13 September 2024).

[2] Luna-Jimenéz, C., Callejas, Z. and Griol, D., 2024, June. Mental-Health Topic Classification employing D-vectors of Large Language Models. In 2024 IEE

[3] Anand, A., Tank, C., Pol, S., Katoch, V., Mehta, S. and Shah, R.R., 2024. Depression Detection and Analysis using Large Language Models on Textual and Audio-Visual Modalities. arXiv preprint arXiv:2407.06125.

[4] Gong, Y., Liu, A.H., Luo, H., Karlinsky, L. and Glass, J., 2023, December. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1-8). IEEE.