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

1. Problem Formulation

Nowadays, a large number of people around the world are suffering from mental health conditions caused by abuse, the lifestyle of an individual, etc., hampering their behavior, physical health, and daily life [1][2]. It can also lead to suicide in worst-case scenarios. Traditional diagnosis methods of mental health issues include several questionnaires, but they are subjective in nature [3]. Nowadays, several researchers are leveraging the non-verbal clues of the patient along with Artificial Intelligence for mental health assessment. A large number of works in the literature are focused mainly on machine learning and deep learning-based solutions.

1.1 Project Objectives

This project aims to develop an audio and text-based LLM for mental health recognition and incorporate the visual modality into the same LLM model.

1.2 Objective of the Mid-Project Review 1

This mid-project review 1 report aims to show the inference results of the modified LTU-AS architecture on the benchmark dataset Extended Distress Analysis Interview Corpus Wizard of Oz dataset (E-DAIC) corpus and our custom dataset named Depression Video Dataset (DVD).

2. Strategy

This time, we have changed our evaluation strategy and benchmarked the architecture on a standard proof of concept depression dataset called E-DAIC. We are focusing on predicting the global label of the patient's interview by assessing mental health disorders such as depression from the individual chunks of the patient's interview and getting the majority vote of those sample chunks as the label for those samples. We have used the dev set and test set of the E-DAIC dataset [4] presented during the Audio/Visual Emotion Challenge (AVEC) 2019 Challenge and the complete OS-DVD dataset [as mentioned in the baseline report]. We have not used the training set of the E-DAIC dataset because we will use it to train the model in the future. Earlier, in baseline results, we focused on predicting the mental health disorder on each chunk of audio files of the OS-DVD dataset, whose results are in the baseline report.

This strategy has changed our evaluation of previous baseline results from chunk level to sample level, which has also changed the results. Even in this, we have used the LLAMA-405B ground truths to compare, but this time, we are using the majority voting across chunks of a sample.

3. About E-DAIC Dataset

The E-DAIC dataset [4] is the extended version of the DAIC-WOZ dataset [5][6] consists of 275 interviews of patients suffering from anxiety, depression, or post-traumatic stress Disorder (PTSD), which is divided into three splits, which are training split, validation split, and test split, containing 163, 56, and 56 interviews, respectively. It includes video features such as head pose, eye gaze, Facial Action Units (FAUs) extracted from OpenFace toolkit, audio features such as MFCCs and eGeMaPS extracted from openSMILE toolkit, other features including Bag of Video Words (BoVW), Bag of Audio Words (BoAW), ResNet and VGG features, text transcripts of the patients in the form of .csv files, audio file of the interview in the form of .wav format and metadata containing demographic data of the participants along with severity level of depression in the form of PHQ scores.

For more details, refer to the following:
1.
https://dl.acm.org/doi/abs/10.1145/3347320.3357688?casa_token=_feDR9UfnZwAAAAA:Sr502L_cFlLiC6rCvkLza8iv3XPF554IWkMC0mHTgi_jUj4_OH-mBUkKMtzCS7bz0M9f28XzHyav
2.
https://pub.deadnet.se/ict.usc.edu/pubs/The%20Distress%20Analysis%20Interview%20Corpus%20of%20human%20and%20computer%20interviews.pdf
3. https://kgeorgila.github.io/publications/devault_aamas14.pdf
4. https://arxiv.org/abs/2407.06125

4. About LTU-AS

LTU-AS [7] is an extension of the LTU model [8] proposed by Gong et al., used for both audio and speech understanding using perception and reasoning properties. The architecture consists of three major parts: Whisper (used for extracting audio features and textual transcript), Time and Layer-Wise Transformer (TLTR) (for capturing more audio events and LLaMA (generates responses for natural language questions). These components help LTU-AS to perform several tasks such as audio event recognition, speech context comprehension, open-ended question answering, etc.

For more details, refer to the following:
1. https://ieeexplore.ieee.org/abstract/document/10389742
2. https://arxiv.org/abs/2305.10790

3.https://docs.google.com/document/d/18EccU9BxaHNOJ6_I3mB8dOgzsaMKs4BTmvjBTw7Z_KM/edit?usp=sharing

5. Methodology

We have used the modified LTU-AS model for classification tasks on the E-DAIC dataset. We have used this framework to analyse subtle emotions and transcribe the audio samples from the E-DAIC dataset, which will help with depression assessment tasks.

Steps:

The steps followed in this process are the same as those used during baseline results, with some changes. They are as follows:

1. Splitting the audio files of the E-DAIC and DVD datasets into several chunks of 30 seconds each.
2. As mentioned during baseline results, we have increased the context window size of LTU-AS from the default 10 seconds to 30 seconds for processing the audio files and getting the outputs.

Prompt used for output Generation:

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

3. As mentioned during baseline results, we get the outputs (classes - **Depressed** / **Not Depressed**) from the audio files, create a pipeline for performing sequential inference on all the audio files, save all the results in the .csv file, and perform majority voting to assign the global class labels.

For more details, refer to the following:
1. https://ieeexplore.ieee.org/abstract/document/10389742
2.
https://docs.google.com/document/d/18EccU9BxaHNOJ6_I3mB8dOgzsaMKs4BTmvjBTw7Z_KM/edit?usp=sharing

6. Evaluation and Results:

The generated outputs from the modified LTU-AS model on each chunk of audio files of the development and test set of the E-DAIC dataset and the complete OS-DVD dataset are stored in a .csv file with the naming convention of the audio file name. Then, we conducted the majority vote on the predicted labels of the chunks of the audio files and assigned the global label to complete the audio file, which was evaluated against the actual labels of the dataset provided by the competition's organisers.

6.1 Results:

We have also tried different prompts to infer the modified LTU architecture this time. One of these prompts is:

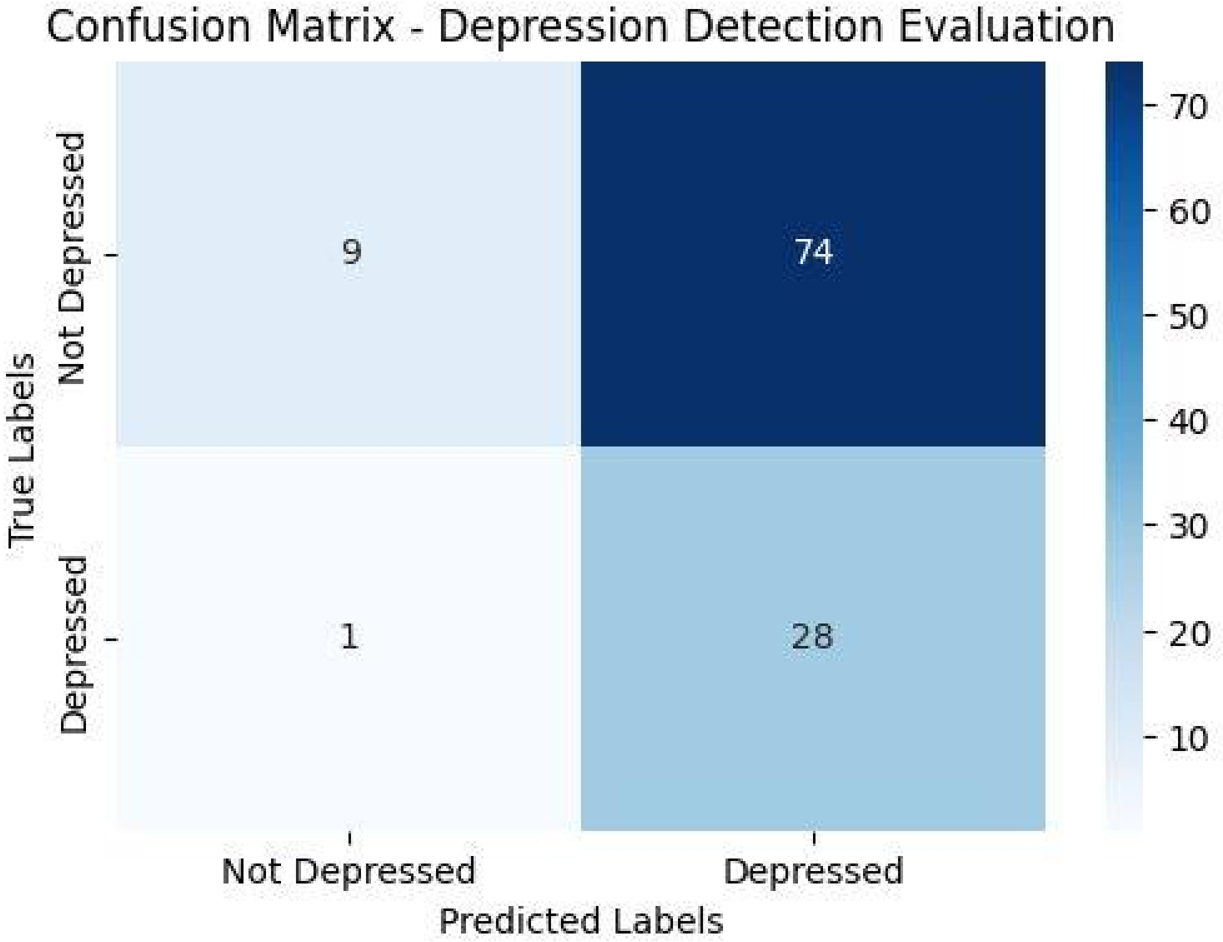
"Based on the conversation in this audio, does the person show symptoms of depression? Provide a one-word answer: Yes if the person is depressed or No if the person is not depressed."

But out of all these prompts, the original prompt worked better. This above prompt fetched an accuracy of around 27%. We also found some hallucinated outputs for this prompt. We have shown the result of the original prompt for these experiments.

6.1.1 Development Set and Test Set of E-DAIC Dataset:

Accuracy: 0.3304
Classification Report:

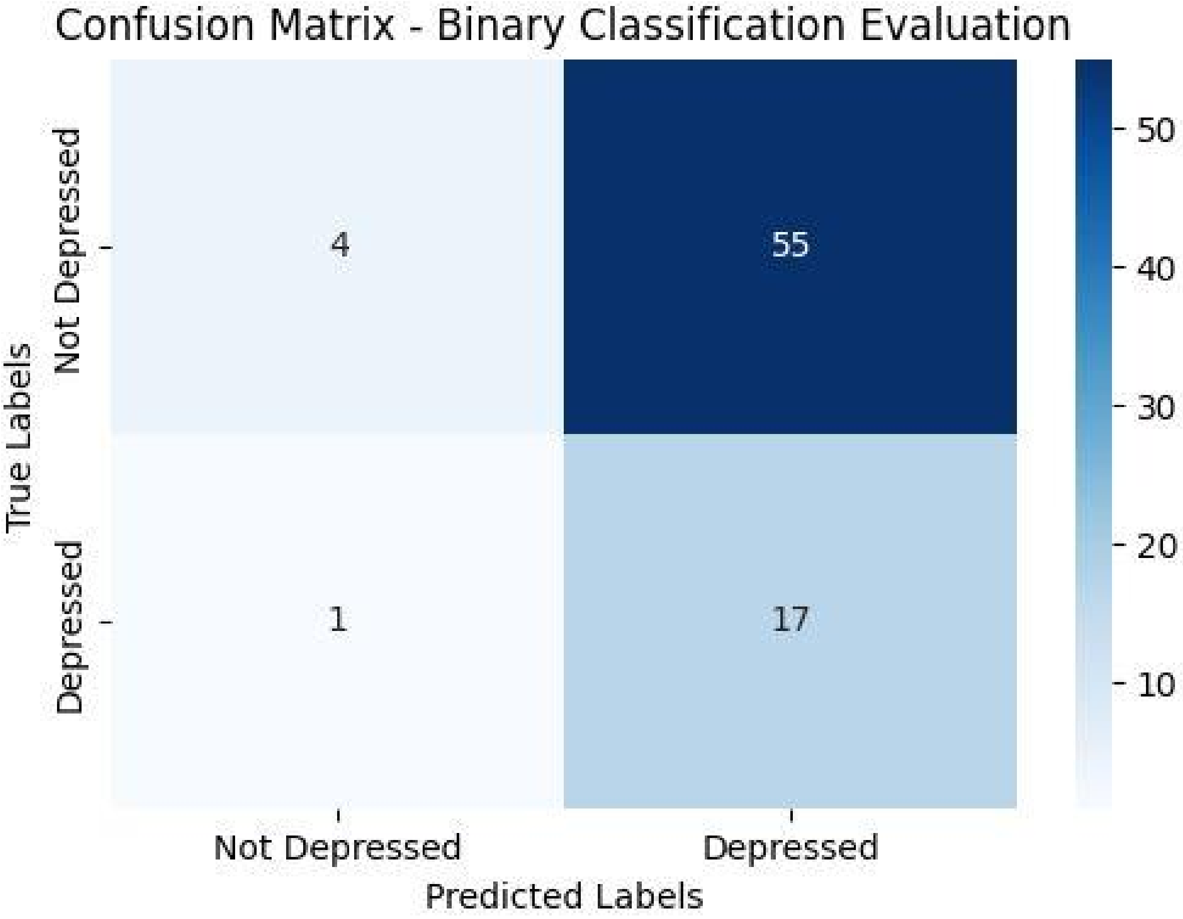
	precision	**recall**	**f1-score**	**support**
Non Depressed	**0.90**	**0.11**	**0.19**	**83**
Depressed	**0.27**	**0.97**	**0.43**	**29**
accuracy			**0.33**	**112**
macro avg	**0.59**	**0.54**	**0.31**	**112**
weighted avg	**0.74**	**0.33**	**0.25**	**112**



6.1.2 OS-DVD Dataset

Accuracy: 0.2727
Classification Report:

	precision	recall	f1-score	support
Not Depressed	**0.80**	**0.07**	**0.12**	**59**
Depressed	**0.24**	**0.94**	**0.38**	**18**
accuracy			**0.27**	**77**
macro avg	**0.52**	**0.51**	**0.25**	**77**
weighted avg	**0.67**	**0.27**	**0.18**	**77**



7. Conclusion and Future Work:

The sample-based experimentation demonstrated an understanding of the marker of the depressed samples better than that of non-depressed samples. In the future, we will improve the policy of majority vote rather than 1:1 mapping; we can try 1:2 mapping of majority vote weights, giving more weight to non-depressed predictions as we noted them far less than the depressed predictions in our architectural inference. We will also modify the architecture to understand the marker of non-depressed samples better. In future work, we will also incorporate the video modality into the architecture in either a complete video or video features format.

References

[1] Luna-Jimenéz, C., Callejas, Z. and Griol, D., 2024, June. Mental-Health Topic Classification employing D-vectors of Large Language Models. In 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS) (pp. 199-204). IEEE.

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

[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] Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.M. and Song, S., 2019, October. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop (pp. 3-12).

[5] Gratch, J., Artstein, R., Lucas, G.M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S. and Traum, D.R., 2014, May. The distress analysis interview corpus of human and computer interviews. In LREC (pp. 3123-3128).

[6] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M. and Lucas, G., 2014, May. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems (pp. 1061-1068).

[7] 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.

[8] Gong, Yuan, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. "Listen, think, and understand." *arXiv preprint arXiv:2305.10790* (2023).