Paper ID: 3510

# Automatic Depression Detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based Model

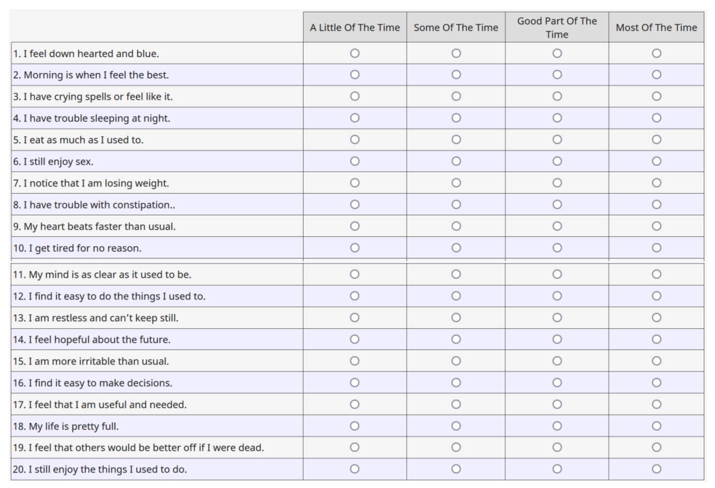
Ying Shen, Huiyu Yang, Lin Lin
School of Software Engineering, Tongji University, P. R. China
{yingshen,2031552,1931542}@tongji.edu.cn

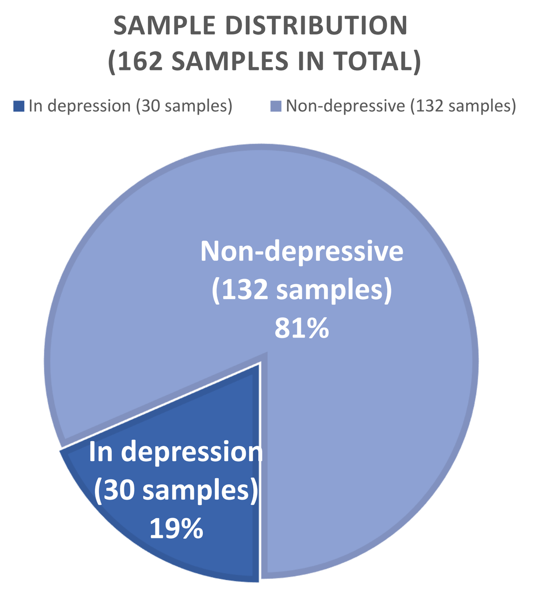
## Abstract

Depression is a global mental health problem, the worst case of which can lead to suicide. An automatic depression detection system provides great help in facilitating depression self-assessment and improving diagnostic accuracy. In this work, we propose a novel depression detection approach utilizing speech characteristics and linguistic contents from participants' interviews. In addition, we establish an Emotional Audio-Textual Depression Corpus (EATD-Corpus) which contains audios and extracted transcripts of responses from depressed and non-depressed volunteers. To the best of our knowledge, EATD-Corpus is the first and only public depression dataset that contains audio and text data in Chinese. Evaluated on two depression datasets, the proposed method achieves the state-of-the-art performances. The outperforming results demonstrate the effectiveness and generalization ability of the proposed method. The source code and EATD-Corpus are available at https://github. com/speechandlanguageprocessing/ICASSP2022-Depression.

## EATD-Corpus: a new Chinese depression dataset

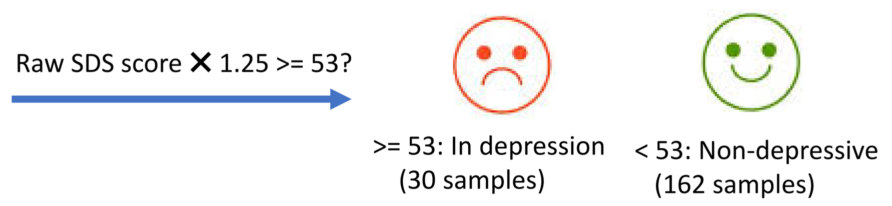
### ➢ Depressive or Not?



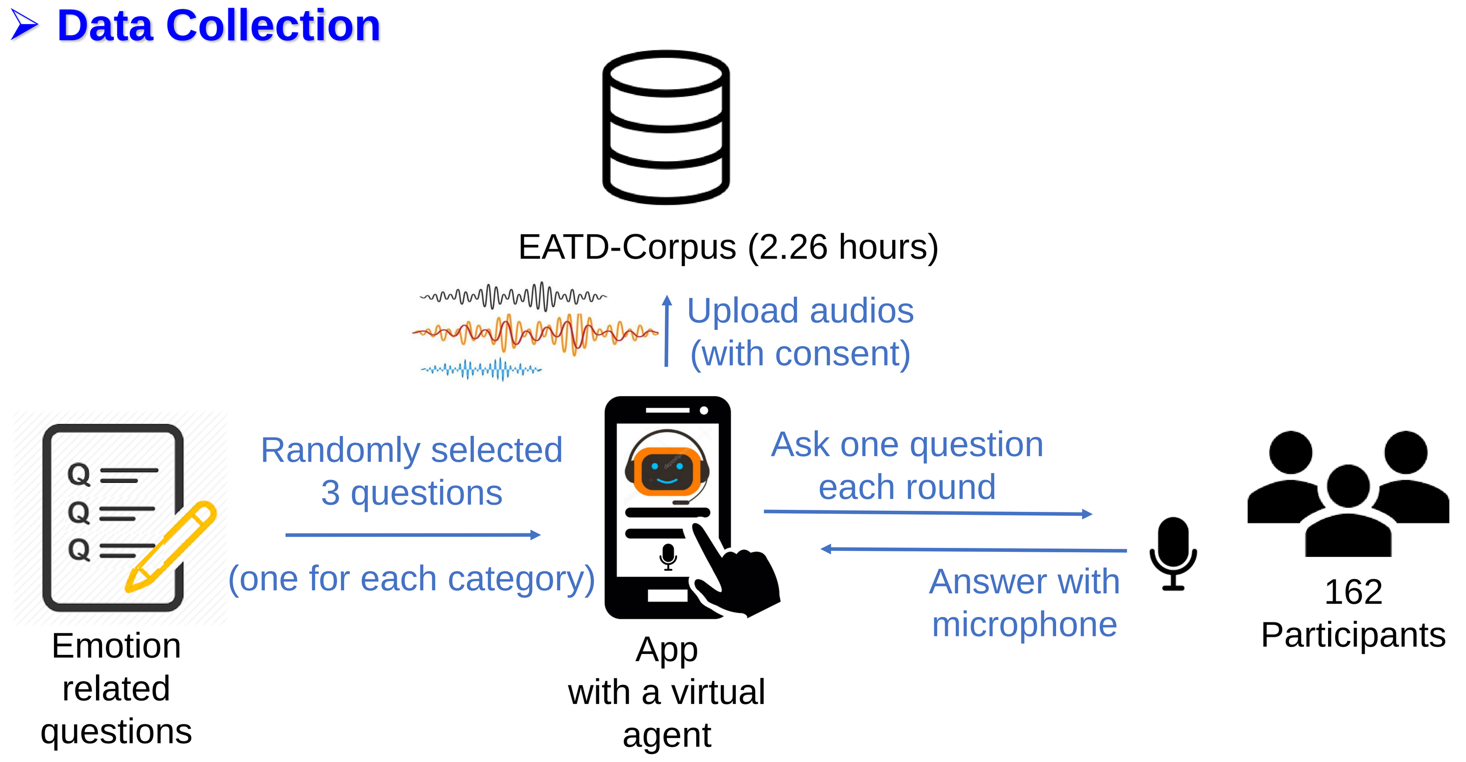



SDS questionnaire (20 items)

- ➢ the pervasive effect
- ➢ the physiological equivalents
- ➢ other disturbances
- ➢ psychomotor activities




### ➢ Data Collection

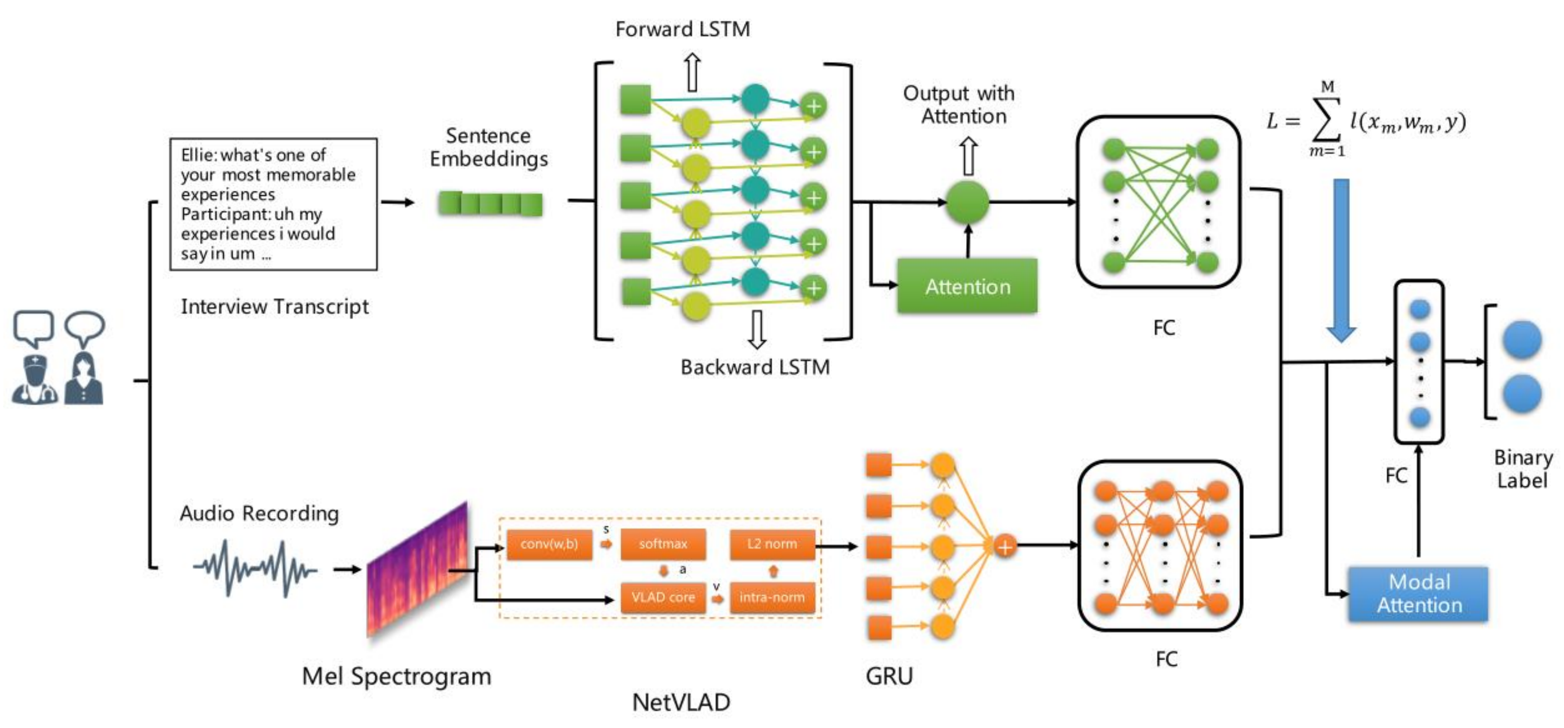



### ➢ Data Pre-processing

Several preprocessing operations have been performed on the collected audios:

- ➢ Mute audios, audios less than 1 second are removed
- ➢ Silent segments at the beginning and the end of each recording are removed
- ➢ Background noises are reduced using RNNoise with default parameters
- ➢ Kaldi is used to extract transcripts from the audios
- ➢ All the transcripts were manually checked and corrected

## A multi-modal depression detection method




### ➢ Features

In our method, text and audio features are used to prediction depression state:

- ➢ Text features
  - extracted using ELMo
  - project transcript sentences into high-dimensional sentence embeddings
- ➢ Audio features
  - Mel spectrograms are extracted from each segments
  - NetVLAD is adopted to aggregate audio embeddings from each segments

### ➢ BiLSTM with Attention Layer (for Text Features)

Attention layer is adopted to emphasize sentences contributes most in depression detection. Attention is defined in Eq. 1:

$$\mathbb{O} = \mathbf{BiLSTM}(\boldsymbol{X}), \boldsymbol{O} = \mathbb{O}_f + \mathbb{O}_b$$
$$\boldsymbol{c} = tanh(\boldsymbol{O}) \times \boldsymbol{w}, y = \boldsymbol{O} \times \boldsymbol{c} \tag{1}$$

where

- ➢ $X$ is the input text features
- ➢ $O$ consists of $O_f$ and $O_b$ representing the forward and backward output of BiLSTM respectively
- ➢ $w$ is the learned weight vector from $O$
- ➢ $c$ is the weighted context
- ➢ $y$ is the final output with attention

**Table 1**. Parameter Settings of BiLSTM Model

| Layer Name | Parameter Settings |
|---|---|
| BiLSTM | Hidden: 128<br>Layers: 2<br>Dropout: 0.5 |
| Attention | |
| Dropout | 0.5 |
| FC1 | Out features: 128<br>Activation: ReLU |
| Dropout | 0.5 |
| FC2 | Out features: 2<br>activation: ReLU |

**Table 2**. Parameter Settings of GRU Model

| Layer Name | Parameter Settings |
|---|---|
| GRU | Hidden: 256<br>Layers: 2<br>Dropout: 0.5 |
| Dropout | 0.5 |
| FC1 | Out features: 256<br>Activation: ReLU |
| Dropout | 0.5 |
| FC2 | Out features: 2<br>activation: Softmax |

### ➢ Gate Recurrent Unit Neural Network (for Audio Features)

- ➢ The GRU model summarizes the audio embeddings to audio representations
- ➢ Consists of two GRU layers, a two-layer FC network that outputs binary labels

### ➢ Multi-modal Fusion

- ➢ Representations from the last layer of the two models are concatenated horizontally.
- ➢ A weight vector is trained to represent the importance of different modalities.
- ➢ The dot product of attention vector and the concatenated representations produce the weighted representation.

$$\mathcal{L} = \sum_{m=\{audio,text\}} \ell_{ce}(\boldsymbol{x}_m, \boldsymbol{\omega}_m, y) \tag{2}$$

$$\ell_{ce} = -\frac{1}{n}\sum_{x}[y \cdot logx + (1-y) \cdot log(1-x)] \tag{3}$$

- ➢ A loss function is derived as defined in Eq. 2, where $m$ is the adopted modality, $\ell$ is the cross entropy loss function defined in Eq. 3.
- ➢ $x_m$ is the representation vectors of modal $m$, $\omega_m$ is the weight with respect to $m$, $y$ is the ground-truth

## Experiments and results

The experiments are performed on DAIC-WoZ and EATD-Corpus dataset.

### ➢ Data Imbalance

- ➢ For DAIC-WoZ dataset, group resampling is performed when training:
  - Every 10 responses of one participant are grouped
  - Samples are randomly selected from different groups until equivalence is achieved
- ➢ For EATD-Corpus, responses are rearranged to increase the size of the depressed class:
  - Each participants answered 3 questions, so there are 6 orders when rearranging

### ➢ Performance Evaluation on DAIC-WoZ Dataset

The performances of our approach together with some existing methods for depression detection are summarized in Table 3.

- ➢ Compared with the methods only adopting audio features:
  - The proposed GRU model yields the highest performance with F1 score equal to 0.77
- ➢ Compared with methods adopting only text features:
  - The proposed BiLSTM model is merely 0.01 worse than the best method
- ➢ Compared with the other method accepting both audio and text features:
  - Our multi-modal fusion method produces the best result with F1 score equal to 0.85

In addition, the Recall values of the proposed single modality models and fusion model are close to 1, indicates that our method can detects most of the depressed participants in practice.

### ➢ Performance Evaluation on EATD-Corpus Dataset

The performances of our approach together with some existing methods for depression detection are summarized in Table 4.

- ➢ When only audio features are considered:
  - the proposed GRU model achieves the best performance with F1 score equals to 0.66
- ➢ For text features:
  - our method yields the highest performance with the best F1 score 0.65
- ➢ Our fusion model exhibits a much higher performance with the F1 score increased to 0.71

Similarly, the Recall values of the fusion model have also been significantly increased to 0.84, which indicates that our method can detect most depressive cases.

**Table 3**. Results of Experiments on DAIC-WoZ dataset

| Features | Models | F1 Score | Recall | Precision |
|---|---|---|---|---|
| Audio | Gaussian Staircase Model [11] | 0.57 | - | - |
| | DepAudioNet [14] | 0.52 | 1.00 | 0.35 |
| | Multi-modal LSTM [13] | 0.63 | 0.56 | **0.71** |
| | SVM | 0.40 | 0.50 | 0.33 |
| | Decision Tree | 0.57 | 0.50 | 0.57 |
| | Proposed GRU model | **0.77** | **1.00** | 0.63 |
| Text | Multi-modal LSTM [13] | 0.67 | 0.80 | 0.57 |
| | Cascade Random Forest [8] | 0.55 | **0.89** | 0.40 |
| | Gaussian Staircase Model [11] | 0.84 | - | - |
| | SVM | 0.53 | 0.42 | 0.71 |
| | Decision Tree | 0.50 | 0.67 | 0.40 |
| | Proposed BiLSTM model | **0.83** | 0.83 | **0.83** |
| Fusion | Multi-modal LSTM [13] | 0.77 | 0.83 | 0.71 |
| | Proposed fusion model | **0.85** | **0.92** | **0.79** |

**Table 4**. Results of Experiments on EATD-Corpus

| Features | Models | F1 Score | Recall | Precision |
|---|---|---|---|---|
| Audio | Multi-modal LSTM [13] | 0.49 | 0.56 | 0.44 |
| | SVM | 0.46 | 0.41 | 0.54 |
| | RF | 0.50 | 0.53 | 0.48 |
| | Decision Tree | 0.45 | 0.44 | 0.47 |
| | Proposed GRU model | **0.66** | **0.78** | **0.57** |
| Text | Multi-modal LSTM [13] | 0.57 | 0.63 | 0.53 |
| | SVM | 0.64 | **1.00** | 0.48 |
| | RF | 0.57 | 0.53 | 0.61 |
| | Decision Tree | 0.49 | 0.43 | 0.59 |
| | Proposed BiLSTM model | **0.65** | 0.66 | **0.65** |
| Fusion | Multi-modal LSTM [13] | 0.57 | 0.67 | 0.49 |
| | Proposed fusion model | **0.71** | **0.84** | **0.62** |