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], Jiaming Tan[1]

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

[2]Department of Academician Chunming Rong Team Innovation, University of Sanya, Sanya, Hainan, China

Abstract: 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. 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.

In recent years, The integration of clinical psychology with machine learning (ML) and deep learning (DL) has advanced depression diagnosis by uncovering patterns in data like EHRs, social media, and wearables. Applications include predicting depressive episodes and personalizing treatments. However, challenges like data privacy, model interpretability, and limited annotated datasets hinder broader clinical adoption [2] [4] [5].

Large Language Models (LLMs), like OpenAI's GPT series, offer transformative potential in mental health. They can analyze linguistic markers to assist in early depression screening, provide scalable and personalized therapeutic support (e.g., CBT), and analyze unstructured data to advance research on depression's etiology and treatment.

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 [1].

2. The Proposed Framework

2.1 DAIC-WOZ dataset

The DAIC-WOZ database, a subset of the Distress Analysis Interview Corpus (DAIC), contains clinical interview recordings designed to assist in diagnosing mental health conditions such as anxiety, depression, and PTSD. The dataset includes 189 samples (audio, video, and questionnaire responses) featuring interviews conducted by "Ellie," a virtual animated interviewer controlled remotely. All data have been transcribed and annotated for verbal and non-verbal features [7].

In this study, we converted audio recordings from DAIC-WOZ into text, followed by data cleaning and feature extraction. These preprocessed texts were used for tasks such as information extraction and sentiment analysis, supporting the automated assessment of mental health conditions. This approach provides deeper insights into patients' psychological states, improving diagnostic and treatment processes.

2.2 Ravdess dataset

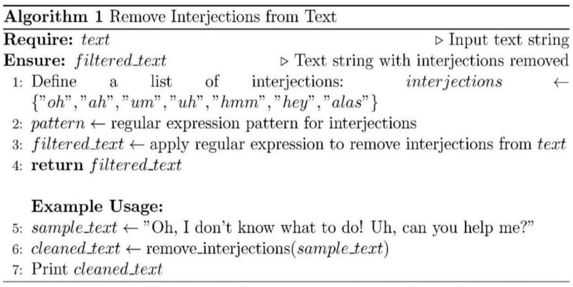
The RAVDESS dataset, widely used in emotion recognition and mental health research, comprises 7,356 recordings by 24 actors (12 male, 12 female) with neutral North American accents. It includes expressions of emotions such as calm, happy, sad, fearful, angry, surprised, and disgusted at two levels of intensity.

For this study, we selected 2,452 audio files: 1,440 from the "Audio_Speech_Actors_01-24" subset and 1,012 from the "Audio_Song_Actors_01-24" subset. These files, stored in standard audio format, were used as the training set. Testing excluded files without sound content, ensuring consistency in analysis [8].

3. Data Preprocessin

3.1 Text Emotion Analysis

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, 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
  Define a list of interjections: $interjections \leftarrow \{"oh","ah","um","uh","hmm","hey","alas"\}$
  $pattern \leftarrow$ regular expression pattern for interjections
  $filtered\_text \leftarrow$ apply regular expression to remove interjections from $text$
  return $filtered\_text$

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

Figure l. Remove Interiections from Text

3.2 Audio Emotion Analysis 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:

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

$0 \leq n < M$ otherwise

Where represents thenth filter coefficient, I_0 is the zero−order modified Bessel function, with a specific value of $\sum_{k=0}^{\infty}\frac{(x/2)^{2k}}{(k!)^2}$, β 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 key technology in audio signal processing that separates speech from non-speech, enhancing the efficiency of tasks like speech recognition and audio analysis. VAD relies on calculating the short-term energy (E) of audio signals, using the formula:

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

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

An energy threshold is set (based on empirical or statistical methods) to signify speech presence.

During detection:

Frames where energy (EiEi) exceeds the threshold are marked as speech:

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

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

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

Consecutive silent frames are identified as non-speech and removed, leaving only the speech segments. This process ensures that subsequent audio tasks focus on relevant speech data while discarding silence.

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 Model Construction

We optimized training efficiency by batching labeled audio segments into fixed sizes with a consistent sequence length of 60. Then we used a Bidirectional Long Short-Term Memory network (BiLSTM) as the core structure. BiLSTM captures temporal information in audio sequences and enhances model representation through bidirectional connections. The hidden state at each time step considers inputs from both past and future steps, improving sentiment recognition accuracy. Input audio features were first mapped to a 512-dimensional hidden representation via a fully connected layer, then fed into the BiLSTM layer. Forward and backward LSTM units computed hidden states separately, which were concatenated to form the final representation, capturing long-term dependencies in audio sequences.

4.3 Training

Table 1. The performance of two different models, BiLSTM and BaseModel

Model	Params(M)	Preprocess 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 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.

We used mini-batch stochastic gradient descent for parameter optimization, dividing data into batches for forward and backward propagation. Validation metrics were evaluated each epoch, with learning rate adjustments or model parameter saving as needed. We compared Adam, AdamW, and SGD optimizers using cross-validation and applied dynamic learning rate adjustments like warmup and cosine annealing. Two learning rate strategies were used: Warmup Cosine Annealing (gradually increasing then smoothly adjusting the rate) and Cosine Annealing (direct decay during stable training). We also fine-tuned a Llama3-8B model with 5-bit precision for depression detection in patient-doctor conversations, improving recognition while reducing computational demands.

4.3.1 LoRA (Low-Rank Adaptation)

LoRA is a parameter-efficient fine-tuning method that adapts pre-trained models to specific tasks by introducing a small number of trainable parameters. It approximates weight updates through low-rank matrix decomposition, reducing the number of parameters to be trained. Two low-rank matrices are introduced, and their product represents the adjustment to the original weights. During training, these matrices are updated to optimize the adapted weight matrix, preserving the original model structure while enabling efficient fine-tuning.

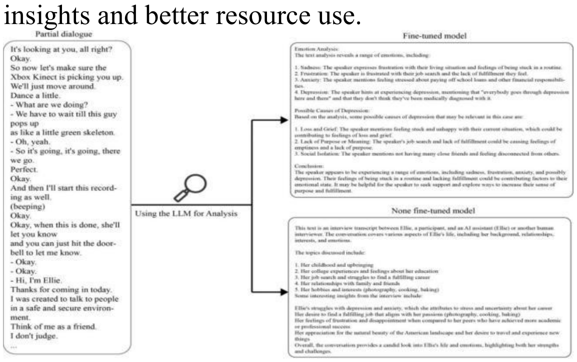
5. Experiment

5.1 Audio Emotion Recognition Testing

This study compared the BiLSTM and BaseModel in audio emotion recognition, using Emotion2Vec and CustomFeature preprocessing. The BiLSTM achieved 85.33% accuracy with Emotion2Vec, the best result. BaseModel, with fewer parameters, reached 81.35% with Emotion2Vec and 68.00% with CustomFeature. This shows the value of pretrained features and the efficiency-performance balance in simpler models. CustomFeature's lower accuracy indicates it struggles with complex audio characteristics. Audio emotion recognition is less accurate than text-based analysis due to audio signal complexity, highlighting the need for better preprocessing and model architectures.

5.2 Textual Emotion Information Extraction Experiment

To protect user data, we ran our algorithm on consumer - grade GPUs instead of using closed - source models like ChatGPT. Even after quantizing the model to 5 bits, performance improved significantly. For example, the fine - tuned Llama model with 5 - bit quantization gave detailed emotional analyses, spotting sadness, frustration, and anxiety, and offered practical advice for better well - being. In contrast, the non - fine - tuned 8B model only provided basic summaries without emotional details or useful tips. The fine - tuned model also used just 4GB of VRAM, showing the benefits of quantization. Meanwhile, the larger, non - fine - tuned 8B model didn't perform as well, proving the need for task - specific fine - tuning. Overall, the fine - tuned Llama model outperformed the non - fine - tuned 8B model in emotional analysis, offering deeper insights and better resource use.



It's looking at you, all right?
Okay.
So now let's make sure the
Xbox Kinect is picking you up.
We'll just move around.
Dance a little.
- What are we doing?
- We have to wait till this guy
pops up
as like a little green skeleton.
- Oh, yeah.
- So it's going, it's going, there
we go.
Perfect.
Okay.
And then I'll start this recording as well.
(beeping)
Okay.
Okay, when this is done, she'll
let you know
and you can just hit the doorbell to let me know.
- Okay.
- Okay.
- Hi, I'm Ellie.
Thanks for coming in today.
I was created to talk to people
in a safe and secure environment.
Think of me as a friend.
I don't judge.
...

Emotion Analysis:
The text analysis reveals a range of emotions, including:

1. Sadness: The speaker expresses frustration with their living situation and feelings of being stuck in a routine.
2. Frustration: The speaker is frustrated with their job search and the lack of fulfillment they feel.
3. Anxiety: The speaker mentions feeling stressed about paying off school loans and other financial responsibilities.
4. Depression: The speaker hints at experiencing depression, mentioning that "everybody goes through depression here and there" and that they don't think they've been medically diagnosed with it.

Possible Causes of Depression:
Based on the analysis, some possible causes of depression that may be relevant in this case are:

1. Loss and Grief: The speaker mentions feeling stuck and unhappy with their current situation, which could be contributing to feelings of loss and grief.
2. Lack of Purpose or Meaning: The speaker's job search and lack of fulfillment could be causing feelings of emptiness and a lack of purpose.
3. Social Isolation: The speaker mentions not having many close friends and feeling disconnected from others.

Conclusion:
The speaker appears to be experiencing a range of emotions, including sadness, frustration, anxiety, and possibly depression. Their feelings of being stuck in a routine and lacking fulfillment could be contributing factors to their emotional state. It may be helpful for the speaker to seek support and explore ways to increase their sense of purpose and fulfillment.

This text is an interview transcript between Ellie, a participant, and an AI assistant (Ellie) or another human interviewer. The conversation covers various aspects of Ellie's life, including her background, relationships, interests, and emotions.

The topics discussed include:

1. Her childhood and upbringing
2. Her college experiences and feelings about her education
3. Her job search and struggles to find a fulfilling career
4. Her relationships with family and friends
5. Her hobbies and interests (photography, cooking, baking)
Some interesting insights from the interview include:

Ellie's struggles with depression and anxiety, which she attributes to stress and uncertainty about her career
Her desire to find a fulfilling job that aligns with her passions (photography, cooking, baking)
Her feelings of frustration and disappointment when compared to her peers who have achieved more academic or professional success
Her appreciation for the natural beauty of the American landscape and her desire to travel and experience new things
Overall, the conversation provides a candid look into Ellie's life and emotions, highlighting both her strengths and challenges.

Figure 2. Pre-Fine-Tuned And Post-Fine-Tuned Models

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.

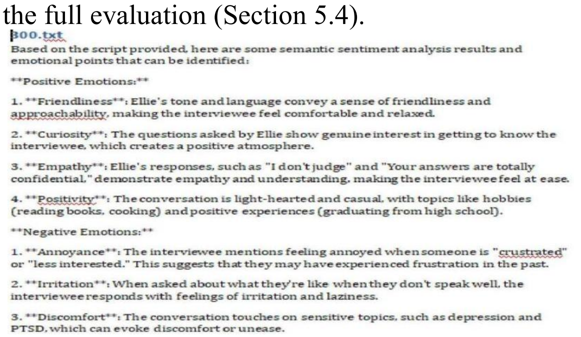
5.3 Comparison Results of Audio Analysis Added to The Mode

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

Experts evaluated the fine-tuned Llama model's text across accuracy, fluency, and consistency using 50 random samples (Figure 3). The model demonstrated high precision in facts and language, with accurate technical terms, though minor errors in numbers and context-specific details were noted. Texts were well-structured and coherent, closely resembling human writing, indicating the model's strong grasp of language structure. In psychological assessments, particularly depression evaluations, the model performed well, with text aligning with clinical standards but needing refinement in symptom details and context. In conclusion, the fine-tuned Llama model generates accurate, fluent, and consistent text for mental health assessments. Minor improvements are needed for clinical use, but it can aid professionals with preliminary diagnoses, boosting efficiency. Future work should focus on enhancing detail accuracy and context. Due to space limits, only some samples are shown, but all 50 were used in the full evaluation (Section 5.4).



300.txt

Based on the script provided, here are some semantic sentiment analysis results and emotional points that can be identified:

Positive Emotions:

1. **Friendliness**: Ellie's tone and language convey a sense of friendliness and approachability, making the interviewee feel comfortable and relaxed.

2. **Curiosity**: The questions asked by Ellie show genuine interest in getting to know the interviewee, which creates a positive atmosphere.

3. **Empathy**: Ellie's responses, such as "I don't judge" and "Your answers are totally confidential," demonstrate empathy and understanding, making the interviewee feel at ease.

4. **Positivity**: The conversation is light-hearted and casual, with topics like hobbies (reading books, cooking) and positive experiences (graduating from high school).

Negative Emotions:

1. **Annoyance**: The interviewee mentions feeling annoyed when someone is "crustrated" or "less interested." This suggests that they may have experienced frustration in the past.

2. **Irritation**: When asked about what they're like when they don't speak well, the interviewee responds with feelings of irritation and laziness.

3. **Discomfort**: The conversation touches on sensitive topics, such as depression and PTSD, which can evoke discomfort or unease.

Figure 3. Sample Quantization Model Results

6. Conclusion

Recent research shows Llama and other large language models have strong potential in handling mental health data, especially for depression. Using the DAIC-WOZ text dataset and RAVDESS audio dataset, these models can better capture the nuances in patients' speech and text, which is key to accurately diagnosing and understanding the emotional states of people with depression. Combining audio and text analysis has greatly improved speech-to-text conversion and the extraction of non - verbal information often missed in traditional text analysis. Fine - tuning the models has also enhanced their ability to understand ambiguous and multi - meaning words, improving information extraction accuracy and summary coherence.

In our experiment, setting the model's max_length to 8192 reduced GPU memory usage and extended contextual memory, allowing the system to handle longer texts without losing coherence, thus improving performance in complex mental health contexts.

However, future research needs to focus on several areas. With evolving medical data privacy laws, training models on sensitive medical data while ensuring user privacy remains a challenge. Also, though models' emotional understanding has improved, their ability to identify and analyze complex emotions and subtle psychological states needs further enhancement. Cultural and linguistic differences also complicate the universal application and accuracy of these models.

Declaration 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.

Acknowledgements

Special Training Program for Young Teachers (Teaching Category): 'Construction and Exploration of Big Data Technology and Application Courses for Non-Computer Science Majors' (SYJPZQ00)

References

[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.

[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).

[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.

[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.

[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.

[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.

[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.

[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).

[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.

[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.

[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.

[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.