

We Care: Multimodal Depression Detection and Knowledge Infused Mental Health Therapeutic Response Generation

Palash Moon and Pushpak Bhattacharyya
Department of Computer Science and Engineering,
Indian Institute of Technology Bombay
palashmoo@cse.iitb.ac.in

Abstract

The detection of depression through non-verbal cues has gained significant attention. Previous research predominantly centred on identifying depression within the confines of controlled laboratory environments, often with the supervision of psychologists or counsellors. Unfortunately, datasets generated in such controlled settings may struggle to account for individual behaviours in real-life situations. In response to this limitation, we present the Extended D-vlog dataset, encompassing a collection of 1,261 YouTube vlogs. Additionally, the emergence of large language models (LLMs) like GPT3.5, and GPT4 has sparked interest in their potential that LLMs can act like mental health professionals. Yet, the readiness of these LLM models to be used in real-life settings is still a concern as they can give wrong responses that can harm the users. We introduce a virtual agent serving as an initial contact for mental health patients, offering Cognitive Behavioral Therapy (CBT)-based responses. It comprises two core functions: 1. Identifying depression in individuals, and 2. Delivering CBT-based therapeutic responses. Our Mistral model achieved impressive scores of **70.1%** and **30.9%** for distortion assessment and classification, along with a Bert score of **88.7%**. Moreover, utilizing the TVLT model on our Multimodal Extended D-vlog Dataset yielded outstanding results, with an impressive F1-score of **67.8%**.

Disclaimer: It is not the intent of this paper to advocate using large language models (LLMs) in therapeutic settings in real life; the work reported is purely a research demonstration

1 Introduction

Depression is a prevalent and significant medical condition. It hurts one's emotional state, thought processes, and behaviour. It manifests as persistent feelings of sadness and diminished interest in previously enjoyed activities. This condition can give

rise to various emotional and physical challenges, affecting one's ability to perform effectively both at work and in personal life. Depression symptoms range from mild to severe and can include persistent sadness, loss of interest in once-enjoyable activities, appetite changes, sleep disturbances, fatigue, psychomotor changes, feelings of worthlessness, cognitive challenges, and, in severe cases, suicidal thoughts. Symptom severity varies, requiring careful clinical evaluation for diagnosis and treatment (Cleveland Clinic).

Motivation: According to the statistics of the World Health Organisation (WHO) (World Health Organization) 3.8% of the world's population experience depression, including 5% of adults less than 60 years of age (4% of men and 6% of women) and 5.7% of adults above 60 years of age. Approximately 280 million people have depression which depression is 50% more common in women than men. Depression is 10% more in pregnant women and women who have just given birth (Evans-Lacko et al., 2018). If the depression is left untreated, it can lead to several serious outcomes such as suicide (Ghosh et al., 2022).¹ Currently, there is a lack of mental health practitioners globally, with a ratio of 1 : 10000 mental health professionals per patient. Due to the shortage of mental health professionals, we aim to reduce the overreliance on them by automating the prediction of depression and providing therapeutic responses to users, which can mitigate distress to some extent.

Virtual Agent: In the realm of mental health support, the notion of therapy chatbots has intrigued both researchers and the public since the introduction of Eliza (Shum et al., 2018) in the 1960s. Recent advancements in large language models (LLMs) like ChatGPT have further fueled this interest. However, concerns have been raised by mental

¹<https://www.who.int/news-room/fact-sheets/detail/depression>

health experts regarding the use of LLMs for therapy as the therapy provided may not be accurate (Mental; Li et al., 2024; Stade et al., 2023; Sharma et al., 2024). Despite this, many researchers have begun exploring LLMs as a means of providing mental health support (Sharma et al., 2023).

Our understanding of how LLMs behave in response to clients seeking mental health support remains limited. It is unclear under what circumstances LLMs prioritize certain behaviours, such as reflecting on client emotions or problem-solving, and to what extent (Chung et al., 2023), (Ma et al., 2023). Given the critical nature of mental health support, it is essential to comprehend LLM behaviour, as undesirable actions could have severe consequences for vulnerable clients. Additionally, identifying desirable and undesirable behaviours can inform the adoption and improvement of LLMs in mental health support.

Our Contributions are:

- Extended D-vlog dataset (**Original no. of videos: 961, Total videos** (after adding 300 videos taken from YouTube to the Original dataset): **1261**) which contains videos of various types such as Major depressive disorder, postmortem disorder, anxiety and videos from different age group and gender which was lacking in the original D-vlog dataset. (see Section 3)
- TVLT (Tang et al., 2022) model for depression detection, which outperforms baseline models by **4.3%** and establishes a new benchmark, on the Extended D-vlog dataset. (see section 6)
- Replacing spectrogram with the combination of spectrogram and wav2vec2 (Baevski et al., 2020) features which capture the vocal cues associated with depression more effectively than spectrogram, which further increases the accuracy by **2.2%** resulting in the final F1-score of **67.8%**. (see section 6)
- To the best of our knowledge, this work is the first to propose a virtual agent that delivers therapeutic responses to users using LLM with domain knowledge as an external knowledge base on mental health.

2 Related Work

With the rise in mental health conditions, there is growing interest in detecting depression. However, there is a shortage of datasets for this purpose, largely due to privacy concerns, limiting public

availability. Among the few publicly accessible datasets, the DAIC-WOZ (Gratch et al., 2014) is notable, featuring clinical interviews in text, audio, and video formats, relying on self-reporting via the PHQ-8 questionnaire. Another dataset, the Pittsburgh dataset, primarily contains audio and video clinical interviews. Despite its small size of 189 samples, the DAIC-WOZ (Gratch et al., 2014) remains valuable for research. The AViD-Corpus, used in AVEC 2013 (Valstar et al., 2013) and 2014 (Valstar et al., 2014) competitions, includes video recordings of various activities with self-reporting conducted in the presence of mental health professionals. While these datasets provide insights into depression patterns, their assembly in controlled environments may not fully represent typical behaviours of depressed individuals.

dataset	Modality	# Subjects	# Samples
DAIC-WOZ	A+V+T	189	189
Pittsburg	A+V	49	130
AViD-Corpus	A+V	292	340
D-vlog	A+V	816	961
E-Dvlog	A+V+T	1016	1261

Table 1: Comparison of various Depression datasets with E-Dvlog (Extended D-vlog). Where A: Audio, V: Video, T: Text.

The use of social media for depression detection is increasingly preferred over clinical interviews due to its ability to capture patient’s authentic behaviour. Unlike supervised interviews, social media datasets reveal atypical behaviors exhibited in daily life but fail to capture "in the wild" behavior—those day-to-day activities that laboratory settings miss. "In the wild" datasets for depression detection are often more useful than those collected in laboratory settings because they record natural behaviors and interactions, providing more accurate and applicable data. These datasets encompass a wide range of contexts and a more diverse population, enhancing the generalizability of the findings. Additionally, "in the wild" datasets capture spontaneous and authentic responses, reducing the influence of artificial settings that can distort behavior in laboratory environments (Sawadogo et al., 2024; Nepal et al., 2024; Opoku Asare et al., 2021; Xezonaki et al., 2020). In recent years, depression detection using text from social media has been focused on (Fatima et al., 2019), (Burdisso et al., 2019), (Chiong et al., 2021). Various ap-

proaches have emerged to detect depression using data from platforms like Twitter, Reddit, and Facebook, focusing on textual-based features such as linguistic characteristics. For instance, (Yang et al., 2018) utilized text and tags from micro-blogs in China to extract behavioural features for depression detection. However, there is a growing need to explore video data and multimodal fusion for more comprehensive detection methods.

Multimodal fusion combines various modalities to predict outcomes, and it’s increasingly used for depression detection. (Haque et al., 2018) utilized 3D facial expressions and spoken language features to detect depression. (Yang et al., 2018) integrated text and video features, employing deep and shallow models for depression estimation. (Ortega et al., 2019) proposed an end-to-end deep neural network integrating speech, facial, and text features for emotional state estimation. Although previous studies have explored depression detection using multimodalities, the combination of multimodal transformer with wav2vec2 features and spectrograms remains unexplored despite its potential for superior results.

Virtual Agents: In recent years with the increase in mental health problems, people have started taking emotional support from text-based platforms such as in (Eysenbach et al., 2004), (De Choudhury and De, 2014), (talkelife. co). there is also a rise in empathetic virtual agents (Saha et al., 2022), which impart empathy in their responses by giving motivational responses and responses with hope and reflections which is seen as important to uplift the spirit of an individual who is seeking support. Additionally, efforts have been made to enhance the therapeutic value of these platforms by incorporating insights (Fitzpatrick et al., 2017), (Xie and Pentina, 2022) encouraging exploration through open-ended questioning, and providing guidance and problem-solving techniques, all aimed at aiding users in their healing process.

3 Datasets

The D-vlog dataset (Yoon et al., 2022) is a collection of depression vlogs of various people posted on YouTube. The D-vlog (Yoon et al., 2022) dataset has 961 vlogs in total out of which 505 are categorized as depressive vlogs and 465 are categorized as Non-depressive vlogs. However, the D-vlog dataset (Yoon et al., 2022) has some limitations, such as the dataset majorly having Major Depressive Dis-

order and lacking other disorders such as Bipolar Disorder, Postmortem Disorder, and Anxiety with depression. Which does make the dataset more generalized. So, we extended the D-vlog dataset (Yoon et al., 2022) by adding 300 more vlogs to the D-vlog dataset (Yoon et al., 2022). The dataset now includes more vlogs on various depressive disorders from different age groups, specifically targeting individuals aged 40 and above, as well as a greater number of vlogs from males, which were previously underrepresented. Refer Figure 1. For more details on data collection (Appendix B).

3.1 Dataset Statistics:

The extended D-vlog dataset has 1261 vlogs with 680 depressive vlogs and 590 non-Depressive vlogs as Shown in below Table 2

	Gender	# Samples
Depression	Male	273
	Female	406
Non-Depression	Male	232
	Female	350

Table 2: Extended D-vlog Statistics

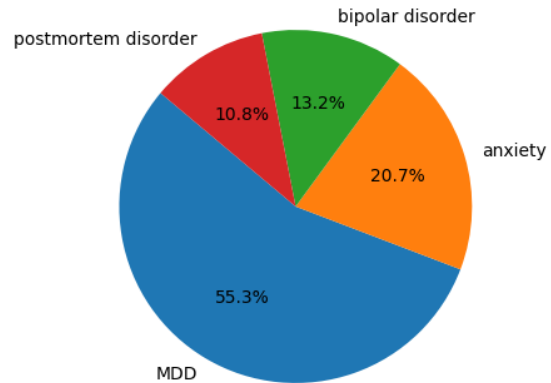


Figure 1: The Above figure shows the distribution of various types of Depressive vlogs. where MDD is Major Depressive Disorder, Bipolar Disorder is also called as Manic Disorder.

3.2 Datasets for Therapeutic Conversations:

Acquiring datasets of therapy conversations poses a significant challenge as they are typically private and rarely shared. Moreover, potential privacy issues may arise when exposing therapy datasets to public LLM APIs as they may contain sensitive client information. Publicly available therapy conversation datasets are limited. Here, we use three

datasets that carefully preprocess publicly available therapy. This ensures high-quality transcripts while maintaining the confidentiality of sensitive personal information. These datasets are 1. High-and-Low-Quality Therapy Conversation Dataset (High-Low Quality) (Pérez-Rosas et al., 2018) 2. HOPE Dataset (Malhotra et al., 2022) 3. Motivate Dataset. We finetune these datasets on the Mistral and LLaMa-2 models to generate therapeutic responses. (Saha et al., 2022). Further details can be seen in the Appendix section B.1.

4 Methodology

The system is comprised of two stages: the first involves detecting depression, and the second focuses on generating a therapeutic response.

Detection: Detection of depression where the video, audio and text are provided as input to the TVLT model for depression detection.

Generation: Provide a therapeutic response to the depressed user. The utterance that was given previously to detect depression. The same text utterance will be fed to the virtual assistant to find the type of distortion classification and after that, we generate the responses.

4.1 Detection

We use **TVLT** (Textless Vision Language Transformer) (Tang et al., 2022), an end-to-end vision and language multimodal transformer model that takes raw video, raw audio, and text as input to the transformer model. TVLT (Tang et al., 2022) is a textless model, which implicitly does not use text, but with the ASR model (whisper) (Radford et al., 2023), we can extract text from the audio segments. The TVLT model is more effective for multimodal classification because the TVLT (Tang et al., 2022) model can capture visual and acoustic information, providing a more comprehensive fused representation of video, audio, and text.

Textual Feature: We make use of the powerful BERT (Kenton and Toutanova, 2019) language model, a pre-trained model described to capture important features from the text. This means we can understand not only the specific details in the text but also the overall context. These BERT embeddings help us understand text thoroughly, making them perfect for tasks like analyzing sentiment or identifying depression. We apply BERT (Kenton and Toutanova, 2019) to our text to get textual embeddings, using specific dimensions ($dt = 786$)

where dt : the dimension of the text.

Audio features: We use a combination of techniques to analyze audio. Firstly, we generate spectrograms using the librosa library (McFee et al., 2015) and extract low-level features. Additionally, we incorporate features from wav2vec2, which is described in (Baevski et al., 2020). The wav2vec2 features include various acoustic attributes such as MFCC (Hossain et al., 2010), spectral (Pachet and Roy, 2007), temporal (Krishnamoorthy and Prasanna, 2011), and prosody (Olwal and Feiner, 2005) features. These features help with identifying the pitch, intonation, and tempo of the audio segment. They are excellent at capturing both local and contextual information from the raw audio waveform. Finally, we compute the average across the spectrogram vector and the wav2vec2 vector to create our final audio representation.

Video Features: Our video processing pipeline involves several essential steps. First, we load the video file using a tool called VideoReader (Frith et al., 2005). Next, we randomly select a subset of frames from the video clip. These frames are then resized and cropped to focus on the subject's frontal view. For extracting visual features, we rely on the powerful ViT (Vision Transformer) model introduced in (Dosovitskiy et al., 2020). This model helps us create what we call "vision embeddings." It does this by breaking down each video frame into smaller 16x16 patches. We then apply a linear projection layer to these patches, resulting in a 768-dimensional patch embedding. This vision embedding module is a critical component of our model. It transforms each video frame or image into a sequence of 768-dimensional vectors.

We have implemented the architecture illustrated in Figure 2, where our TVLT (Tang et al., 2022) transformer model comprises a 12-layer encoder and an 8-layer decoder. To obtain the fused representation of all three modalities, We exclusively utilize the encoder portion of the model to generate fused representations for depression prediction tasks. Our evaluation, conducted on the extended D-vlog (Yoon et al., 2022) dataset comprising 1261 video clips from 1016 speakers, involves transcription using an ASR model with manual error correction. We split the data into a 7:1:2 train-validation-test ratio and employ weighted accuracy (WA) and F1-score metrics. Additionally, we add task-specific heads on top of the encoder representation and train the model using binary cross-entropy

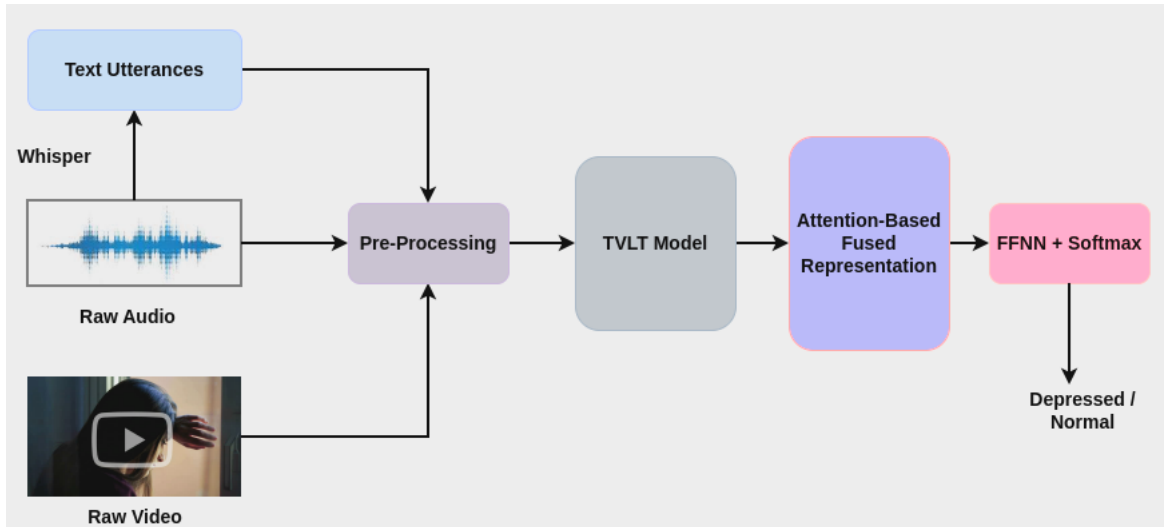


Figure 2: In the Above **Architecture** we leverage three different modalities such as video, audio and text where text is extracted from the audio segment using the Whisper ASR Model. We then preprocess all three modalities and pass them to the model where we get the fused representation of all three modalities. This fused representation is then passed to the feed-forward Neural Network with a softmax function to determine whether the individual exhibits signs of Depression or is in a Normal state.

loss for each downstream task.

$$L(y, \hat{y}) = - [y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \quad (1)$$

where y : True label and \hat{y} : Predicted label.

4.2 Prompting with LLMs:

Cognitive distortions are irrational or biased ways of thinking that negatively impact emotions and behavior. During our discussion with the psychologist, we learned that identifying the ABCs—Activation Events, Beliefs, and Consequences—is crucial for recognizing and categorizing these distortions. Identifying these distortions in depressed individuals is key to improving mental health by changing negative thought patterns (Ota et al., 2020; Kendall et al., 1990; Wang et al., 2023; Lefebvre, 1981).

- **Activation Event (A):** Identifies situations triggering emotional responses.
- **Beliefs (B):** Refers to the patient’s thoughts about the activating event.
- **Consequences (C):** Indicates the impact on the individual’s life.

The ABC model breaks down triggers, thoughts, and effects, helping to develop strategies to challenge and change negative thoughts. This approach

is essential for improving emotional and behavioral outcomes and is widely used in cognitive-behavioral therapy (CBT).

Analyzing the ABCs makes it easier to understand the distortions and their underlying reasons (Dryden, 2012), (Lam, 2008). Identifying these distortions and the reasons behind them can help challenge the distorted beliefs by asking why the individual feels that way, and reassuring them that these beliefs are normal. Ultimately, this can lead to a therapeutic response such as cognitive reconstruction. To determine whether ABC’s generated are correct we performed a human evaluation on 200 samples. Additional information is provided in Appendix D. After generating the ABCs, we input them along with an additional few shot prompts to the Mistral-7B-Instruct-v0.2 model². By doing so, we determine whether the assessment exhibits cognitive distortion and identify the specific type of distortion present. This process enables us to offer the appropriate therapeutic response to the user based on the type of distortion identified.

We use the RAG (Lewis et al., 2020) pipeline incorporating domain-specific documents as an external knowledge base. This external knowledge is employed to validate and correct the responses generated and fine-tunes the mistral model (Jiang et al., 2023) which was fine-tuned on the motivate

²<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

(Saha et al., 2022) and hope (Malhotra et al., 2022) dataset. While generating responses to user queries, we utilize a system prompt as given in (Appendix C.1)

5 Experiments

Detection: To obtain a fused representation of audio, video and text modalities, we employ trained text-based TVLT (Tang et al., 2022) model on the video dataset and subsequently fine-tune on the extended D-vlog (Yoon et al., 2022) dataset. We split our dataset into train, valid and test sets in the ratio of 7:1:2. Details are in the Appendix B.2

Distortion Identification: We employ the "Mistral-7B-Instruct-v0.2" model to prompt and determine two things directly: firstly, whether an individual exhibits cognitive distortions based on provided context, and secondly, if so, to identify the specific type of cognitive distortion present on the extended D-vlog test dataset. We use few-shot chain-of-thought (Wei et al., 2022) techniques to identify cognitive distortions, including ABC (Dryden, 2012) prompts and pinpointing the distorted parts. Additionally, we explore providing reasoning for the distorted portions identified. Prompt details are given in the Appendix C.2. Results can be seen in Table 7.

Response Generation: We use pre-trained Mistral-7B models (Jiang et al., 2023), fine-tuned on hope (Malhotra et al., 2022) and motivation data (Saha et al., 2022), employing PEFT QLoRA (Detmers et al., 2024)- a method that combines 4-bit quantization with low-rank adapters for improved memory usage and computational efficiency—to generate therapeutic responses. Additionally, we implemented a chain-of-thought (Wei et al., 2022) with an (Lewis et al., 2020) RAG pipeline to ensure accurate responses without generating false information, utilizing Adam’s Optimizer with a learning rate set to 0.00025, known for its superior results. we leverage the Mistral-7b as a Large language model, utilizing the pre-trained RAG model "thenlper/gte-large" from the Hugging Face library. The chunk size used here is 256 and employs the vectorStoredIndex as an indexing mechanism for the storage and retrieval of embeddings from documents.

6 Result and Discussion

In this section, we will cover the results on the extended D-vlog Dataset (Yoon et al., 2022), the

clinical Diac-woz dataset and results on distortion classification and response generation.

6.1 Result on extended D-vlog dataset

To analyse the importance of each modality for depression detection, we trained our model on each modality separately and reported the results in Table 3 below. We discovered that the audio modality

Modalities	F1-scores
T	0.572
A	0.601
V	0.567
V + A	0.631
V + T	0.628
A + T	0.634
V + A + T	0.656

Table 3: Results obtained on the Extended D-vlogs dataset via experiments with Video (V), Audio (A), Textual (T).

outperforms other modalities in terms of F1-Score, indicating its significance in depression detection. This suggests that individuals with depression exhibit distinct speech patterns. Although audio features outweigh visual ones, combining both modalities results in superior performance compared to using audio alone. Additionally, combining audio and text modalities surpasses using audio alone. Finally, incorporating all three modalities yields the best results, highlighting the effectiveness of considering audio, visual, and textual features and their interactions in depression detection.

Modalities	F1-scores
V + A + T	0.656
V + A + T(Mask)	0.663
V + A(W2V2+Spect) + T	0.678
V(Mask) + A + T	0.661

Table 4: Results obtained on the Extended D-vlogs dataset via experiments with Video (V), Audio (A), Textual (T). T(Mask) is text with word-masking, V(Mask) is Video frames with frame-masking and A(W2V2+Spect) is Audio with wav2vec2 +spectrogram features.

Applying frame masking to video data, alongside audio and text modalities, enhances performance by 0.005. These results highlight the effectiveness of incorporating diverse modalities. The table (Table 5) underscores the importance of leveraging text, video, and audio modalities with wav2vec2

Model Type	Model	Precision	Recall	F1-Score
Fusion Baseline	Concat	62.51	63.21	61.1
	Add	59.11	60.38	58.1
	Multiply	63.48	64.15	63.09
Depression Detector	Cross-Attention	65.4	65.5	65.4
Our Model	TVLT Model	67.3	68.3	67.8

Table 5: Performance comparison of our model and various baseline models on the extended D-vlog dataset for depression detection.

(Baevski et al., 2020) features and spectrograms, leading to an impressive F1-score of 67.8%.

We extensively evaluated the TVLT (Tang et al., 2022) model’s performance on the D-vlog (Yoon et al., 2022) dataset, comparing it with several baseline models to gauge its effectiveness in depression detection. The TVLT (Tang et al., 2022) model outperformed the Cross Attention State-of-the-Art model by 2.2%, establishing itself as the new benchmark for the D-vlog (Yoon et al., 2022) dataset.

6.2 Result on the clinical dataset: DAIC-WOZ

We tested our proposed model for depression detection on the clinically labelled DAIC-WOZ dataset, using the same feature extraction process. We con-

Train	Test	Precision	Recall	F1-score
DW	DV	62.14	62.38	62.26
DV	DV	66.40	66.57	66.48
DW	DW	64.57	54.63	59.19
DV	DW	69.45	57.26	62.77

Table 6: Results between extended D-Vlog and DAIC-WOZ datasets. DV and DW denote D-Vlog and DAIC-WOZ, respectively

ducted four experiments with our model, including training and testing with extended D-Vlog, training with DAIC-WOZ and testing with extended D-Vlog, training and testing with DAIC-WOZ, and training with extended D-Vlog and testing with DAIC-WOZ. The results showed that the model trained with extended D-Vlog improved depression detection performance in both datasets. This suggests that D-Vlog’s features, captured in daily life, are more useful than those in the DAIC-WOZ dataset, developed in a laboratory setting.

Methods	DA F1-W	DC F1-W
Mistral	62.4	21.5
Mistral+FCOT	63.9	22.3
Mistral+FCOT+ABC	65.6	27.8
Mistral+FCOT+ABCD	67.3	29.0
Mistral+FCOT+ABCDCR	70.1	30.9
ChatGPT + FCOT + ABC	57.6	20.4
ChatGPT + FCOT + ABCD	59.1	21.0
ChatGPT + FCOT + ABCDCR	63.5	23.6

Table 7: DA: Distortion Assessment, DC: Distortion Classification, F1-W: F1-weighted, Fcot: Few-shot chain-of-thought, A: Activation Event, B: Belief, C: Consequences, D: Distorted Part, RAG: Retrieval Augmented Generation

6.3 Results of cognitive distortion:

We utilize the Mistral-7b (Jiang et al., 2023) model to assess the F1-score for distortion assessment and classification across the ten types of distortion. Notably, we discover that integrating the ABC (Dryden, 2012) framework from Cognitive Behavioral Therapy (CBT), which identifies Activation Event (A), Beliefs (B), and Consequences (C), notably enhances both the F1-score for distortion assessment and classification. Furthermore, by identifying the distorted segment within the context and providing reasoning behind its classification, we further enhance the F1-score for assessment and classification to 70.1 and 30.9, respectively. We compared results from Instruct-Mistral-v.02 with ChatGPT and found Mistral performed well for distortion assessment and classification. Mistral could identify distorted parts while ChatGPT couldn’t discern them from context. One of the reasons could be that the mistral-7b model was instruction finetuned on the hugging-face repository having distortion-

specific dataset. whereas the chatGPT is designed to be a general purpose for conversational AI.

Ablation Study: We explore various settings

Methods	DA F1-W	DC F1-W
Mistral+FCOT+A	51.9	20.6
Mistral+FCOT+B	65.0	22.0
Mistral+FCOT+C	63.4	23.5

Table 8: DA: Distortion Assessment, DC: Distortion Classification, F1-W: F1-weighted, Fcot: Few-shot chain-of-thought, A: Activation Event, B: Belief, C: Consequences

to demonstrate the effectiveness of incorporating ABCs and see the impact of each on the assessment and classification, which shows that the beliefs and consequences are important measures for cognitive distortion. As shown in the Table 8.

6.4 Results on Response generation:

We have devised a prompt employing Cognitive Behavioral Therapy (CBT) techniques to craft therapeutic responses. Using the Mistral (Jiang et al., 2023) prompt, we generate responses and compare them to the therapist’s provided ground truth. Our analysis reveals a semantic similarity of 88.7% and 86.7% with Mistral (Jiang et al., 2023) and LLaMa-2 (Touvron et al., 2023) prompts respectively. This indicates that the generated responses closely align with the ground truth, affirming their semantic similarity.

Models	BLEU-4	ROUGE-L	BERT Score
Mistral	25	23.5	88.7
LLaMA	21.6	18.8	86.7

Table 9: Results on Bleu-4 score, Rouge-L score, and BertScore were evaluated for both the fine-tuned Mistral (Jiang et al., 2023) model and the LLaMA model.

Human Evaluation: We requested evaluations from 2 annotators to assess therapeutic responses generated by the Mistral-7b and LLaMA-2-7b models based on three criteria: Reflective Inquiry, Challenging Thoughts, and Cognitive Restructuring. Each criterion was rated on a scale of 1 to 10, and average scores were calculated for each model. Table 10 presents the scores for the Mistral-7b model, where Annotator-1 (A-1) and Annotator-2 (A-2)

provided their ratings. Similarly, Table 11 displays the scores for the LLaMA-2-7b model, with ratings provided by the same annotators. These evaluations offer insights into how each model performed across the specified therapeutic criteria as judged by human annotators.

Criterion	A-1	A-2
Reflective Inquiry	7.9	8.8
Challenging Thoughts	8.0	8.9
Cognitive Restructuring	8.9	8.0

Table 10: The Human Evaluation scores from annotators for the therapeutic responses generated by the Mistral-7b model, with A-1 representing Annotator-1 and A-2 representing Annotator-2.

Criterion	A-1	A-2
Reflective Inquiry	8.1	8.9
Challenging Thoughts	7.9	8.1
Cognitive Restructuring	8.3	7.9

Table 11: The Human Evaluation scores from annotators for the therapeutic responses generated by the LLaMA-2-7b model, with A-1 representing Annotator-1 and A-2 representing Annotator-2.

7 Summary, Conclusion & Future Work

This study introduced an Extended D-vlog dataset with 1261 videos, including vlogs by individuals with depression (680 videos) and those without (590 videos). Our goal is to detect depression in non-verbal and non-clinical vlogs using a TVLT (Tang et al., 2022) model, a multimodal transformer. We utilized text, video, and audio data, with visual embeddings from the Vit model and audio features from wav2vec2 (Baevski et al., 2020) and spectrograms. The TVLT (Tang et al., 2022) model, incorporating all modalities, achieved an F1-score of 65.6, which improved to 66.3 with text word masking and 66.1 with frame masking. Our TVLT model, along with wav2vec2 (Baevski et al., 2020) and spectrogram features, outperformed all baseline models on the D-vlog dataset and set a new benchmark on the Extended D-vlog dataset. Our Mistral model achieved impressive F1 scores of **70.1** and **30.9** for distortion assessment and classification, along with a Bert score of **88.7**. In future, we are planning to create an LLM-based Psychologist Agent to converse with the user.

References

- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.
- Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.
- Raymond Chiong, Gregorious Satia Budhi, and Sandeep Dhakal. 2021. Combining sentiment lexicons and content-based features for depression detection. *IEEE Intelligent Systems*, 36(6):99–105.
- Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311.13857*.
- Cleveland Clinic. [Depression](#).
- Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Windy Dryden. 2012. The “abcs” of rebt i: A preliminary study of errors and confusions in counselling and psychotherapy textbooks. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, 30:133–172.
- Sara Evans-Lacko, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, and et al. 2018. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the who world mental health (wmh) surveys. *Psychological Medicine*, 48(9):1560–1571.
- Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *Bmj*, 328(7449):1166.
- Iram Fatima, Burhan Ud Din Abbasi, Sharifullah Khan, Majed Al-Saeed, Hafiz Farooq Ahmad, and Rafia Mumtaz. 2019. Prediction of postpartum depression using machine learning techniques from social media text. *Expert Systems*, 36(4):e12409.
- Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.
- Simon Frith, Andrew Goodwin, and Lawrence Grossberg. 2005. *Sound and vision: The music video reader*. Routledge.
- Saptarshi Ghosh, Asif Ekbal, and Pushpak Bhat-tacharyya. 2022. A multitask framework to detect depression, sentiment, and multi-label emotion from suicide notes. *Cognitive Computation*, pages 1–20.
- Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Reykjavik.
- Albert Haque, Michelle Guo, Adam S Miner, and Li Fei-Fei. 2018. Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*.
- Md. Afzal Hossan, Sheeraz Memon, and Mark A Gregory. 2010. A novel approach for mfcc feature extraction. In *2010 4th International Conference on Signal Processing and Communication Systems*, pages 1–5.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Philip C Kendall, Kevin D Stark, and Therese Adam. 1990. Cognitive deficit or cognitive distortion in childhood depression. *Journal of Abnormal Child Psychology*, 18:255–270.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.
- P Krishnamoorthy and SR Mahadeva Prasanna. 2011. Enhancement of noisy speech by temporal and spectral processing. *Speech Communication*, 53(2):154–174.
- Danny CK Lam. 2008. *Cognitive behaviour therapy: A practical guide to helping people take control*. Routledge.
- Mark F Lefebvre. 1981. Cognitive distortion and cognitive errors in depressed psychiatric and low back pain patients. *Journal of consulting and clinical psychology*, 49(4):517.

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024. Automatic evaluation for mental health counseling using llms. *arXiv preprint arXiv:2402.11958*.
- Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1105. American Medical Informatics Association.
- Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 735–745.
- Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25.
- Knowledge Infused Mental. Health therapeutic response generation.
- Subigya Nepal, Arvind Pillai, Weichen Wang, Tess Griffin, Amanda C Collins, Michael Heinz, Damien Lekkas, Shayan Mirjafari, Matthew Nemesure, George Price, et al. 2024. Moodcapture: Depression detection using in-the-wild smartphone images. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18.
- Alex Olwal and Steven Feiner. 2005. Interaction techniques using prosodic features of speech and audio localization. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 284–286.
- Kennedy Opoku Asare, Aku Visuri, Julio Vega, and Denzil Ferreira. 2021. Me in the wild: An exploratory study using smartphones to detect the onset of depression. In *International Conference on Wireless Mobile Communication and Healthcare*, pages 121–145. Springer.
- Juan DS Ortega, Mohammed Senoussaoui, Eric Granger, Marco Pedersoli, Patrick Cardinal, and Alessandro L Koerich. 2019. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint arXiv:1907.03196*.
- Maki Ota, Shinya Takeda, Shenghong Pu, Hiroshi Matsumura, Takayuki Araki, Naoko Hosoda, Yoko Yamamoto, Aya Sakahihara, and Koichi Kaneko. 2020. The relationship between cognitive distortion, depressive symptoms, and social adaptation: A survey in japan. *Journal of affective disorders*, 265:453–459.
- François Pachet and Pierre Roy. 2007. Exploring billions of audio features. In *2007 international workshop on content-based multimedia indexing*, pages 227–235. IEEE.
- Verónica Pérez-Rosas, Xueting Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. Analyzing the quality of counseling conversations: the tell-tale signs of high-quality counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.
- Tulika Saha, Saichethan Reddy, Anindya Das, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A shoulder to cry on: towards a motivational virtual assistant for assuaging mental agony. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 2436–2449.
- Moctar Abdoul Latif Sawadogo, Furkan Pala, Gurkirat Singh, Imen Selmi, Pauline Puteaux, and Alice Othmani. 2024. Ptsd in the wild: a video database for studying post-traumatic stress disorder recognition in unconstrained environments. *Multimedia Tools and Applications*, 83(14):42861–42883.
- Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.
- Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–29.
- Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19:10–26.
- Elizabeth Stadel, Shannon Wiltsey Stirman, Lyle H Ungar, David Bryce Yaden, H Andrew Schwartz, João Sedoc, Robb Willer, Robert DeRubeis, et al. 2023. Artificial intelligence will change the future of psychotherapy: A proposal for responsible, psychologist-led development.
- Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. 2022. Tvl: Textless vision-language transformer. *Advances in Neural Information Processing Systems*, 35:9617–9632.

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Michel Valstar, Björn Schuller, Kirsty Smith, Timur Al-maev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 3–10.
- Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10.
- Bichen Wang, Yanyan Zhao, Xin Lu, and Bing Qin. 2023. Cognitive distortion based explainable depression detection and analysis technologies for the adolescent internet users on social media. *Frontiers in Public Health*, 10:1045777.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- World Health Organization. [Depression](#).
- Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews. *arXiv preprint arXiv:2006.08336*.
- Tianling Xie and Iryna Pentina. 2022. Attachment theory as a framework to understand relationships with social chatbots: a case study of replika.
- Le Yang, Dongmei Jiang, and Hichem Sahli. 2018. Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures. *IEEE Transactions on Affective Computing*, 12(1):239–253.
- Leon Yin and Megan Brown. 2018. [Smappnyu/youtube-data-api](#).
- Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234.

A TVLT Model:

Textless Vision-Language Transformer (TVLT), is a model designed for vision-and-language representation learning using raw visual and audio inputs. Unlike traditional approaches, TVLT employs homogeneous transformer blocks with minimal modality-specific design and does not rely on text-specific modules such as tokenization or automatic speech recognition (ASR). TVLT is trained using masked autoencoding to reconstruct masked patches of continuous video frames and audio spectrograms, as well as contrastive modelling to align video and audio. Experiments demonstrate that TVLT achieves comparable performance to text-based models across various multimodal tasks, including visual question answering, image retrieval, video retrieval, and multimodal sentiment analysis. Additionally, TVLT offers significantly faster inference speed (28x) and requires only one-third of the parameters. These results suggest the feasibility of learning compact and efficient visual-linguistic representations directly from low-level visual and audio signals, without relying on pre-existing text data.

A.1 Pretaining details:

- **HowTo100M:** We used HowTo100M, a dataset containing 136M video clips of a total of 134,472 hours from 1.22M YouTube videos to pre-train our model. Our vanilla TVLT is pre-trained directly using the frame and audio stream of the video clips. Our text-based TVLT is trained using the frame and caption stream of the video. The captions are automatically generated ASR provided in the dataset. We used 0.92M videos for pretraining, as some links to the videos were invalid to download.
- **YTTemporal180M:** YTTemporal180M includes 180M video segments from 6M YouTube videos that spans multiple domains, and topics, including instructional videos from HowTo100M, lifestyle vlogs of everyday events from the VLOG dataset, and YouTube’s auto-suggested videos for popular topics like ‘science’ or ‘home improvement’.

B Dataset Collection:

We have collected the dataset vlogs using certain keywords using YouTube API (Yin and Brown,

2018) and downloaded them using the yt-dlp package³.

Depressive vlogs: ‘depression daily vlog’, ‘depression journey’, ‘depression vlog’, ‘depression episode vlog’, ‘depression video diary’, ‘my depression diary’, and ‘my depression story’, ‘postpartum depression vlogs’, ‘Anxiety vlogs’.

Non-Depressive vlogs: ‘daily vlog’, ‘grwm (get ready with me) vlog’, ‘haul vlog’, ‘how to vlog’, ‘day of vlog’, ‘talking vlog’, etc.

We used the same approach to collect the dataset as used in the D-vlog dataset (Yoon et al., 2022). We focused our analysis on vlogs featuring content creators who have a documented history of depression, currently manifesting symptoms of the condition. We specifically excluded vlogs that solely discussed having a bad day without a deeper connection to depressive experiences.

B.1 Datasets for Therapeutic Conversations:

1. **High-and-Low-Quality Therapy Conversation Dataset (High-Low Quality):** The initial dataset, established by (Pérez-Rosas et al., 2018), encompasses 259 therapy dialogues, predominantly centring on evidence-based motivational interviewing (MI) therapy. Assessing the conversations by MI psychotherapy principles, the authors identify 155 transcripts of high quality and 104 of low quality within the dataset. Both high-quality and low-quality therapy dialogues conducted by human therapists are utilized to examine desirable and undesirable conversational behaviours.
2. **HOPE Dataset:** The second dataset from (Malhotra et al., 2022) was used to study dialogue acts in therapy. This dataset contains 212 therapy transcripts and includes conversations employing different types of therapy techniques (e.g., MI, Cognitive Behavioral Therapy).
3. **MotiVAte Dataset:** The MotiVAte Dataset (Saha et al., 2022) contains 7076 dyadic conversations with support seekers who have one of the four mental disorders: MDD, OCD, Anxiety or PTSD.

³<https://github.com/yt-dlp/yt-dlp/wiki/Installation>

B.2 Experiments setup: Detection

Gender	Train	Valid	Test
Male	354	51	100
Female	530	74	152

Table 12: Number of vlogs in Train, Valid and Test Split of Extended D-vlog dataset

For training the model, we utilized Adam’s Optimizer with learning rates ranging from 0.0001 to 0.00001 and batch sizes of 32 and 64. The model underwent four iterations with different seed values, each taking approximately three hours to train on the Nvidia RTX A6000. Binary cross-entropy served as our chosen loss function for the depression detection task, and F1 scores were reported based on the test set results.

The extended D-vlog dataset exhibits more representation of Female vlogs as compared to Male vlogs within the Depressed category, reflecting a high prevalence of depression among Females. In the non-depressive category similar trend is observed with more female representation than male vlogs as predominantly "get ready with me vlogs", and "Haul vlogs" are uploaded by females.

C Prompt Details

C.1 Response Generation Prompt

"Act like a mental health therapist skilled in Cognitive Behavior Therapy (CBT). Your client presents a cognitive distortion, and your task is to guide them towards healthier thinking. Your response should involve three key steps: 1. Reflective Inquiry: Acknowledge the distortion without judgment, exploring it with empathy and understanding. 2. Challenging Thoughts: Gently question the distorted thinking, uncovering its roots and promoting alternative perspectives. 3. Cognitive Restructuring: Offer practical strategies for reframing thoughts and fostering self-compassion, empowering your client to reshape their mindset."

C.2 Prompt for Identification of Distortion

"you are a mental health therapist who uses Cognitive Behavioral Therapy (CBT) to give responses. Understand the following definitions: Activating Event: This represents the specific situations or events that trigger emotional responses. Beliefs: These are the patient’s thoughts and

interpretations regarding the mentioned Activating Event. Consequences: What effect has happened due to the Activating Event on a person’s life? Cognitive Distortion: A cognitive distortion is an exaggerated or irrational thought pattern involved in the onset or perpetuation of psychopathological states, such as depression and anxiety. Your task is to use Cognitive Behavioral Therapy (CBT), analyze the given question to identify the Activating Event, Belief, Consequences, Distortion Part in the Question. Follow the steps below: 1. Identify the Activating Event: Pinpoint the specific situation triggering emotional responses. 2. Explore the Belief: Examine underlying thoughts, distinguishing between Rational and Irrational beliefs. Tell if it has Rational Belief or Irrational Belief. 3. Assess the Consequences: Evaluate emotional, behavioural, and physiological outcomes resulting from beliefs. 4. Identify the Distorted part or sentence from the Question itself if present else none. 5. Using Question, Activation Event, Belief, Consequences and Distorted Part identify the Cognitive Distortion category from the above types if present else indicate none. 6. Give a reason why you choose for a particular Cognitive Distortion and why not for the other Cognitive Distortion. 6. Provide an Assessment: if the case of Cognitive distortion provides yes else no. The Assessment should be "yes" or "no" only. Followed by the types of cognitive distortion

1. Emotional Reasoning: Believing “I feel that way, so it must be true”.
2. Overgeneralization: Concluding with limited and often unnegative experience.
3. Mental Filter: Focusing only on limited negative aspects and not the excessive positive ones.
4. Should Statements: Expecting things or personal behaviour should be a certain way.
5. All or Nothing: Binary thought pattern. Considering anything short of perfection as a failure.
6. Mind Reading: Concluding that others are reacting negatively to you, without any basis.
7. Fortune Telling: Predicting that an event will always result in the worst possible outcome.

8. Magnification: Exaggerating or Catastrophizing the outcome of certain events or behaviour.
9. Personalization: Holding oneself personally responsible for events beyond one’s control.
10. Labeling: Attaching labels to oneself or others (ex: “loser”, “perfect”)."

D Human Evaluation

We used only one human evaluator who has an idea about cognitive distortion and its types and we have shared with him 200 sample forms for the evaluation of the correctness of A: Activation event, B: beliefs, C: Consequences, D: Distorted part. The Percentage tells what percentage of time the model has given the correct value of the activation event, Beliefs, Consequences and distortion.

Models	Percentage
Activation Event	68%
Beliefs	52.5%
Consequences	63.2%
Distorted	41.2%

Table 13

E Qualitative Analysis

In this section, we demonstrate how the integration of wav2vec2 (Baevski et al., 2020) features significantly enhances our TVLT (Tang et al., 2022) model’s ability to accurately detect depression. By gaining a deeper understanding of vocal cues embedded in the audio data, our model’s performance is notably improved. In contrast to the TVLT (Tang et al., 2022) model relying solely on spectrogram data, the inclusion of wav2vec2 (Baevski et al., 2020) features equips our model to make predictions that would be challenging otherwise. The provided table comprises carefully chosen examples where our model correctly identifies depression. In the first instance, the girl’s facial expressions exhibit relative consistency i.e. their not much change in her facial expression. However, the analysis of her audio contains a monotone tone, and low pitch and is filled with cries also the textual utterance is a clear example that she is distressed. The TVLT (Tang et al., 2022) model, enhanced with wav2vec2 (Baevski et al., 2020) and spectrogram features, accurately predict this case, whereas the

model using only spectrogram data falters. This underscores the pivotal role of wav2vec2 (Baevski et al., 2020) features in depression detection this is because wav2vec2 can understand the vocal cues that audio features with spectrogram are failing to understand. In the second example, we observe a girl who simultaneously displays both a smile and tears. The audio content provides a clear indication of her depression, supported by textual information and visual information. However, the model without wav2vec2 (Baevski et al., 2020) features fails to provide an accurate prediction, while the model with wav2vec2 (Baevski et al., 2020) correctly identifies the depressive nature of this example. This highlights the capability of wav2vec2 (Baevski et al., 2020) features to extract crucial information from audio segments, enriching the model’s understanding of depression cues. In the third example, the woman displays minimal to no variation in her facial expressions, and her audio content appears unremarkable, lacking any discernible markers typically associated with depression. However, upon analyzing the textual content, we can identify indications of depression, which our model struggles to predict accurately.

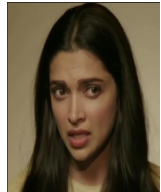
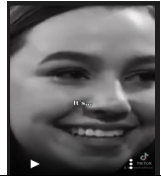
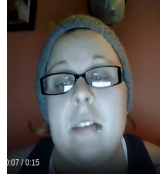
Utterance	Ground Truth	Prediction (w2v2 + spect)	Prediction w/o (w2v2 + spect)	Video frames
I knew what I was feeling, but I don't think I was able to communicate entirely what I was feeling. Like I knew I had this pittish feeling in my stomach. I knew that I'd be scared to wake up. I didn't want to wake up. Yeah, I think waking up was tough because I didn't want to face a day.	Depression	Depression	Normal	
Some days, it's hard to just move. It's... I like it. I, yeah, it's hard to get out of bed. It's hard to even go downstairs to get something to eat.	Depression	Depression	Normal	
No concept of time, no sense of feeling. Have I become cold, dead to the world, where I once mattered? I can't even remember when I was important to someone last. Everything has escaped me. Deeper I fall into a void.	Depression	Normal	Normal	

Table 14: A **Qualitative analysis**, In the given instances, the model, equipped with both wav2vec2 and spectrogram features, effectively detects depression through audio analysis. In the first example, despite seemingly normal facial expressions, the model accurately detects depression. In the second case, the model succeeds in identifying depression even when the individual smiles while crying, whereas the model relying solely on spectrogram data falls short in these situations. In the third scenario, the woman's facial expressions and audio do not exhibit evident signs of depression, while text analysis reveals potential indicators that challenge our model's accuracy, resulting in an incorrect prediction.