EasyChair Preprint

№ 14991

Whispers of Sound: Enhancing Information Extraction from Depression Patients' Unstructured Data Through Audio and Text Emotion Recognition and Llama Fine-tuning

Lin Gan, Xiaoyang Gao, Yifan Huang and Tao Yang

September 21, 2024

Whispers of Sound：Enhancing Information Extraction from Depression Patients' Unstructured Data through Audio and Text Emotion Recognition and Llama Fine-tuning

Lin Gan[1,2]*，Xiaoyang Gao[1], Yifan Huang[1], Tao Yang[1,2]*

[1] Department of Information & Intelligence Engineering, University of Sanya

[2]Department of Academician Chunming Rong Team Innovation

*Correspondence to

Department of Information & Intelligence Engineering, University of Sanya

ABASTRCT

Mental health issues present significant global challenges, affecting over 20% of adults at some point in their lives. While large language models have shown promise in various fields, their application in mental health remains underexplored. This study assesses how effectively these models can be applied to mental health, using the DAIC-WOZ text datasets and RAVDESS audio datasets. Given the challenges of missing non-verbal cues and ambiguous terms in text data, audio data was incorporated during training to address these gaps. This integration enhanced the models' ability to comprehend, extract, and summarize complex information, particularly in depression assessments. Additionally, technical optimizations, such as increasing the model's max_length to 8192, reduced GPU memory usage by 40%-50% and improved context processing, leading to substantial gains in handling complex mental health data.

Keywords: Llama, Fine-tuning, Mental Health, Depression, Audio,Text

1. INTRODUCTION

Depression is a significant global health concern that affects millions of individuals across various demographics, leading to considerable social, economic, and health-related impacts. According to the World Health Organization (WHO), depression is one of the leading causes of disability worldwide, with over 264 million people affected.[1] The condition is associated with decreased productivity, increased morbidity and mortality, and immense personal and societal burdens. Despite its prevalence, depression often remains underdiagnosed and undertreated due to various factors, including stigma and a lack of accessible and effective treatment options. Moreover, the heterogeneity of depression, manifesting through a diverse range of symptoms and severity levels, complicates its early detection and management.

In recent years, the intersection of clinical psychology and technology, particularly in the realms of machine learning (ML) and deep learning (DL), has opened new vistas for understanding, diagnosing depression. These technologies have demonstrated potential in identifying intricate patterns within large datasets, including electronic health records (EHRs), social media interactions, and wearable device data, offering insights into individual and collective mental

[1] Xu, M., Yin, X., & Gong, Y. (2023). Lifestyle Factors in the Association of Shift Work and Depression and Anxiety. JAMA Network Open, 6(8), e2328798. https://doi.org/10.1001/jamanetworkopen.2023.28798

health trends.[2] Research in this area has explored various applications, from predicting depressive episodes based on linguistic and behavioral cues to personalizing treatment recommendations. Despite the significant progress, challenges such as data privacy, model interpretability, and the need for large, annotated datasets continue to pose limitations to the widespread adoption of these technologies in clinical settings.

The advent of Large Language Models (LLMs), such as GPT (Generative Pre-trained Transformer) series by OpenAI, offers groundbreaking opportunities in the field of mental health. These models, trained on extensive corpora of text, can generate human-like text, comprehend context, and engage in meaningful dialogues. In the context of depression, LLMs hold the promise for several applications. Firstly, they could serve as initial screening tools, analyzing speech or written texts to identify linguistic markers indicative of depression. This capability could augment traditional diagnostic processes, enabling early identification of individuals at risk. Secondly, LLMs can provide scalable and personalized support, offering therapeutic interventions or guiding users through cognitive-behavioral therapy (CBT) exercises. Moreover, by analyzing vast amounts of unstructured data from diverse sources, LLMs can contribute to the ongoing research into the etiology and progression of depression, potentially uncovering novel treatment avenues.

Furthermore, the application of LLMs in the medical domain is not limited to mental health. These models are being increasingly employed for medical literature review, patient education, and to enhance the communication between healthcare providers and patients, encapsulating a broader shift towards AI-augmented healthcare services. However, the integration of LLMs into clinical practice necessitates rigorous validation studies, ethical considerations, especially regarding patient confidentiality and data security, and the development of robust regulatory frameworks.

In this study was to identify and extract specific information from conversations between patients with depression and healthcare providers, such as emotional attitudes, topic categorization, and key phrases, to facilitate further analysis or application-oriented research., we downloaded and utilized the DAIC-WOZ (Depression, Anxiety, and Stress Scales-Warwick-Edinburgh Mental Well-being Scale) dataset, which comprises 189 audio files related to depressive episodes. Initially, these audio files were converted into text format. Subsequently, advanced large-scale language models were employed to conduct structured data extraction from the converted texts and identify its emotion.Additionally, we utilized the RAVDESS dataset along with Audio_Speech_Actors_01-24 for audio emotion recognition, implementing multi-method information processing and analysis.

2. RELATED WORK

LLM, These models are trained on vast amounts of text data using techniques derived from deep learning, primarily transformers—a type of neural network architecture that relies heavily on self-attention mechanisms. The working principle behind these models is fundamentally based on probability distribution. They choose subsequent words based on the likelihood calculated from the trained neural network, making predictions more accurate as the network consumes more representative data. This pre-training is followed by fine-tuning, where the model is adapted to specific tasks with task-specific data, which enhances its ability to perform particular functions. Nawshad Farruque et al. delve into the complexities of modeling depression symptoms using

[2] Gan L, Guo Y, Yang T. Machine Learning for Depression Detection on Web and Social Media: A Systematic Review[J]. International Journal on Semantic Web and Information Systems (IJSWIS), 2024, 20(1): 1-28.

social media texts through a state-of-the-art semi-supervised learning (SSL) approach. The study is anchored on the development of a novel SSL framework that initially leverages a large pre-trained language model fine-tuned on a specially curated clinician-annotated dataset for depression symptoms detection (DSD). This model is complemented by a Zero-Shot learning method to effectively capture depression signals from user-generated content. A unique aspect of their research is the iterative retraining of this model with data extracted from a robust, self-curated Depressive Tweets Repository (DTR), compiled from tweets of self-disclosed depressed users. This repository not only preserves a naturalistic distribution of depressive symptoms but also enriches the model's responsiveness to diverse depressive expressions. Their findings reveal substantial improvements in model accuracy for detecting depression, demonstrating the potential of integrating advanced machine learning techniques with clinical insights to enhance mental health diagnostics in digital environments.[3]

Llama, The core of Llama's computing principle lies in its use of self-attention mechanisms that evaluate the relevance of each part of the text to the rest, allowing it to generate highly contextualized outputs. This design enables the model to manage and extract relationships across a vast range of data inputs, which is vital for applications requiring a deep understanding of context, such as in healthcare. In the case of Rotary Positional Encoding (see Figure 1), in the context of Rotary Positional Encoding (RoPE), the conventional method of incorporating positional encodings into the q, k, and v vectors through direct addition or concatenation is supplanted. In RoPE, when either q or k is represented by a two-dimensional vector located at position m, positional information is integrated using rotary coding. This method circumvents the straightforward addition or concatenation of positional encodings, instead embedding positional information directly into the vector representation through rotary coding.

$$RoPE(x_m^{(1)}, x_m^{(2)}, m) = \begin{pmatrix} cosm\theta & -sinm\theta \\ sinm\theta & cosm\theta \end{pmatrix} \begin{pmatrix} x_m^{(1)} \\ x_m^{(2)} \end{pmatrix} = \begin{pmatrix} x_m^{(1)} cosm\theta - x_m^{(2)} sinm\theta \\ x_m^{(2)} sinm\theta + x_m^{(1)} cosm\theta \end{pmatrix} \quad (1)$$

Assuming q or k has two dimensions and is located at position m, positional information is integrated into the vector through rotary encoding here. θ is a constant angle used to define the rotation magnitude for each position. When adding positional information in this manner, the specific steps involved in calculating the attention for q and k are as follows: First, the angle θ is multiplied by the position m to determine the rotation angle for that position. Then, this angle is used to apply a matrix rotation, encoding the positional information into each dimension of q and k.

[3] Farruque, N., Goebel, R., Sivapalan, S. et al. Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach. Lang Resources & Evaluation (2024). https://doi.org/10.1007/s10579-024-09720-4

$$
\begin{gathered}
< RoPE(x_m^{(1)}), x_m^{(2)}), m), RoPE(x_n^{(1)}), x_n^{(2)}), n) >= \\
(x_m^{(1)} cosm\theta - x_m^{(2)} sinm\theta)(x_n^{(1)} cosn\theta - x_n^{(2)} sinn\theta) = \\
(x_m^{(2)} cosm\theta + x_m^{(1)} sinm\theta)((x_n^{(2)} cosn\theta + (x_n^{(1)} cosn\theta) = \\
x_m^{(1)} x_n^{(1)} (cosm\theta cosn\theta + sinm\theta sinn\theta) + \\
x_m^{(1)} x_n^{(2)} (-cosm\theta sinn\theta + sinm\theta cosn\theta) + \\
x_m^{(2)} x_n^{(1)} (-sinm\theta cosn\theta + cosm\theta sinn\theta) + \\
x_m^{(2)} x_n^{(2)} (sinm\theta sinn\theta + cosm\theta cosn\theta) = \\
x_m^{(1)} x_n^{(1)} cos(m-n)\theta + x_m^{(1)} x_n^{(2)} sin(m-n)\theta + \\
-x_m^{(2)} x_n^{(1)} sin(m-n)\theta + x_m^{(2)} x_n^{(1)} cos(m-n)\theta = \\
(x_m^{(1)} cos(m-n)\theta - x_m^{(2)} sin(m-n)\theta) x_n^{(1)} + \\
(x_m^{(2)} cos(m-n)m\theta + x_m^{(1)} sin(m-n)\theta) x_n^{(2)} = \\
< RoPE(x_m^{(1)}), x_m^{(2)}), m-n), RoPE(x_n^{(1)}), x_n^{(2)}), 0)
\end{gathered}
\tag{2}
$$

This method of rotary encoding not only maintains the continuity of the encoding and relative positional information but also ensures stability and efficiency in the computation process due to its multiplicative properties. Unlike traditional addition or concatenation methods, it models positional information directly in the feature space through rotational transformations.

Yunxiang Li and their team recently investigated methods to improve the precision of medical consultations delivered by large language models (LLMs). They modified the LLaMA model by incorporating a dataset containing 100,000 anonymized patient-doctor dialogues from a popular online medical consultation platform. To ensure privacy, all identifying information was removed from these conversations. Furthermore, the researchers added a self-directed information retrieval function to the model, allowing it to consult real-time online sources, such as Wikipedia, along with validated data from offline medical databases. This enhancement enabled the model to source and utilize up-to-date and accurate medical information.[4]

[4] Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus. 2023 Jun 24;15(6):e40895. doi: 10.7759/cureus.40895. PMID: 37492832; PMCID: PMC10364849.

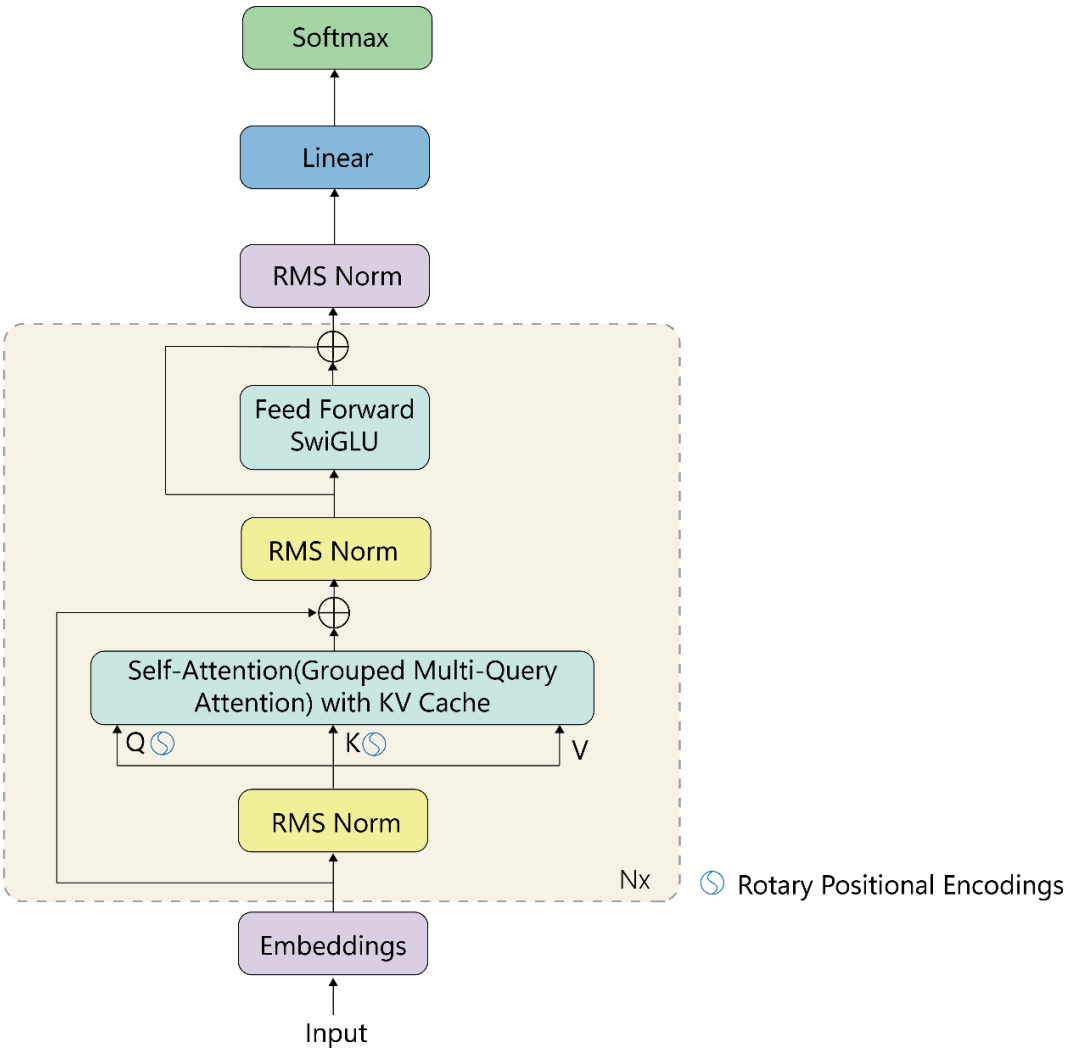


Figure 1. Llama Architecture Diagram

Health care in AI, AI is playing an increasingly important role in patient management and prognosis. In the area of disease diagnosis, AI, particularly through deep learning models such as convolutional neural networks (CNNs), has become capable of efficiently processing and analyzing large volumes of medical imaging data. For instance, Esteva and colleagues (2017) reported in Nature that their model could compete with dermatologists in diagnosing skin cancer. Their research highlights the potential of machine learning models in visual diagnostic tasks.[5] Additionally, De Fauw and others (2018) demonstrated a deep learning system designed for diagnosing eye diseases, capable of automatically identifying clinical abnormalities in retinal optical coherence tomography (OCT) images.[6] Google's DeepMind team has developed an AI system that can estimate kidney function to help doctors predict the risk of acute kidney injury. This tool, by analyzing historical and real-time patient data, supports doctors in making more precise clinical decisions.

---

[5] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb 2;542(7639):115-118. doi: 10.1038/nature21056. Epub 2017 Jan 25. Erratum in: Nature. 2017 Jun 28;546(7660):686. PMID: 28117445; PMCID: PMC8382232.

[6] De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., ⋯ Hughes, C. O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24(9), 1342 – 1350. https://doi.org/10.1038/s41591-018-0107-6

3. THE PROPOSED FRAMEWORK

3.1 DAIC-WOZ dataset

The DAIC-WOZ database[7], a segment of the Distress Analysis Interview Corpus (DAIC), primarily encompasses clinical interview recordings aimed at assisting in diagnosing mental health issues such as anxiety, depression, and Post-Traumatic Stress Disorder (PTSD). This database contains 189 data samples, numbered from 300 to 492. Such interview data are utilized to train computational agents capable of autonomously conducting interviews and identifying mental illnesses through verbal and non-verbal indicators. This dataset includes audio and video records as well as extensive questionnaire responses. A notable feature within the corpus is the inclusion of interviews conducted by 'Ellie', an animated virtual interviewer controlled by a real interviewer from another room. All data have been transcribed and meticulously annotated for both verbal and non-verbal features.

In our experiments, audio data from the DAIC-WOZ database were transformed into text format. These text data were subsequently subjected to thorough cleaning and feature extraction to ensure data quality and usability. Utilizing these preprocessed data, we implemented downstream tasks for information extraction and sentiment analysis. This process not only enhanced the utility of the data but also provided significant technological support for the automated identification and assessment of mental health conditions. Through these advanced analyses, we are able to more deeply understand patients' psychological states, thereby advancing the diagnosis and treatment processes for mental health disorders.

3.2 RAVDESS dataset

The RAVDESS dataset has been widely employed in various fields, including emotion recognition, human-computer interactions, and mental health assessments. It offers valuable support for diverse research pursuits such as the detection of emotional content in speech and the analysis of facial expressions. This study uses the RAVDESS dataset, which contains a total of 7,356 files. The dataset includes recordings from 24 professional actors (12 male, 12 female) with neutral North American accents, who recite two lexically matched statements. The emotions expressed include calm, happy, sad, fearful, angry, surprised, and disgusted, each produced at two levels of emotional intensity and with a neutral expression. For this research, we selected 1,440 audio files from the "Audio_Speech_Actors_01-24" subset and 1,012 audio files from the "Audio_Song_Actors_01-24" subset, totaling 2,452 audio files as the training set. The testing was conducted on files excluding pure video (those with no sound content). All audio samples are stored in standard audio format and have a uniform sampling .[8]

3.3 Text Emotion Analysis

Data Preprocessing

We transcribed 189 pieces of audio from the DAIC-WOZ database into text data, utilizing Feishu Minutes from Lark for real-time audio-to-text conversion. Within the 189 pieces of text data,

---

[7] The DAIC-WOZ database is the Depression Analysis Interview Corpus.Official wensite is https://dcapswoz.ict.usc.edu/

[8] RAVDESS：Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391

certain modal words and verbal fillers or non-verbal vocalizations are present in the conversations, such as 'well,' 'actually,' 'basically,' 'obviously,' 'Umm,' 'Uh,' 'Ah,' and 'Er.' Consequently, during the data cleaning process, we eliminate redundant modal words from the dialogue by employing Regular Expressions in Python.

Algorithm 1 Remove Interjections from Text

Require: $text$ ▷ Input text string
Ensure: $filtered\_text$ ▷ Text string with interjections removed
1: Define a list of interjections: $interjections \leftarrow \{"oh","ah","um","uh","hmm","hey","alas"\}$
2: $pattern \leftarrow$ regular expression pattern for interjections
3: $filtered\_text \leftarrow$ apply regular expression to remove interjections from $text$
4: **return** $filtered\_text$

Example Usage:
5: $sample\_text \leftarrow$ "Oh, I don't know what to do! Uh, can you help me?"
6: $cleaned\_text \leftarrow$ remove_interjections($sample\_text$)
7: Print $cleaned\_text$

3.4 Audio Emotion Analysis

Data Preprocessing

Resampling

To enhance the model's generalization, we resampled the audio data to standardize sampling rates. Resampling adjusts the audio signal's sampling rate, ensuring uniformity during training. This involves interpolation (adding points) and decimation (reducing points). We used the Kaiser window method for FIR filtering, balancing side lobe reduction and main lobe width to control transition bandwidth and peak amplitude distortion. The calculation is as follows:

$$h[n] = \begin{cases} \dfrac{I_0\left(\beta\sqrt{1-\left(\frac{\sin(\pi n/M)}{\sin(\pi\alpha/M)}\right)^2}\right)}{I_0(\beta)} & 0 \le n < M \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Where represents the nth filter coefficient，I_0 is the zero−order modified Bessel function, with a specific value of $\sum_{k=0}^{\infty}\frac{(x/2)^{2k}}{(k!)^2}$，$\beta$ is determined based on the required stopband attenuation and transition bandwidth. α is a parameter that relates the transition bandwidth to the filter length, determining the position of the transition band edges relative to the Nyquist frequency. FIR filter design methods mainly include the window function method and the frequency sampling method.

The window function design method involves selecting an ideal filter and applying a suitable window function to achieve a finite-length impulse response digital system. The key is balancing a narrow main lobe for high frequency resolution and low side lobes to reduce interference.

In audio signal processing, adjusting the main lobe width and side lobe attenuation is critical. A narrow main lobe enhances frequency resolution, suppresses aliasing, and preserves audio details, improving emotion recognition accuracy. Low side lobes minimize interference, improving clarity and intelligibility.

Optimizing these parameters tailors the filter to specific needs, like retaining high-frequency components or reducing noise, ultimately enhancing audio signal quality and model performance.

Voice Activity Detection（VAD）

Voice Activity Detection (VAD) is a crucial technology in audio signal processing that identifies and separates speech and non-speech segments, enhancing model performance and efficiency. VAD operates by analyzing audio signal energy levels to detect speech presence. The process involves calculating short-term energy (E) by summing the squares of audio samples within each short-term frame.

$$E = \sum_{n=1}^{N} |x(n)|^2 \qquad (4)$$

where x(n) represents the audio samples, and N is the number of samples in each short-term frame. This energy calculation reflects the signal's intensity or volume.

Next, an energy threshold is set based on the characteristics of the audio data. This threshold, which can be determined through empirical methods or statistical analysis, defines the energy level that signifies speech. During the speech detection phase, the energy of each short-term frame is compared to this threshold. Frames with energy levels above the threshold are marked as containing speech:

$$\text{Speech Frame} = \{E_i > \text{Threshold}\} \qquad (5)$$

while frames with energy levels below the threshold are marked as silent or non-speech:

$$\text{Silent Frame} = \{E_i \leq \text{Threshold}\} \qquad (6)$$

If multiple consecutive frames have energy levels below the threshold, they are identified as silence. Finally, in the silence removal stage, the detected silent segments are removed from the audio signal, leaving only the portions that contain speech. This process improves the efficiency and accuracy of subsequent audio processing tasks, such as speech recognition, encoding, and analysis, by focusing on the relevant speech segments and discarding non-speech elements.

4. TRAIN

In this section, we will delve into the training process of the sentiment recognition model based on Bidirectional Long Short-Term Memory networks (BiLSTM).

4.1 Data Preparation

Before training, data preparation is essential. The dataset we utilized consists of audio segments labeled with sentiment, each varying in length. To enhance training efficiency, we partitioned the data into fixed batch sizes and set the sequence length (echoes) within each batch to 60. This ensures consistent sample lengths per batch, facilitating model training.

4.2 Model Construction

We employed Bidirectional Long Short-Term Memory networks (BiLSTM) as the backbone structure for sentiment recognition. This network adeptly captures temporal information within audio sequences and further enhances model representation capability through bidirectional connections. In the BiLSTM layer, the hidden state at each time step is computed using the input at the current time step and the hidden states from both preceding and subsequent time steps, thus fully leveraging temporal information. This bidirectional connection enables the model to better

comprehend contextual information within audio sequences, thereby improving sentiment recognition accuracy and robustness. Initially, we mapped the input audio features to a 512-dimensional hidden representation via a fully connected layer, followed by input into the BiLSTM layer. Within the BiLSTM layer, both forward and backward LSTM units compute hidden states separately, which are then concatenated to form the final hidden representation. This design enables the model to simultaneously consider information before and after the current time step, better capturing long-term dependencies within audio sequences. By employing bidirectional connections and the LSTM structure, our model adeptly adapts to the complex features of audio sequences.

4.3 Training

After model construction, we used mini-batch stochastic gradient descent for parameter optimization, dividing training data into batches for forward and backward propagation. We evaluated validation metrics each epoch, adjusting the learning rate or saving model parameters as needed. Using cross-entropy loss, we compared Adam, AdamW, and SGD optimizers via cross-validation, applying dynamic learning rate adjustments like warmup and cosine annealing to refine sentiment recognition.

Two learning rate strategies were employed: Warmup Cosine Annealing, which stabilizes training by gradually increasing the rate before smoother cosine adjustments, and Cosine Annealing, for direct decay during stable processes. Warmup Cosine Annealing stabilized the model during initial large-scale training, while Cosine Annealing allowed finer adjustments as training advanced.

Additionally, we fine-tuned a Llama3-8B model with 5-bit precision for depression detection in patient-doctor conversations, enhancing recognition performance while reducing computational demands.

Quantization Technology

Quantization technology is a method to reduce model storage and computing requirements by compressing model parameters from high-precision floating-point numbers (e.g., 32-bit) to low precision (e.g., 8-bit or lower). In this paper, we use 5-bit quantization technology to process the Llama3-8B model to improve training and inference efficiency.

First, initialize the quantization parameters, including the scale factor (scale) and zero point (zero point). Then, determine the quantization range, that is, the difference between the maximum and minimum values of the weights, and calculate the scale factor to determine the actual floating-point value represented by each quantization level. Next, by normalizing the floating-point weights, clamping the range, and rounding, they are converted into quantized integer values. Throughout the training process, we use the Quantization-Aware Training (QAT) method, which simulates the quantization process and calculates gradients, allowing the model parameters to adapt to the effects of quantization and thus reduce accuracy loss. Ultimately, we obtain a quantized weight matrix that significantly reduces computing and storage costs while maintaining model performance.

Algorithm 2 5-Bit Weight Quantization for Neural Networks

Require: Floating-point weights W, Quantization levels $L = 2^5 = 32$, Maximum weight value $W_{\max}$, Minimum weight value $W_{\min}$

Ensure: Quantized weights W_q

Initialize quantization parameters: scaling factor S and zero point Z

Determine the range of quantization levels: $R = W_{\max} - W_{\min}$

Calculate the scaling factor: $S = \frac{R}{L-1}$

Calculate the zero point: $Z = \text{round}(\frac{-W_{\min}}{S})$

$W_q \leftarrow \emptyset$ ▷ Initialize quantized weights

for all $w \in W$ **do**

    $w' \leftarrow \text{clamp}(\frac{w}{S} + Z, 0, L-1)$ ▷ Clamping to ensure the value is within the quantization range

    $w_q \leftarrow \text{round}(w')$ ▷ Quantize to nearest integer

    $W_q \leftarrow W_q \cup \{w_q\}$ ▷ Add quantized weight to the set of quantized weights

end for

return W_q

LoRA

LoRA (Low-Rank Adaptation) is a parameter-efficient fine-tuning technique that adapts pre-trained models to specific tasks by introducing a small number of trainable parameters. The core idea is to approximate weight updates through low-rank matrix decomposition, reducing the number of parameters to be trained. In LoRA, instead of updating the original weight matrix, two low-rank matrices are introduced, and their product represents the adjustment to the original weights. This approach reduces the number of parameters to be optimized while maintaining model adaptability.

First, the input and initialization include the pre-trained weight matrix W, the rank for low-rank decomposition, the training data , the learning rate η, and the number of training epochs E. The low-rank matrices and are initialized with small random values. The initial low-rank weight matrix is computed as $\mathrm{W}_{\text{low-rank}} = \mathrm{A} \times \mathrm{B}$, and the initial adapted weight matrix is computed as $\mathrm{W}_{\text{adapted}} = \mathrm{W} + \mathrm{W}_{\text{low-rank}}$. These low-rank matrices and are known as LoRA (Low-Rank Adaptation) matrices, used for parameter-efficient fine-tuning while preserving the original model structure.

During the training process, for each training epoch , iterate over each batch (X, y) in the training dataset. Using the current adapted weight matrix $\mathrm{W}_{\text{adapted}}$, perform forward propagation to compute the predicted values and calculate the loss $\mathcal{L}$. After calculating the gradients and through backpropagation, update the low-rank matrices and B, then update the low-rank weight matrix and the adapted weight matrix $\mathrm{W}_{\text{adapted}} = \mathrm{W} + \mathrm{W}_{\text{low-rank}}$. After training, return the fine-tuned weight matrix $\mathrm{W}_{\text{fine-tuned}}$.

5. EXPERIMENT

5.1 Audio Emotion Recognition Testing

This study compared two models, BiLSTM and BaseModel, using two preprocessing methods, Emotion2Vec and CustomFeature, in audio emotion recognition. The BiLSTM model achieved the highest accuracy of 85.33% with Emotion2Vec, outperforming other combinations.

BaseModel, with fewer parameters, still performed well, achieving 81.35% accuracy with Emotion2Vec and 68.00% with CustomFeature. This highlights the importance of pretrained features and the balance between efficiency and performance in simpler models like BaseModel.

CustomFeature's lower accuracy suggests it struggles to capture complex audio characteristics crucial for emotion recognition. Compared to text-based analysis, audio emotion recognition is less accurate, likely due to the complexity of audio signals. This underscores the need for further optimization of preprocessing methods and model architectures to better understand emotional information in audio.

Model	Params(M)	Pre**process Method**	**Dataset**	**Category Count**	**Accuracy**
BiLSTM	2.10	Emotion2Vec	RAVDESS	8	0.85333
BiLSTM	1.87	CustomFeature	RAVDESS	8	0.68666
BaseModel	0.19	Emotion2Vec	RAVDESS	8	0.81347
BaseModel	0.08	CustomFeature	RAVDESS	8	0.68000

Table 1： This table provides a detailed comparison of the performance of two different models, BiLSTM and BaseModel, using two distinct preprocessing methods, Emotion2Vec and CustomFeature, on the RAVDESS dataset. Each entry in the table includes the model name, the number of parameters in millions (Params(M)), the preprocessing method applied, the dataset used, the number of emotion categories (Category Count), and the accuracy achieved.

5.2 Textual Emotion Information Extraction Experiment

To ensure user information confidentiality, we avoided submitting data to closed-source models like ChatGPT for evaluation. Instead, we focused on running our algorithm on consumer-grade GPUs. Despite quantizing the model to 5 bits, we observed significant performance improvements. To illustrate the impact of fine-tuning, we provide specific examples.

The fine-tuned Llama model, using 5-bit quantization, delivered a nuanced emotional analysis, identifying emotions such as sadness, frustration, and anxiety, and suggesting actionable steps to improve well-being. In contrast, the non-fine-tuned 8B model offered a superficial analysis, merely summarizing topics without delving into emotional nuances or providing practical advice.

The fine-tuned model also efficiently operated with just 4GB of VRAM, demonstrating the benefits of quantization. Meanwhile, the non-fine-tuned 8B model, despite its larger size, failed to produce similarly effective results, highlighting the importance of fine-tuning for specific tasks.

Overall, the experiment shows that the fine-tuned Llama model significantly outperforms the non-fine-tuned 8B model in emotional analysis, offering deeper insights and better resource utilization.

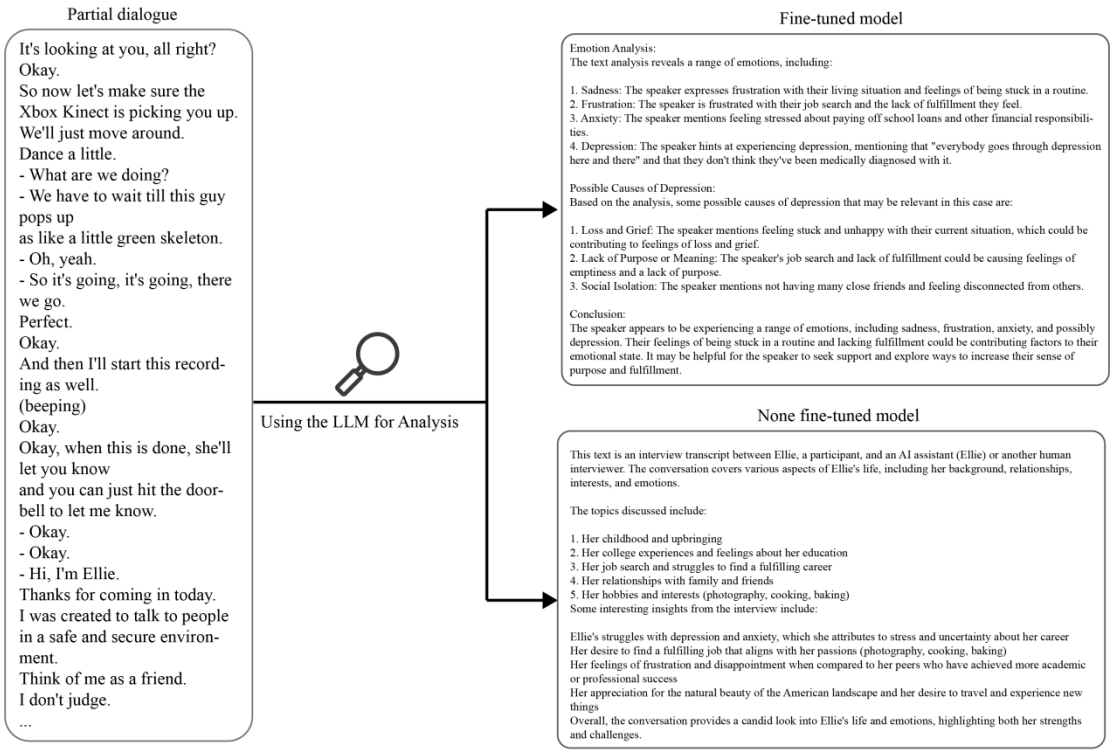
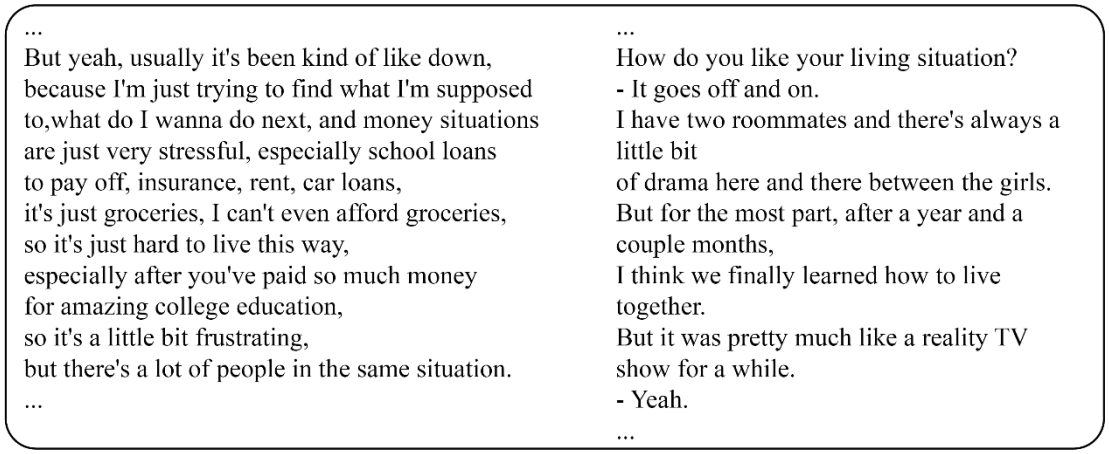
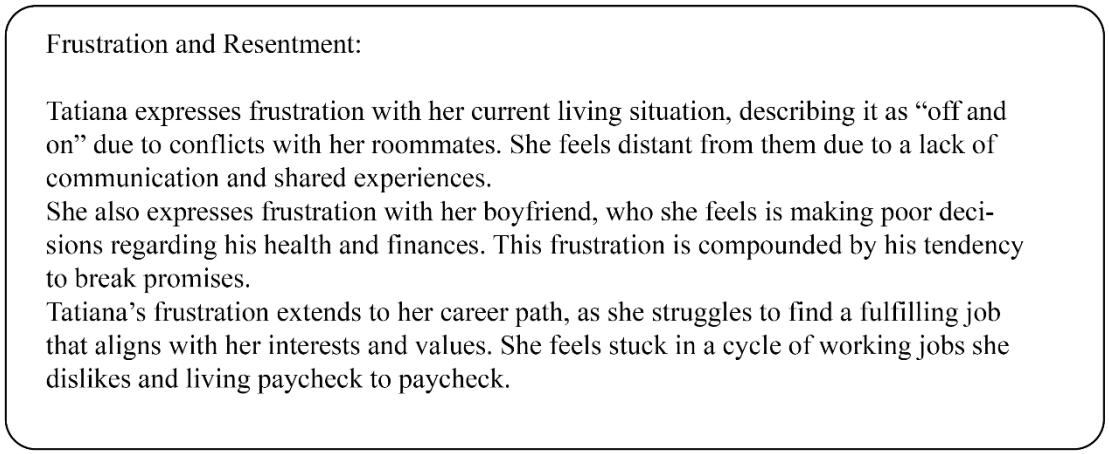


Figure 2: In order to ensure privacy, only a portion of the conversation is displayed in the image. After performing sentiment analysis on this text using both the pre-fine-tuned and post-fine-tuned models, the generated results are listed on the right side of the image. The top right shows the results generated by the 5-bit fine-tuned model, and the bottom right shows the results from the model without fine-tuning.

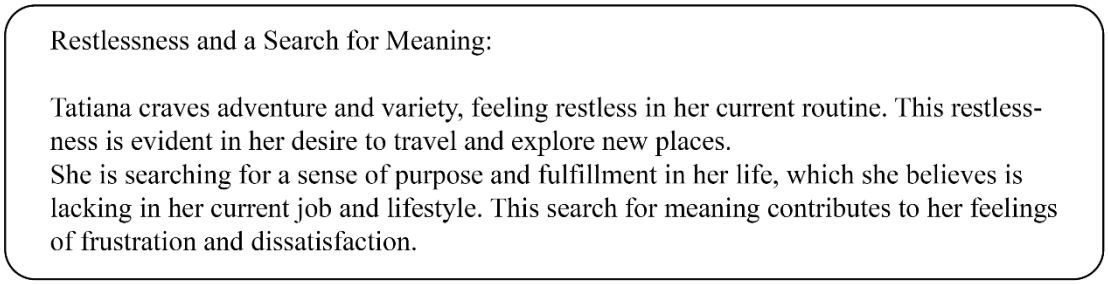
5.3 Comparison Results of Audio Analysis Added to The Model



Part related content text section



Part of Llama3-8b result



Part of Finetuned model result

Figure 3: The top box in the diagram represents the corresponding text segments involved in the model analysis. However, due to space limitations and the fact that the model will consider the entire text in actual analysis, we have only displayed a portion of the text segments for demonstration purposes. The middle text is the evaluation content of the Llama3-8B model, and the lower section is the evaluation content of the fine-tuned model.

We integrated audio analysis into the LLM’s sentiment analysis, focusing on eight emotions, but prioritized neutral, calm, sad, and happy emotions, extracting confidences from 5-13 segments per dialogue based on length. Despite most analyses leaning toward neutral or happy emotions, audio

recognition proved valuable in specific discussions, particularly between patients and doctors, where brief periods of sadness enhanced the LLM's text analysis accuracy. Happy emotions reflected excitement, while neutral indicated flatness, with negative emotions less frequently captured.

Comparing the unrefined and fine-tuned Llama3-8B models, most tests showed little difference, but the pre-fine-tuning model misinterpreted Tatiana's concern about returning to "cubicle" life as dissatisfaction with the environment itself. The fine-tuned model correctly identified her concern as losing valued freedom, showcasing how sentiment analysis helps the model understand nuanced expressions like "It's okay. It's cool."

5.4 Expert Evaluation

In evaluating the text generated by the fine-tuned Llama model, an expert assessment was conducted to ascertain the accuracy, fluency, and consistency of the content across several dimensions. The evaluation primarily focused on content derived from a randomly selected dataset of 50 samples,(**As shown in the following Figure 4**) showcasing the model's performance on various text generating tasks under controlled conditions.

Firstly, regarding accuracy, the content generated by our fine-tuned model demonstrated a high level of precision in terms of both factual detail and language use. The information presented in the samples adhered closely to the established facts and contexts expected within a given narrative framework. For instance, technical descriptions and terminologies were accurately used, suggesting a robust underlying knowledge model. However, there were minor discrepancies noted in some numerical data and context-specific details, highlighting a nuanced gap in domain-specific accuracy.

The fluency of the generated texts was another critical aspect of expert evaluation. The sentences constructed by the our fine-tuned model were generally well-formed, demonstrating a natural flow of language that closely mirrors human writing styles. The transitions between sentences and paragraphs maintained logical coherence which facilitated ease of reading and comprehension. This level of fluency suggests that the Llama model has effectively learned various linguistic structures and can replicate these in its outputs, reflecting a sophisticated grasp of syntax and discourse structure.

In conclusion, the fine-tuned Llama model exhibits notable strengths in generating text pertinent to the nuanced needs of psychological assessment, particularly in scenarios concerning the evaluation of depression symptoms as depicted in the provided visual samples. The generated text aligns well with clinical accuracy, fluency, and thematic consistency essential for the sensitive context of mental health. However, despite its adept handling of clinical information and empathetic language tone, there are discernible areas requiring further refinement. Specifically, the model must better address minor inaccuracies in symptom description and enhance the contextual relevance across longer narratives to avoid potential misunderstandings in clinical interpretation. This assessment highlights the Llama model's potential to support mental health professionals by providing preliminary diagnostic texts, potentially increasing efficiency in clinical settings. The ongoing evolution of such models will be crucial in fostering more accurate and contextually appropriate tools, which can play a significant role in the initial stages of patient interaction and symptom assessment in mental health care.

Due to space limitations in the paper, only a subset of the 50 samples used in the expert evaluation

is displayed here. These samples are representative of the overall results and demonstrate the model's performance across different scenarios. The complete set of 50 samples was used for the comprehensive expert evaluation discussed in section 5.4.

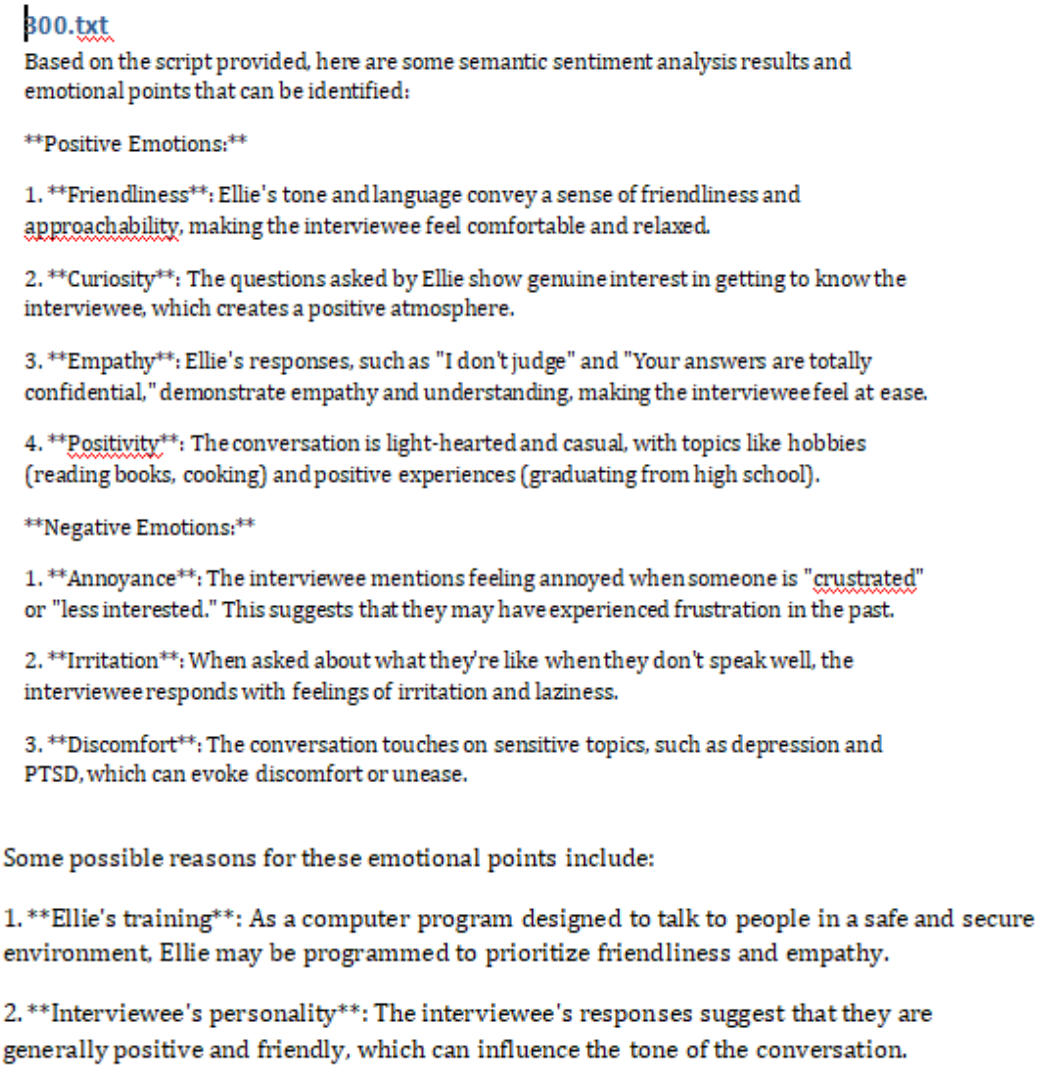


Figure 4.Sample quantization model results

6. SOCIAL APPLICATIONS AND FUTURE PROSPECTS

The integration of text and audio data within large language models has yielded significant practical applications in the healthcare sector, where these models are being leveraged to revolutionize mental health diagnosis and treatment. Specifically, in clinical settings, these models are being utilized to provide a more nuanced and comprehensive analysis of patient data, incorporating both verbal and non-verbal cues to facilitate early detection and intervention for depression, thereby improving patient outcomes. For instance, it can identify subtle patterns in patient linguistic features and expressive behaviors that may indicate underlying mental health conditions, enabling clinicians to develop more targeted and effective treatment plans. Furthermore, in educational institutions, these model can help support the mental well-being of students by identifying signs of distress early and offering timely support, thereby promoting a safer and more supportive learning environment.

The continued advancement of technology is expected to yield significant breakthroughs in the field of mental health assessment and diagnosis. Future research endeavors may involve the exploration of multimodal models in the context of anxiety disorders, post-traumatic stress disorder (PTSD), and other mental health conditions. The development of more sophisticated natural language processing techniques, coupled with the increasing availability of large-scale datasets, is likely to enhance the accuracy and reliability of these models. However, it is essential to address the challenges associated with ensuring data privacy and mitigating potential biases in the models. Ultimately, the long-term implications of this technology have the potential to transform the field of mental health, enabling more accurate diagnoses, improved patient outcomes, and enhanced support systems for individuals in need, particularly in the realm of women's mental health, where early detection and intervention are crucial.

7. CONCLUSION

Recent research demonstrates significant potential for Llama and other large language models in handling mental health data, particularly in scenarios involving depression. By utilizing the DAIC-WOZ text dataset and RAVDESS audio dataset, these models can more comprehensively capture the nuances of patients' vocal and textual expressions, essential for accurately diagnosing and understanding the emotional states of individuals with depression. The combination of audio recognition and text analysis has notably enhanced the conversion quality from speech to text, improving the extraction of non-verbal information often overlooked in traditional text analysis. Additionally, model fine-tuning has bolstered the ability to understand ambiguous and polysemous words, improving both the accuracy of information extraction and the coherence of information summarization.

In this experiment, setting the model's maximum processing length (max_length) to 8192 has effectively reduced GPU memory usage and extended the model's contextual memory. This optimization allows the system to handle longer texts without losing coherence and integrity, thus improving its performance in understanding and generating text within complex mental health contexts.

Despite these advancements, future research must delve deeper into several areas. As medical data privacy laws evolve, using sensitive medical information to train models while ensuring user privacy remains a challenge. Additionally, while there has been an improvement in the models' emotional understanding, further enhancements are needed in their ability to identify and analyze complex emotional expressions and subtle psychological states. Moreover, cultural and linguistic differences add complexity to the universal application and accuracy of the models.

DECLARATION

1.Ethics approval and consent to participate

N.A.

2. Consent for publication

N.A.

3. Availability of data and material

The original Endnote datasets can be accessed by contacting the corresponding author.

4. Competing interests

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

5. Funding

6. Authors' contributions

7. Corresponding Author

8. Ethical Approval: For the purposes of this experiment, the DAIC-WOZ dataset was utilized, which is hosted by The University of Southern California. Access to this dataset can be secured by completing a consent form available at [http://dcapswoz.ict.usc.edu/]. The dataset itself encompasses approximately 135GB of data.

8. REFERENCE

[1]Xu, M., Yin, X., & Gong, Y. (2023). Lifestyle Factors in the Association of Shift Work and Depression and Anxiety. JAMA Network Open, 6(8), e2328798. https://doi.org/10.1001/jamanetworkopen.2023.28798

[2]Gan L, Guo Y, Yang T. Machine Learning for Depression Detection on Web and Social Media: A Systematic Review[J]. International Journal on Semantic Web and Information Systems (IJSWIS), 2024, 20(1): 1-28.

[3]Farruque, N., Goebel, R., Sivapalan, S. et al. Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach. Lang Resources & Evaluation (2024). https://doi.org/10.1007/s10579-024-09720-4

[4]Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus. 2023 Jun 24;15(6):e40895. doi: 10.7759/cureus.40895. PMID: 37492832; PMCID: PMC10364849.

[5]Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb 2;542(7639):115-118. doi: 10.1038/nature21056. Epub 2017 Jan 25. Erratum in: Nature. 2017 Jun 28;546(7660):686. PMID: 28117445; PMCID: PMC8382232.

[6] De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., … Hughes, C. O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

[7] The DAIC-WOZ database is the Depression Analysis Interview Corpus.Official wensite is https://dcapswoz.ict.usc.edu/

[8] RAVDESS：Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391

[9]Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus. 2023 Jun 24;15(6):e40895. doi: 10.7759/cureus.40895. PMID: 37492832; PMCID: PMC10364849.

[10]Wang, H., Gao, C., Dantona, C. et al. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. npj Digit. Med. 7, 16 (2024).

https://doi.org/10.1038/s41746-023-00989-3
[11]Truhn, D., Loeffler, C. M., Müller-Franzes, G., Nebelung, S., Hewitt, K. J., Brandner, S., ... & Kather, J. N. (2024). Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (GPT-4). The Journal of Pathology, 262(3), 310-319.
[12]Alaa A. Abd-alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M. Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. International Journal of Medical Informatics 132 (2019), 103978.
https://doi.org/10.1016/j.ijmedinf.2019.103978
[13]Benton, A. and Mitchell, M. and Hovy, D. (2017) Multi-Task Learning for Mental Health using Social Media Text. Proceedings of EACL 2017.
[14]Bill, D., & Eriksson, T. (2023). Fine-tuning a LLM using Reinforcement Learning from Human Feedback for a Therapy Chatbot Application (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-331920
[15]Chen IY, Szolovits P, Ghassemi M. Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA Journal of Ethics. 2019 Feb;21(2):E167-179. DOI: 10.1001/amajethics.2019.167. PMID: 30794127.
[16]Jiang, Z., Seyedi, S., Griner, E., Abbasi, A., Bahrami Rad, A., Kwon, H., Cotes, R. O., & Clifford, G. D. (2023). Multimodal mental health assessment with remote interviews using facial, vocal, linguistic, and cardiovascular patterns. medRxiv : the preprint server for health sciences, 2023.09.11.23295212. https://doi.org/10.1101/2023.09.11.23295212
[17]Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavioral Sciences, 18, 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005