ISSN 2616-5775 Vol. 6, Issue 5: 85-90, DOI: 10.25236/AJCIS.2023.060511

Research of Using Deep Learning Language Model to Classify Depression by Level

Ziyang Liu*

Cate School, Carpinteria, California, America
**Corresponding author*

Abstract: *This article presents a multimodal neural network method that can process audio and text data simultaneously. The method uses BiLSTM and BiGRU network structures and has broad clinical and public application prospects. It has significant advantages in depression screening with high accuracy, low cost, and fast speed. The method can be applied to the whole population, especially those who are not easily accessible to healthcare. Additionally, it can be used as a fast and effective monitoring tool for continuous monitoring of the deterioration and improvement of depression.*

Keywords: *Multimodal neural network, Depression screening, BiLSTM, BiGRU, Continuous monitoring*

1. Introduction

Following the onset of the pandemic, people's lives have shifted dramatically in all aspects, bringing new world-wide challenges to be solved. One of them being the increase in major depressive disorder (addressed in the following essay as 'depression') rates, especially among those with the least access to any kinds of health care (Including the youngest age groups, oldest age groups, lowest income groups, etc.) [1]. Moreover, despite this trend, and despite depression being one of the most treatable mental illnesses, treatment for depression has not increased by proportion [2]. This suggests that a growing number of the population are suffering from untreated depression, possibly due to the complicated and potentially expensive or time-consuming diagnosing process, which leads to great risk of suicide and substance abuse otherwise avoidable, especially among those more socioeconomically vulnerable. This paper recognizes the potential dangers of the current situation and aims to address this challenge from the approach of identifying depression with machine or preprogrammed networks. This method could standardize and simplify the process of diagnosis and supervision of depression, allowing for greater availability, accessibility, and uniformity in assessment by reducing multiple in-office clinical visits, facilitating accurate measurement and identification, and quickening the evaluation of treatment.

2. Despression

Depression is a common psychological disorder, affecting about 5% of adults worldwide. Simultaneously, the harm of depression is enormous. Depression will cause huge damage to patients' physical and mental health. According to relevant research, one out of four depression patients will commit suicide, and more than half of depression patients will have suicidal tendencies. At the same time, depression will also make patients' emotions worse, cause serious psychological barriers, long-term negative emotions, destroy people's spirit, suffer from psychological suffering, interfere with social functions, and bring death shadows. Symptoms of depression include long-term feelings of sadness, loss of interest or pleasure, and changes in sleep and appetite. According to DSM-5, depression can be classified as mild, moderate, and severe. The symptoms of mild depression are mild and do not affect daily life; the symptoms of moderate depression are more severe and can affect daily life; the symptoms of severe depression are the most severe and can affect daily life and may lead to suicide. In addition, some papers explore the classification and evaluation methods of depression. For example, a paper proposes using Hamilton Depression Rating Scale (HAMD) to define the severity of depression and suggests dividing patients into different severity groups.

3. Related work

Previous research done in this field includes various methods of detection from different aspects.

ISSN 2616-5775 Vol. 6, Issue 5: 85-90, DOI: 10.25236/AJCIS.2023.060511

Yang et al.'s multimodal regression model for predicting PHQ-8 socres used the combination of a predicted PHQ-8 score generated by audio and video features processed by a 2-layer CNN network and a 2-layer FC network, and a binary classification of depressed or not depressed from textual data through Random Forest classification [3].

Haque et al. measured depression symptom severity (from predicted PHQ scores and binary classifications) from audios, linguistic signals, and 3D facial expressions processed into a multi-modal sentence embedding and through a causal CNN network [4].

Alhanai et al investigated three approaches, respectively context-free modeling, weighted modeling, and sequence modeling with a multi-modal LSTM model based on textual and audio features, for binary depression detection and concluded that the weighted approach has the best precision while the multi-modal one has the best f1, recall, MAE, and RMSE [5].

Burdisso et al. focused on using textual evidence with a newly constructed SS3 model for early risk depression detection and classification meant for use on social media [6]. Finally, Shen et al. used the same dataset as the experiment in this paper and proposed a binary depression detection model with a bidirectional LSTM network processing textual input and a GRU network for audio input [7].

The experiment in this paper specifically focuses on detection as well as classification of the severeness of depression using text and audio data using neural networks and machine learning. While such automatic detection devices cannot yet be used alone to diagnose depression, it may be used for screening and classification, potentially reducing the cost and time consumption of the diagnosing process. The remaining sections are presented in the following order. This section discusses related literature, section 2 introduces the methods of the experiment, section 3 analyzes the result and final product of the experiment as well as proposes its potential usages, and section 4 summarizes the paper with conclusions of this study.

4. Methods

4.1. Dataset

This experiment utilized the EATD-Corpus dataset [8], which contains the audio and text transcripts of 30 individuals with depression and 132 individuals without, each also labeled with a SDS score, which is a 20-items questionnaire for depression screening. Each of the 162 volunteers answered the same 3 questions asked by a virtual interviewer and both the questions and the answers were in Chinese. A total of 2.26 hours of audio data was collected with transcripts. The audio data were preprocessed to exclude mute audios, audios less than 1 second, background noise, and silent segments at the start and end of recordings. The transcripts were extracted using Kaldi and manually proofread.

4.2. Data Processing

Before being inputted into the model, both audio and textual data were processed into features and labels. Textual features are high-dimensional sentence embeddings extracted from the transcripts through ELMo.

Audio features are resampled to 16 kHz mono then computed into a sectrogram using magnitudes of the Short-Time Fourier Transform with a window size of 25 *ms*, a window hop of 10 ms, and a periodic Hann window. A mel spectrogram is computed by mapping the spectrogram to 64 mel bins covering the range 125-7500 Hz. A stabilized log mel spectrogram is computed by applying log (mel-spectrum + 0.01) where the offset is used to avoid taking a logarithm of zero. These features are then framed into non-overlapping examples of 0.96 seconds, where each example covers 64 mel bands and 96 frames of 10 ms each.

The labels of the data, originally an SDS score, was processed into 0 for a score less than 53 or without depression, and 1 for a score equal to or greater than 53, indicating depression, for binary classifications. And the labels are accordingly adjusted into 0, 1, 2, 3 for the 4-classes classification. 0 meaning no depression (<53), 1 mild depression (53-62), 2 moderate (63-72), and 3 severe depression (>72), according to the classification suggested by Yuan et al. [9].

ISSN 2616-5775 Vol. 6, Issue 5: 85-90, DOI: 10.25236/AJCIS.2023.060511

4.3. Model Text

The BiLSTM (Bidirectional Long Short-Term Memory) model [10] is adopted for text features processing.

BiLSTM is composed of bidirectional long short-term memory networks, used in the field of Deep Learning. It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, therefore often performs well with sequential data such as human languages.

BiLSTM is a type of recurrent neural network (RNN) that is widely used in natural language processing (NLP) tasks.

The main difference between a regular LSTM and a BiLSTM is that a BiLSTM processes the input sequence in both forward and backward directions. This means that the model has access to information from both past and future inputs at any given time step, which can be beneficial in tasks such as named entity recognition and sentiment analysis.

The architecture of a BiLSTM consists of two separate LSTM layers, one processing the input sequence in a forward direction, and the other processing it in a backward direction. The output of the two layers is concatenated at each time step, resulting in a final output that incorporates both past and future information.

Table 1: Parameter Settings of BiLSTM Model

Layer Name	Parameter settings
BiLSTM	Input: 1024 Hidden: 128 Layers: 2 Dropout: 0 Output: 256
FC	Out features: 128 Activation: ReLU
FC	Out features: 2 Activation: Softmax

The detailed configuration of the proposed BiLSTM has been listed in Table 1. The model consists of two BiLSTM layers, the output of which is fed into a two-layer FC network, outputting a binary prediction of whether the participant is in depression.

4.4. Model Audio

Table 2: Parameter Settings of BiGRU Model

Layer Name	Parameter settings
BiGRU	Input: 128 Hidden: 256 Layers: 1 Dropout: 0 Output: 256
FC	Out features: 256 Activation: ReLU
FC	Out features: 2 Activation: Softmax

A bi-directional GRU (Gated Recurrent Unit) model [11] is a type of neural network architecture commonly used for sequential data processing tasks, such as natural language processing and speech recognition.In a traditional GRU model, the input sequence is processed from left to right by a series of recurrent neural network layers. Each layer processes one element of the sequence at a time, and uses the output from the previous layer as input. The final output is typically taken from the last layer.

On the other hand, processes the input sequence in both directions simultaneously. It consists of two separate GRU layers, one processing the input sequence from left to right and the other from right to left. The output from each layer is concatenated at each time step, and the final output is obtained by combining the outputs from both layers.This architecture allows the model to capture both past and future context when making predictions, making it particularly effective for tasks that require understanding of context, such as machine translation and sentiment analysis.

Bi-directional GRU model is used for audio features extraction. The GRU model summarizes the audio embeddings to audio representations. Detailed configuration of the GRU model is shown in Table 2. The proposed GRU model consists of one GRU layer, followed by a two-layer FC network that outputs

ISSN 2616-5775 Vol. 6, Issue 5: 85-90, DOI: 10.25236/AJCIS.2023.060511

binary labels predicting the presence of depression.

4.5. Model Multimodal Fusion

Audio and textual data is integrated by concatenating the outputs of the audio BiGRU model and textual BiLSTM model and processing it through a final layer of the FC network to output a binary classification of depressed or not.

The above mentioned audio and text fusion model is further developed into a 4-class classification model. It is achieved by dividing the SDS score in the original label into 4 categories as described in the data processing section and adjusting the parameters of the model accordingly, as shown in Figure 1.

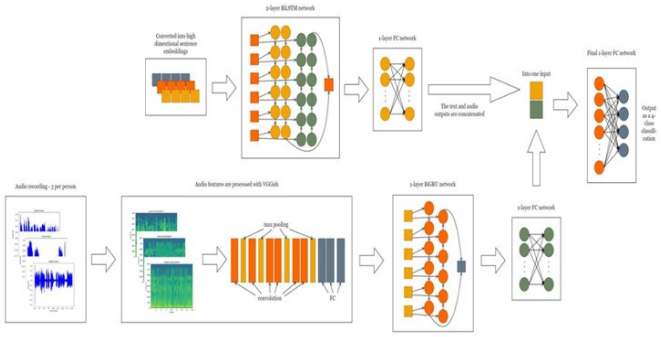

Figure 1: Flow diagram.

5. Experiment & Result

5.1. Text Only

The model for analyzing text was created from comparisons between the performances of various machine learning methods, the LSTM network, and the BiLSTM network.

5.2. Audio Only

The model for analyzing audio was created from comparisons between the performances of various machine learning methods, the GRU network, and the BiGRU network.

5.3. Text & Audio Fusion

The model for analyzing both text and audio was created with comparisons between different parameters and categories of classification.

5.4. Multi-class Classification Fusion

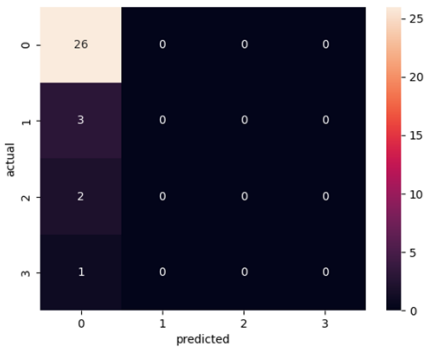

Figure 2: Thermodynamic diagram, and the best accuracy: 81.25, training loss: 0.6858118176460266, validation loss: 0.33681491017341614

ISSN 2616-5775 Vol. 6, Issue 5: 85-90, DOI: 10.25236/AJCIS.2023.060511

The experiment consists of comparison of the performance of the audio, text, and fusion models. In the text transcripts of EATD-Corpus, responses to the same question are concatenated and encoded as the average of all three layers embeddings from ELMo. A matrix of 3×1024 is obtained for each participant, where 3 is the number of questions. VGGish is used in audio processing to generate 128-dimensional audio embeddings from extracted audio data. After extracting audio and text embeddings, the dataset is divided into 2 groups (2:8), the 20% one used for validation and the remaining 80% used for training. A BiGRU model and a BiLSTM model is trained. Then, the 128- dimensional text and 128-dimensional audio representations are concatenated horizontally and fed into the multi-modal network which produces binary or 4-levels labels according to need. The proposed BiGRU model and BiLSTM model are also trained separately for comparison. Also included in the comparison are various machine learning methods, all of which use 5-fold cross validation. The experimental results are shown in Figure 2.

6. Conclusions

In this paper a multi-modal neural network method that processes audio as well as textual data is presented. It implements the BiLSTM and the BiGRU network.

This model can potentially be useful in clinical as well as public settings. It could possibly serve as an accurate but inexpensive and fast method for depression screening that could reach whole populations, including groups without much access to healthcare. Another potential use is as a rapid and economic monitoring tool, used consistently to monitor improvements or deteriorations of depression.

However, it should be noted that this model's depression identification was based on SDS scores, which although is commonly used in treatment settings and clinical trials, is not the same as a formal diagnosis of depression. This model is meant to augment existing clinical methods and not issue a formal diagnosis.

In conclusion, this paper presents a multi-modal machine learning method which combines techniques from audio processing and text processing. It is hoped this work will inspire others to build AI-based tools for understanding mental health disorders beyond depression.

References

[1] Lazar, M. and Davenport, L. (2018) "Barriers to health care access for low income families: A review of literature," Journal of Community Health Nursing, 35(1), pp. 28–37. Available at: https://doi.org/10.1080/07370016.2018.1404832.

[2] The Lancet: Prevalence and treatment of depressive disorders in China. Available at: https://news.medlive.cn/psy/info-progress/show-181716_60.html (Accessed: April 9, 2022).

[3] Yang, L. et al. (2017) "Hybrid depression classification and estimation from audio video and text information," Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge [Preprint]. Available at: https://doi.org/10.1145/3133944.3133950.

[4] de Melo, W.C., Granger, E. and Hadid, A. (2019) "Combining global and local convolutional 3D networks for detecting depression from facial expressions," 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) [Preprint]. Available at: https: // doi. org/ 10. 1109/fg.2019.8756568.

[5] Al Hanai, T., Ghassemi, M. and Glass, J. (2018) "Detecting depression with audio/text sequence modeling of interviews," Interspeech 2018 [Preprint]. Available at: https: // doi. org/ 10. 21437/ interspeech. 2018-2522.

[6] Burdisso, S.G., Errecalde, M. and Montes-y-Gómez, M. (2019) "A text classification framework for simple and effective early depression detection over social media streams," Expert Systems with Applications, 133, pp. 182–197. Available at: https://doi.org/10.1016/j.eswa. 2019.05.023.

[7] Shen, T. et al. (2018) "Cross-domain depression detection via harvesting social media," Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence [Preprint]. Available at: https://doi.org/10.24963/ijcai.2018/223.

[8] Shen, Y., Yang, H. and Lin, L. (2022) "Automatic depression detection: An emotional audio-textual corpus and a GRU/BILSTM-based model," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [Preprint]. Available at: https: // doi. org/ 10. 1109/ icassp43922. 2022.9746569.

[9] Bouchard-Cannon, P. et al. (2013) "The circadian molecular clock regulates adult hippocampal neurogenesis by controlling the timing of cell-cycle entry and exit," Cell Reports, 5(4), pp. 961–973. Available at: https://doi.org/10.1016/j.celrep.2013.10.037.

ISSN 2616-5775 Vol. 6, Issue 5: 85-90, DOI: 10.25236/AJCIS.2023.060511

[10] Kim, S.-W., Choi, S.-P. (2018) "Research on joint models for Korean word spacing and pos (part-of-speech) tagging based on bidirectional LSTM-CRF," Journal of KIISE, 45(8), pp. 792–800. Available at: https://doi.org/10.5626/jok.2018.45.8.792.
[11] Zhang, R. et al. (2022) "Bi-directional gated recurrent unit recurrent neural networks for failure prognosis of proton exchange membrane fuel cells," International Journal of Hydrogen Energy, 47(77), pp. 33027–33038. Available at: https://doi.org/10.1016/j.ijhydene.2022.07.188.