International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n1.32
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Depression Detection with Novel Large Language Model Methods

Alan Lu [1], Phillip Koehn [2]

[1] Saratoga High School, 95070, Saratoga, United States
[2] Johns Hopkins University, Baltimore, MD 21218, United States

ABSTRACT

Recent advancements in machine learning and natural language processing have shown significant potential in aiding mental health diagnostics. In this paper, we specifically provide novel approaches to depression identification using large language models (LLMs). First, we employ the LLM2Vec architecture for text-only classification. Additionally, we explore the potential of multimodal LLMs by incorporating text, audio, and visual inputs, aiming to enhance diagnostic accuracy through a more comprehensive analysis of multimodal data. Our results demonstrate that text-only models perform well above baseline performances, while multimodal data shows promise for achieving a more nuanced understanding of a patient's mental state. This work lays the groundwork for future research in developing more robust and effective tools for mental health assessment using LLMs.

KEYWORDS

Machine Learning; Multimodal analysis; Large Language Models; Depression Detection

1. INTRODUCTION

Mental illness affects over 20% of U.S. adults, with approximately 7.1% suffering from major depressive disorder (MDD) [1]. Depression is closely linked to serious health conditions like heart disease, cancer, and diabetes [2]. The suicide rate in the U.S. has increased by 35.2% over the past two decades, with nearly half of these cases associated with MDD [3,4]. Despite the prevalence and severity of MDD, only 16.5% of affected individuals receive minimally adequate treatment [5].

To bridge the treatment gap and enhance outcomes for individuals with major depressive disorder (MDD), pioneering diagnostic techniques are critical. Traditional diagnostic practices for depression typically depend on self-reported symptoms and clinical evaluations, which may be skewed by personal biases and underreporting. However, advancements in machine learning (ML) and natural language processing (NLP) present promising alternatives that facilitate more objective and earlier detection of depression. For instance, the established field of sentiment analysis in NLP research aligns closely with identifying mental health conditions. By examining textual data from various sources such as speech transcripts, social media content, or written responses, ML models can discern patterns and indicators of depression or other mental health issues. Furthermore, the incorporation of multimodal data combining text, audio, and visual inputs can offer a more comprehensive understanding of a patient's condition.

2. RELATED WORK

Research in text-based analysis for mental health aims to identify language and speech patterns that are correlated with mental health disorders. For example, specific words or phrases may be linked to emotions or sentiments, offering valuable insights into an individual's mental state [6]. Additionally, word vectors—numerical representations of words—can be extracted and analyzed using various machine learning classifiers, such as Support Vector Machines (SVMs) and Naive Bayes [8]. Recent advancements in more sophisticated algorithms, including convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, have further enhanced the accuracy and effectiveness of these analyses [7]. These methods improve the detection and classification of mental health conditions by capturing complex patterns and temporal dependencies in language data. Recent work by Hadzic et al. [11] compared the performance of BERT, GPT-3.5, and GPT-4 on depression detection tasks, demonstrating that GPT-4 outperforms earlier models even without fine-tuning, thus highlighting its potential in automated mental health diagnostics. Lan et al. [10] similarly employed large language models (LLMs) to analyze users' post histories, significantly improving upon previous results.

Other data modalities have also been explored. For instance, Reece and Danforth [9] used vector embedding techniques, along with face detection and colorimetric analysis, to show that Instagram photos posted by individuals with depression tend to be darker and greyer. Similarly, certain vocal features used to identify individuals are also markers of depression, suggesting an overlap between speaker identification and mental health diagnosis through voice analysis. This finding highlights the potential of using speech analysis for early depression detection while considering its implications for personal identity recognition.

3. METHODOLOGY

3.1. Dataset

We use the Distress Analysis Interview Corpus (DAIC) from the USC Institute for Creative Technologies. This dataset is typically regarded as a standard for depression detection research, consisting of text, audio, and video data from interviews between 196 patients and a virtual interviewer. Furthermore, each patient's Patient Health Questionnaire (PHQ 8) score was recorded. This standard medical form assesses mental health on a 24 point scale; a score of 10 or above indicates depressive symptoms.

Table 1. Examples of patient diagnosis based on the self-reported PHQ 8 score. The question-wise score distribution is also provided

ID	Diagnosis	PHQ8 Score	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
304	0	6	0	1	1	2	2	0	0	0
319	1	13	2	1	2	1	1	2	3	1

3.2. Text Based Analysis

The text transcripts for the 196 volunteers each consist of several predetermined questions and the volunteer's responses, including transcribed interruptions such as coughs and filler words.

Table 2. Example patient transcript of interview

Speaker	Text
Ellie	hi i'm ellie thanks for coming in today…
Ellie	what are some things you really like about la
Participant	i like the weather i like the opportunities um yes
Ellie	how easy was it for you to get used to living in la
Participant	um it took a minute somewhat easy…

We preprocess the text transcripts of each conversation, removing the virtual interviewer's speech and leaving the patient's responses. Special characters and other irrelevant phrases were then removed.

We first use the newly released open-source Llama 3 8B model by Meta. Specifically, we explore the capabilities of two models: the specialized model for sequence classification and Llama 3 8B Instruct, which is fine tuned for user prompting and textual generation. To ensure a higher degree of accuracy, we prompt the model to give results for each of the 8 PHQ categories, and we give the overall diagnosis based on the sum of these scores.

Based on the relatively small size of the dataset, we utilize several methods to preprocess the data for fine tuning. These include splitting each text transcript into multiple pieces, then upsampling the minority positive test cases to create a more balanced dataset.

Alternatively, we also use the generative version of Llama with few shot prompting instead of fine tuning the model. The model is shown a few examples with a brief explanation of the diagnoses and is then asked to similarly evaluate the results of each test case.

Given the smaller size of our dataset, we also attempt to use state of the art generative LLMs such as ChatGPT 4o to generate more test cases that are like those used in the study. We prompt the model to paraphrase each test case multiple times, giving us artificial data to work with.

As a further improvement upon the Llama 3 model, we use LLM2Vec [13], a technique that transforms decoder-only models like Llama into effective text encoders. This approach allows models traditionally designed for text generation to improve in classification tasks as well. Consequently, applying this encoder-decoder architecture theoretically will improve the performance of the decoder-only Llama 3 model when analyzing the patient transcripts.

The LLM2Vec process consists of three distinct steps. Firstly, the attention mechanism of the LLM is modified. Traditionally, these models use causal attention, which restricts tokens from attending to future tokens, aligning with the left-to-right nature of text generation tasks. However, to repurpose these models for tasks requiring a comprehensive understanding of the entire input sequence, such as text encoding, LLM2Vec replaces this causal attention with a fully bidirectional attention mechanism. This change allows every token to consider the entire sequence when forming its representation. While this alteration is essential for converting the model into a text encoder, it initially disrupts the model's performance because it contradicts the unidirectional training it previously underwent.

To rectify the performance drop caused by enabling bidirectional attention, the LLM2Vec approach introduces the Masked Next Token Prediction (MNTP) task. MNTP is inspired by the training objectives of bidirectional models like BERT, where some tokens in the input are masked, and the model is trained to predict these masked tokens. This task encourages the model to utilize both past and future contexts, aligning its training with the newly enabled bidirectional attention. By doing so, MNTP reorients the model to perform effectively as a text encoder, allowing it to harness the full sequence for predictions and thus improving its understanding and representation of the input text.

Lastly, unsupervised contrastive learning is applied, specifically using SimCSE (Simple Contrastive Learning of Sentence Embeddings). In this phase, the model is trained to maximize the similarity between different representations of the same input sequence while minimizing the similarity with

representations of different sequences. This contrastive learning task forces the model to focus on the most salient features of the input that are consistent across variations, thereby refining its ability to capture the underlying meaning and context of the text. This step significantly boosts the quality of the text embeddings, making the decoder-only LLM competitive with traditional encoder-based models in various text encoding tasks.

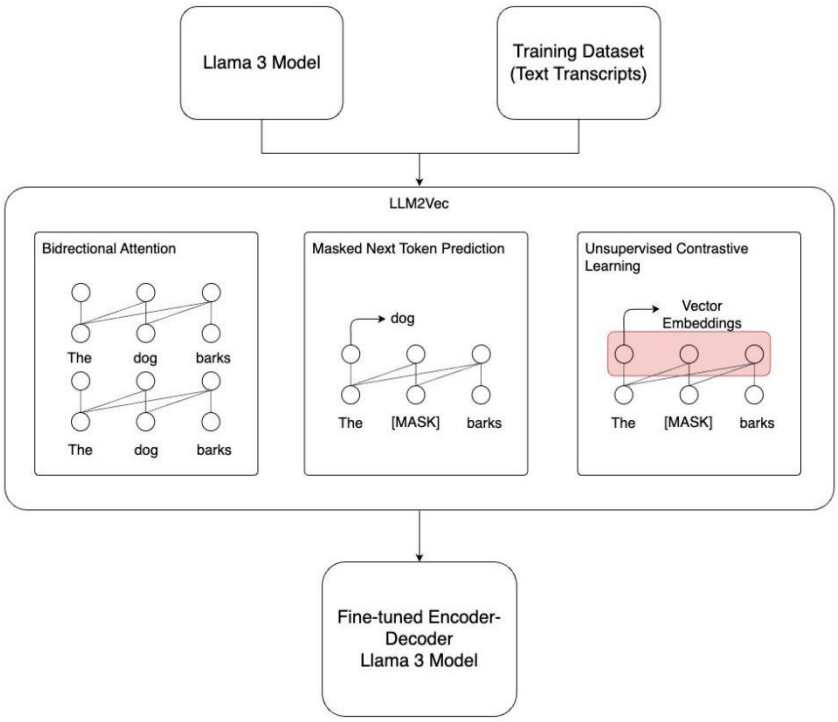


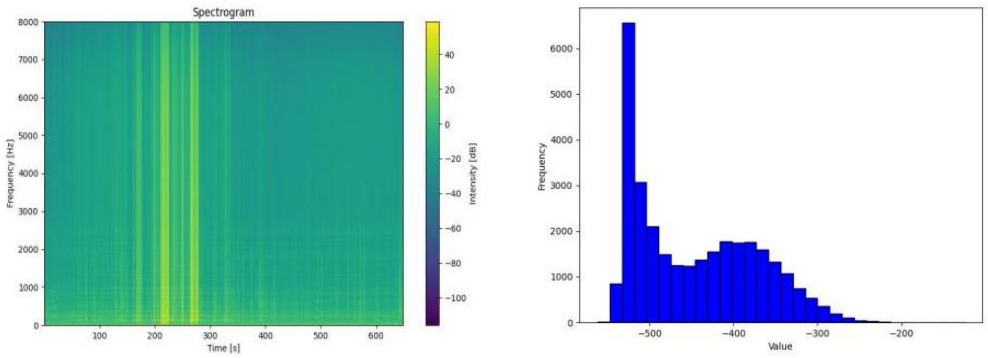
Figure 1. LLM2Vec fine-tuning process

3.3. Multimodal Analysis

Beyond text data, audio and visual modalities may provide crucial non-verbal cues that enhance the understanding of a person's emotional state. Variations in speech patterns such as tone and pitch often reflect underlying mood changes associated with depression. Similarly, visual data such as facial expressions and body language provide characterizations that may not be evident in text alone.

Thus, we consider incorporating multimodal data into our classification process. Besides text data, the DAIC-WOZ dataset also includes audio recordings of each interview. However, although raw video data is not provided due to privacy concerns, the text files for several extracted features from the videos are provided, such as pose and gaze. These features were extracted using OpenFace technology, an advanced open-source toolkit designed for facial behavior analysis [12]. OpenFace specializes in extracting facial landmarks, estimating head poses, recognizing facial action units, and other features that are crucial for detailed facial analysis.

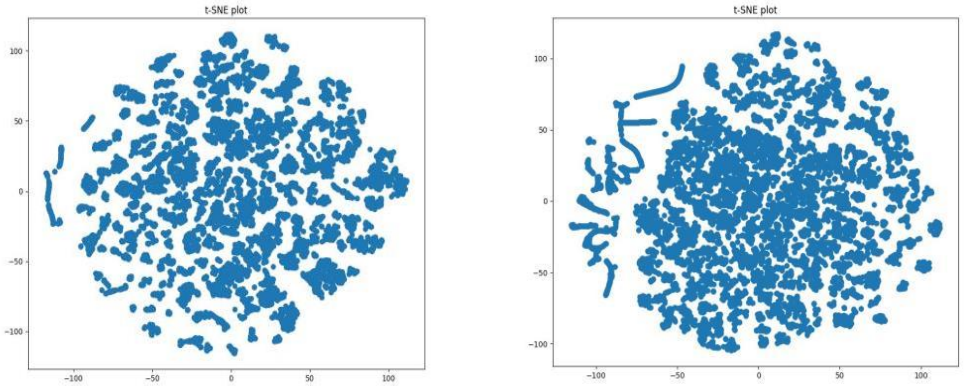
For the multimodal portion of our experimentation, we utilize the open source Idefics2 by Huggingface. This model takes a combination of images and text queries as input. This model has been trained on image and textual input, so we input the audio features in a visual manner. To do so, we experiment with several different methods. We plot the spectrograms for each audio file and try splitting the audio into multiple parts to balance the data, similarly to the text only analysis. We also attempt to plot specific features, including those such as shimmer that have been shown previously to correlate with sentiment and potentially depressive symptoms.



Generated audio spectrogram First MFCC feature distribution plot

Figure 2. Examples of experimented conversions from audio to visual data

Similarly, we decide to convert the extracted video features into image data; inputting these features simply as appended text would not work well due to the Idefics2 model having little prior training to these features. We consider the extracted Facial Action Units (AUs), which represent components of facial expressions such as the Inner Brow Raiser and Chin Raiser Action Units. We utilize t-SNE dimensional reduction to turn the many dimensional data for each frame into a plotted 2D point. We then combine this image with the audio image to feed into the multimodal model.



Non-depressed patient Depressed patient

Figure 3. Examples of generated t-SNE plots of each frame for two patients

Along with these generated images, the text portion of the data is appended to each query, effectively fusing the three modalities together.

4. EXPERIMENTS AND RESULTS

4.1. Metrics

For our experiments, we evaluate the performance of various approaches using standard classification metrics of accuracy, precision, recall, and F1 score. These metrics provide a comprehensive view of model performance, enabling us to assess not only the correctness of predictions but also the balance between precision and recall, which is particularly important in the context of depression detection where false positives and false negatives have significant implications.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+TN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$$

Figure 4: Calculations of metrics. TP, TN, FP, and FN are the total number of true positives, true negatives, false positives, and false negatives, respectively.

4.2. Computation Details

All experiments were conducted on a virtual environment equipped with an NVIDIA A100 GPU, known for its high computational power but also limited by memory constraints in handling large models. To efficiently utilize this hardware, we implemented Quantized Low Rank Adaptation (QLoRA), a technique that reduces the memory footprint of the model during both loading and training phases. We employ the bitsandbytes package to facilitate 8-bit model loading, significantly optimizing memory usage without compromising performance. This approach allowed us to work with larger models, such as the 8B parameter version of the Llama model, specifically chosen for its balance between model complexity and memory efficiency.

We conduct model training and fine-tuning using the Transformers library by Huggingface, which provides robust tools and pre-trained models for various NLP tasks. We choose the 8B version of Llama to manage memory usage effectively while still leveraging a model with substantial parameters to capture complex patterns in the data.

We use the DAIC-WOZ dataset for our experiments, which is a well-known benchmark in depression detection tasks. We divide the dataset into training and test splits, consisting of 107 and 47 datapoints, respectively. By maintaining these splits throughout all experiments, we ensure consistency and comparability of results across different models and approaches.

As a baseline, we evaluate the performance of the 110M parameter version of BERT, a widely recognized model in NLP, particularly for classification tasks. We fine tune BERT on the DAIC-WOZ training set and assess its performance on the test set using the same metrics mentioned earlier. Despite limitations on resources, we attempt to follow the approach of Hadzic et al. [11] to the best of our ability.

Below, we record the best results for each model out of the several discussed methods for processing the provided text, audio, and video data below. Specifically, the performance of the text only models is best when using a balanced dataset with the inclusion of the artificially generated data by ChatGPT 4o. The performance of the multimodal model is best when all three modalities were inputted together, with audio data being processed as individual features and video data being graphed using t-SNE plots.

Table 3. Experimental results of tested models and comparison with baseline

Model	Accuracy	Precision	Recall	F1
BERT (Baseline)	59.6%	38.1%	57.1%	45.7%
Llama-3-8B (Few shot prompting)	68.1%	47.1%	57.1%	51.6%
Llama-3-8B with LLM2Vec	74.5%	55.6%	71.4%	62.5%
Idefics2 (Images + text)	72.3%	52.9%	64.3%	58.1%

As shown, the experimental results indicate that the Llama-3-8B model, when enhanced with the LLM2Vec architecture, achieved a significant improvement in performance across all metrics compared to the baseline BERT model. The Idefics2 model, integrating images and text data, saw slight improvements in metrics over the raw Llama-3-8B model.

5. DISCUSSION

The enhancement of the LLM2Vec architecture by transforming Llama 3 into an encoder-decoder model demonstrates a significant boost in performance for text-only classification tasks, particularly in the early detection of depression. This architectural modification enables the model to capture and utilize contextual information more effectively, which is crucial for identifying subtle linguistic patterns associated with mental health conditions. Future research could build upon this advancement by integrating multimodal data, such as audio and visual cues, into the LLM2Vec framework to further enhance diagnostic accuracy. Investigating these areas may potentially broaden the accuracy and applicability of the LLM2Vec architecture, allowing it to be used as a more robust tool in mental health detection.

Although the performance of the Idefics2 multimodal model only marginally improved upon the Llama model's, our work shows promise in capturing a more holistic view of a patient's mental state by incorporating not only text but also audio and visual data. We attribute this less significant result to that the model is not pre-trained on specific forms of data like spectrograms or specialized graphs; nevertheless, our performance highlights the inherent versatility and potential of these models. Our findings indicate that the capabilities of LLMs to integrate and process multifaceted data show significant potential for enhancing mental health diagnostics.

Due to limited computational resources, we implemented several optimizations, such as Quantized Low Rank Adaptation (QLoRA) and 8-bit model loading using the bitsandbytes package, to reduce memory usage and ensure our models could be trained effectively. While these optimizations allowed us to work with large models like Llama-3-8B, they may have constrained the models' performance. As a result, our findings, though significant, may not match the results of other state-of-the-art studies that benefit from more extensive resources and fine-tuning capabilities. For instance, Hadzic et al. [11] applied more in-depth fine-tuning to their BERT model, which led to it outperforming our BERT baseline. Despite this, we chose to present our BERT results to ensure consistency in our training process, thereby making the comparison between different models in our study more valid and reliable. This approach allows us to maintain a controlled experimental environment, providing clearer insights into the relative performance of the models we evaluated.

Furthermore, we acknowledge that the overall performance and improvements demonstrated by these models is limited the size of the DAIC-WOZ dataset. The small dataset size restricts the models' ability to generalize and fully leverage their capabilities, particularly in the context of fine-tuning and multimodal integration. Collecting comprehensive data that encompasses all three modalities is particularly challenging in regulated environments. Thus, future research could potentially explore synthetic data augmentation techniques or partnerships with larger clinical studies to expand the dataset size. Advancing data privacy and sharing protocols could additionally allow more extensive data collection, enabling more conclusive studies and more robust models.

6. CONCLUSION

Our study demonstrates the significant potential of LLMs in early depression detection. The LLM2Vec architecture, which transforms a decoder-only model into an encoder-decoder model, demonstrated substantial performance gains, underscoring the value of architectural modifications in text-based classification tasks. Furthermore, we integrate multimodal LLMs by combining text, audio, and visual data, achieving notable improvements in diagnostic accuracy. Future work should focus on expanding datasets, enhancing multimodal integration, and exploring synthetic data augmentation to further advance the capabilities of LLMs in mental health diagnostics.

REFERENCES

[1] National Institute of Mental Health. “Mental Illness.” National Institute of Mental Health, U.S. Department of Health and Human Services, 2023, https://www.nimh.nih.gov/health/statistics/ mental-illness. Accessed 4 Aug. 2024.

[2] Depression and Bipolar Support Alliance. “Depression Statistics.” Depression and Bipolar Support Alliance, 2024, www.dbsalliance.org/education/depression/statistics. Accessed 4 Aug. 2024.

[3] National Institute of Mental Health. “Suicide.” National Institute of Mental Health, U.S. Department of Health and Human Services, 2024, https://www.nimh.nih.gov/health/statistics/suicide. Accessed 4 Aug. 2024.

[4] Depression and Bipolar Support Alliance. “Suicide Statistics.” Depression and Bipolar Support Alliance, https://www.dbsalliance.org/crisis/suicide-prevention-information/suicide-statistics. Accessed 4 Aug. 2024.

[5] Thornicroft, Graham, et al. “Undertreatment of People with Major Depressive Disorder in 21 Countries.” The British Journal of Psychiatry, vol. 210, no. 2, 2017, pp. 119-124. PubMed, https: //pubmed.ncbi.nlm.nih.gov/27908899. Accessed 6 Aug. 2024.

[6] Islam, Md. Rafiqul, et al. “Depression Detection from Social Network Data Using Machine Learning Techniques.” 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6111060/. Accessed 7 Aug. 2024.

[7] Kim, Jina, et al. “A Deep Learning Model for Detecting Mental Illness from User Content on Social

[8] Media.” Scientific Reports, vol. 10, no. 11846, 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC7367301/. Accessed 7 Aug. 2024.

[9] AlSagri, Hatoon, and Mourad Ykhlef. “Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features.” arXiv, 2020, arxiv.org/pdf/2003.04763. Accessed 12 Aug. 2024.

[10] Reece, Andrew G., and Christopher M. Danforth. “Instagram Photos Reveal Predictive Markers of Depression.” EPJ Data Science, vol. 6, no. 1, 2017, article 15, https://epjdatascience.springeropen. com/articles/10.1140/epjds/s13688-017-0110-z. Accessed 12 Aug. 2024.

[11] Lan, Xiaochong, et al. “Depression Detection on Social Media with Large Language Models.” arXiv, 2024, https://arxiv.org/html/2403.10750v1. Accessed 12 Aug. 2024.

[12] Hadzic, Besim, et al. “Enhancing Early Depression Detection with AI: A Comparative Use of NLP Models.” SICE Journal of Control, Measurement, and System Integration, vol. 17, no. 1, 2024, pp. 135–143, https://www.tandfonline.com/doi/full/10.1080/18824889.2024.2342624. Accessed 14 Aug. 2024.

[13] Baltruˇsaitis, Tadas, et al. “OpenFace: an open source facial behavior analysis toolkit.” Proceedings of the Winter Conference on Applications of Computer Vision. 2016, https://www.cl.cam.ac.uk/research/rainbow/projects/openface/wacv2016.pdf. Accessed 14 Aug. 2024.

[14] BehnamGhader, Parishad, et al. “LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders.” arXiv, 2024, https://arxiv.org/html/2404.05961v1. Accessed 15 Aug. 2024.