


Article

Towards Automatic Depression Detection: A BiLSTM/1D CNN-Based Model

Lin Lin, Xuri Chen, Ying Shen * and Lin Zhang 

School of Software Engineering, Tongji University, Shanghai 201804, China; 1931542@tongji.edu.cn (L.L.); xurichen@tongji.edu.cn (X.C.); cslinzhang@tongji.edu.cn (L.Z.)

* Correspondence: yingshen@tongji.edu.cn

Received: 19 October 2020; Accepted: 2 December 2020; Published: 4 December 2020

Featured Application: The proposed automatic depression detection method aims at: (1) supporting clinical diagnosis with objective and quantitative measurements and (2) providing a quick, effective, and economic self depressive assessment.

Abstract: Depression is a global mental health problem, the worst cases of which can lead to self-injury or suicide. An automatic depression detection system is of great help in facilitating clinical diagnosis and early intervention of depression. In this work, we propose a new automatic depression detection method utilizing speech signals and linguistic content from patient interviews. Specifically, the proposed method consists of three components, which include a Bidirectional Long Short-Term Memory (BiLSTM) network with an attention layer to deal with linguistic content, a One-Dimensional Convolutional Neural Network (1D CNN) to deal with speech signals, and a fully connected network integrating the outputs of the previous two models to assess the depressive state. Evaluated on two publicly available datasets, our method achieves state-of-the-art performance compared with the existing methods. In addition, our method utilizes audio and text features simultaneously. Therefore, it can get rid of the misleading information provided by the patients. As a conclusion, our method can automatically evaluate the depression state and does not require an expert to conduct the psychological evaluation on site. Our method greatly improves the detection accuracy, as well as the efficiency.

Keywords: automatic depression detection; multi-modal fusion; deep learning; BiLSTM; 1D-CNN

1. Introduction

Depression is a global mental disease, whose key features include disruption in emotion experience, communication, and self regulation [1]. More than 264 million people in the world are suffering from depression. In the worst cases, depression can lead to self-harm or even suicide activities. According to the World Health Organization (WHO) reports, about 800,000 people die from severe depression every year [2]. Previous studies have suggested that intervening in the early episode of depression may be crucial to halting the development of depression symptoms [3]. However, early intervention of the disease may be difficult under certain circumstances. Firstly, traditional treatments for depression such as psychotherapy or pharmacological are timely, costly, and sometimes ineffective [4]. For individuals with financial difficulties, the cost of diagnosis and treatment is a heavy burden, which makes patients unwilling to consult a doctor. Secondly, physicians usually assess the severity of depression based on clinical interviews, rating scales, and self-assessments [5]. However, fearing public stigma and other negative consequences brought by the diagnosis, patients sometimes hide their true conditions from

psychologists intentionally [6]. Public stigma includes prejudice, stereotypical beliefs, and discriminatory behaviors towards the depressed person [7], which are the major factors inhibiting individuals with depressive symptoms from seeking help [8,9]. As a result, psychologists cannot even correctly assess the severity of depression, let alone make effective interventions. In view of these, an efficient automatic depression detection system becomes a necessity, which can help potential patients assess their conditions privately and enhance their willingness to ask for psychologists' help. In addition, an effective automatic depression detection system provides great help to psychologists in the process of diagnosis, especially when the patients are intentionally misleading. Therefore, early intervention can be achieved by providing an objective depression detection system to both mental health professionals and patients [10].

Studies have shown that both nonverbal and verbal behaviors are affected by depression, including facial expressions [11], prosody [1,12–14], syntax [15], and semantics [16]. Motivated by these theories and studies, an automatic depression detection system analogizes clinical diagnosis in which verbal representations, facial expressions, and vocal characteristics are analyzed. Currently, approaches to automatic depression detection often utilize information from multiple sources, such as audio, videos, and text extracted from responses [17–20].

The process of automatic depression detection consists of two steps. In the first step, response audio and/or videos are collected from the participants that are asked a set of questions. Sometimes, text content is also extracted from the audio and videos to improve diagnostic accuracy. In the second step, depression severity is automatically analyzed and assessed using algorithms based on the collected information, which includes voice features, reply content, etc.

Although researchers have made some progress in improving diagnostic accuracy, great difficulties still exist in practice. First of all, videos of clinical interviews may not be available due to the privacy problem. Secondly, patients may incorrectly report their mental states unconsciously or intentionally, which can mislead the diagnosis. Thirdly, how to extract and fuse representative features from different sources needs more investigation. Therefore, automatic depression detection remains a challenging task for researchers in this field.

In this work, we investigate the problem of automatic depression detection and introduce an automatic depression detection method based on the BiLSTM and 1D CNN models to predict the presence of depression, as well as to assess the severity of depressive symptoms. Evaluated on two publicly available datasets, namely the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WoZ) dataset and the Audio-Visual Depressive Language Corpus (AViD-Corpus) dataset, our method achieves a high performance in the tasks of depression detection and depression severity assessment.

Based on the proposed method, our long-term goals include: (1) providing a quick, effective, and economic self depressive assessment and (2) supporting the clinical diagnosis with objective and quantitative measurements.

2. Related Work

Automatic depression detection did not attract much attention until 2009 [21]. Cohn et al. [21] extracted manual Facial Action Coding System (FACS) features, Active Appearance Modeling (AAM), and pitches to measure facial and vocal expression. They used simple classifiers (i.e., SVM and logistic regression) to assess depression severity and reached a high accuracy. Since then, automatic depression detection based on machine learning techniques has aroused increasing interest from researchers [11,22–26].

In general, methods of automatic depression detection first extract different types of features from interview audio/videos of patients who are asked a set of carefully crafted questions from different topics. Then, models are trained using the extracted features to predict the presence of depression or to assess the severity of depression for the patients.

Early studies of automatic depression detection devoted great efforts to extracting effective features from closely correlated questions designed for interviews. In [27], Arroll et al. demonstrated that some key questions (e.g., “Do you need help?”) can improve diagnostic accuracy. Yang et al. [28] conducted content analysis of transcripts to manually select depression related questions. They constructed a decision tree based on the selected questions to predict the patients’ depressive conditions. Similarly, Sun et al. [29] analyzed the interview transcripts of participants and extracted text features from the questions related to certain topics (e.g., sleep quality, introversion, parents, etc.). They used random forest to detect depression tendency. Gong and Poellabauer [30] performed context-aware analysis with topic modeling to preserve important temporal details in long interviews. Utilizing the Gaussian staircase regressor, Williamson et al. [17] achieved a good performance by analyzing the semantic context to obtain coarse depressive descriptors.

Inspired by the emerging deep learning techniques, Mendels et al. [31] proposed a single hybrid deep model, which was trained jointly with the acoustic and linguistic features. They found that using deep networks to fuse features could achieve better results for deception detection. Deep learning approaches to fuse multi-modal features in depression detection appear to be particularly promising. Yang et al. [18] proposed a depression detection method based on the Deep Convolution Neural Network (DCNN) and the Deep Neural Network (DNN). They also designed a set of new video and audio descriptors to train the model. Alhanai et al. [19] used the Long Short-Term Memory (LSTM) network to assess depression tendency. They used text features and audio features that were highly correlated with depression severity. Lam et al. [32] proposed a data augmentation procedure based on topic modeling. They first labeled each participant’s transcript and audio with corresponding topics. Then, for each participant, a subset of the labeled data was extracted and combined to make a new training sample. They used a deep learning model, namely Transformer, for text feature modeling and a deep One-Dimensional Convolutional Neural Network (1D CNN) for audio feature modeling. The combination of the two models achieved state-of-the-art performance. Based on LSTM and CNN, Ma et al. [33] encoded the depression related temporal clues in the vocal modality to predict the presence of depression. To address the data imbalance issue, they performed random over-sampling on audio recordings. Haque [20] utilized a causal CNN to predict depressive symptom severity by summarizing embeddings, which captured long-term acoustic, visual, and linguistic elements. In [34], Dinkel et al. put forward a text-based multi-task BiLSTM model with pre-trained word embeddings and acquired a high performance on the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WoZ) dataset.

Our Motivations and Contributions

However, current methods have their own weaknesses. To better predict depression severity, the methods proposed in [17,28–30] have to design/select a set of effective questions (e.g., questions related to sleep quality, post-traumatic stress disorder, etc.) that can better reveal patients’ conditions. These questions are closely related to psychologists’ expertise, which is not quite easy to obtain. The method in [32] categorized the queries into seven main topics referring to the Patient Health Questionnaire (PHQ-8) questions. The PHQ-8 questions refer to the 8 questions that constitute the PHQ-8 questionnaire (see supplementary material). The eight questions are designed to investigate how often in the past two weeks the patient suffered from loss of pleasure, low mood, sleep disturbance, lack of energy, eating disorders, low self-esteem, trouble concentrating, and slow movements. The features used for depression detection were extracted from patients’ responses sorted by the topics. However, in practice, patients may refuse to respond to the selected questions/topics. If any one of the selected questions/topics is not referred to during the clinical interview, these methods will not be applicable because they cannot construct a complete feature set. As a conclusion, when applied in evaluating new patients’ conditions, these methods

may fail. To avoid relying on the clinicians' expertise and to increase the generalization capability, some methods did not adopt question/topic selection procedures [19,20,33]. Instead, they extracted features from responses to the universal questions. However, their performances are not comparable to their counterparts, which are topic related.

The above-mentioned drawbacks prompted us to propose a new effective depression detection method with a better generalization ability, a higher performance, and less dependency on expertise.

In this work, we propose a new multi-modal depression detection method. Our method utilizes audio features extracted from interview audio, as well text features extracted from patients' responses. It exploits the powerful learning ability of 1D CNN and the bidirectional long short-term memory model with an attention layer to summarize embeddings from audio and text features, respectively. In order to be applicable to various datasets and independent of expertise, the proposed method extracts audio/text features from universal questions instead of manually selected questions/topics. To achieve a better performance, the proposed method integrates two types of embeddings using a multi-modal fusion network and obtains good performances on the two evaluated datasets.

Compared with the method described in [32] which also used deep neural networks to fuse different modalities, our framework is different in the adoption of an effective data balancing method, as well as two more powerful deep learning models. More importantly, our proposed method can be applied to various depression related datasets, while the method in [32] was designed only for topic related depression datasets.

Our major contributions are summarized as follows:

(1) We propose a deep learning-based depression detection model accepting speech signals and linguistic content. The proposed method utilizes a 1D CNN model and a BiLSTM model with an attention layer to process audio features and text features, respectively. The two types of embeddings summarized from the 1D CNN and BiLSTM models are concatenated and passed to one-layer fully connected (FC) networks, one of which predicts the presence of depression and the other the depression severity of the patients. The proposed method has no restrictions on the types of questions to be asked in an interview and thus obtains a better generalization capability. These make our method applicable in various situations.

(2) A key issue of the problem is how to design and extract representative features. Through extensive comparison experiments, we selected a set of efficient audio features and text features. The features were extracted from universal questions instead of manually selected questions/topics. This obviates the necessity of expertise in the proposed depression detection method. It should be noted that our approach uses low-level descriptors such as Mel spectrograms and sentence embeddings. Compared with prior approaches, which employed high-level descriptors such as word frequency [30,35], our method has a better discriminative power.

(3) In most depression related datasets, the number of non-depressed individuals far exceeds that of the depressed ones, the ratio of which is around 3:1 to 8:1. If audio/text samples are sampled from each individual with equal frequency, this will result in an imbalanced dataset and introduce bias into the model training process. To address the data imbalance issue, we propose a new data sampling method to increase the number of samples in the minority class, i.e., samples of the depressed individuals. The presented solution resamples transcripts and audio clips non-redundantly from the depressed class until the samples from the depressed class and the non-depressed class are balanced. The advantage of the presented sampling solution lies in that it increases the robustness of our model and improves the prediction accuracy.

3. Proposed Approach and Models

Our approach consists of a 1D CNN model and a BiLSTM model with an attention layer, which summarize the inputs from different modalities into embeddings. Instead of manually selecting and organizing features to preserve discriminative information, 1D CNN and BiLSTM with an attention layer are able to learn the important characteristics from voice descriptors and sentence inputs.

Multimodal information integration is achieved by the horizontal concatenation of embeddings. The concatenated audio/text embeddings capture short-term, as well as long-term acoustic and linguistic characteristics to distinguish the depressed patients and healthy individuals. In the end, embeddings from 1D CNN and BiLSTM models are concatenated and delivered to two one-layer FC networks, one of which outputs a label indicating the presence of depression and the other of which assesses the depression severity.

The framework of the proposed method is shown in Figure 1.

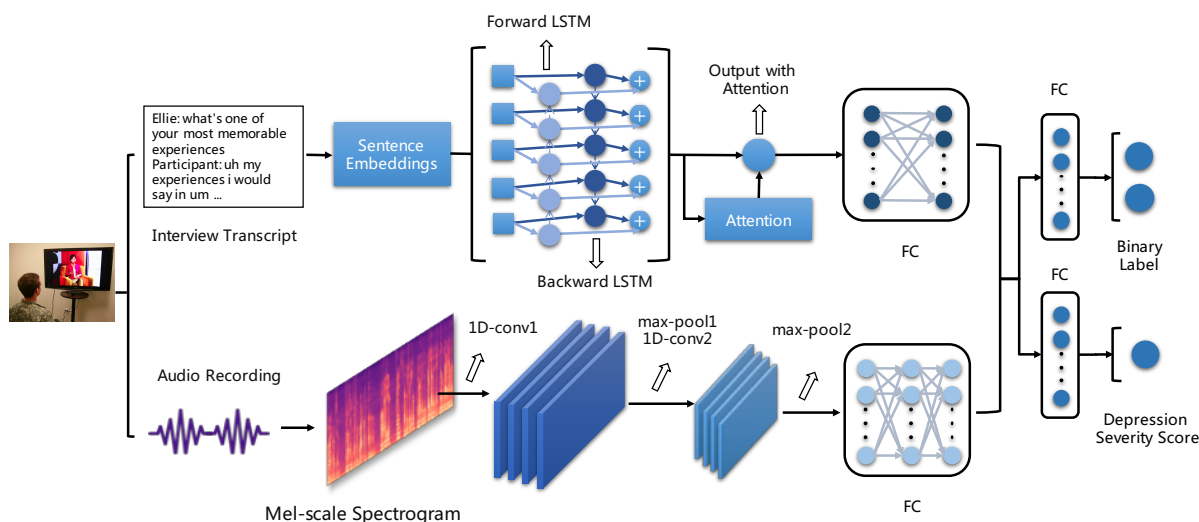


Figure 1. Framework of the proposed model. Text features are trained using a BiLSTM model with an attention layer, and audio features are trained with a 1D CNN model. Outputs from both BiLSTM and 1D CNN models are concatenated and fed into two fully connected networks (FC), one of which generates labels indicating whether the interviewees are depressed or not, and the other predicts depression severity scores, respectively.

3.1. Audio Features

The commonly used voice descriptors in depression detection include the Mel Frequency Cepstrum Coefficient (MFCC), 79 COVAREP features [36] and Mel spectrograms.

Compared with MFCC and COVAREP features, Mel spectrograms produce better results in comparison experiments (see Section 6.3 for more details). Because Mel spectrograms can preserve detailed information that exhibits a better discrimination ability, it was finally adopted in our approach.

Mel spectrograms are computed by multiplying short-time Fourier transform magnitude coefficients with the corresponding Mel filters [33]. Thus, it can be regarded as a non-linear transformation of the spectrogram, which maintains high level details in sound. The relationship between the normal spectrogram and Mel spectrograms is depicted in Equation (1), where f is the frequency of the spectrogram.

$$mel(f) = 2595 \times \log_{10}(1 + f/700) \quad (1)$$

A visualization example of the spectrogram and Mel spectrograms is shown in Figure 2.

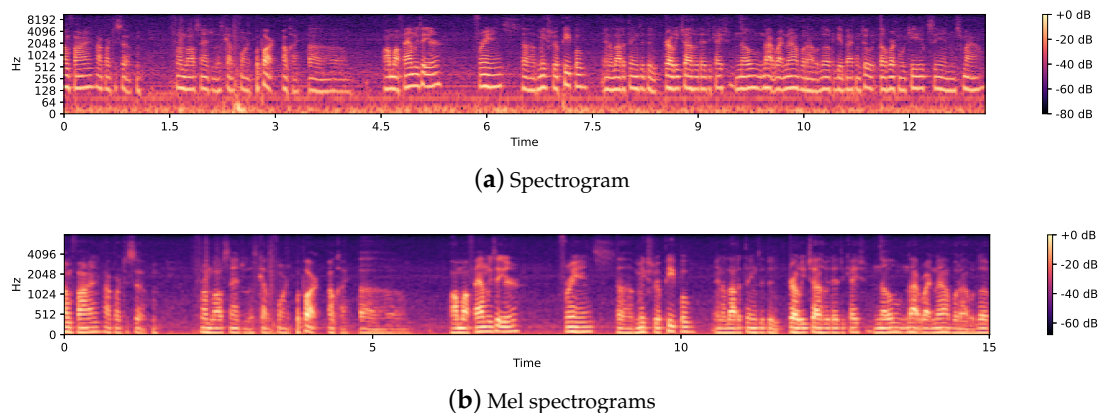


Figure 2. An example of (a) spectrogram and (b) Mel spectrogram features generated from a clip of audio.

3.2. Text Features

Since depressive text data are few and not easy to obtain, it is hard to train word embeddings from scratch. An alternative solution is to use a pre-trained language model on the large datasets. Two popular language models, Bert and ELMo, are investigated. ELMo encodes different types of syntactic and semantic information about words-in-context [37]. Bert pre-trains deep bidirectional word representations by jointly conditioning on both left and right contexts in all layers [38]. In the feature comparison experiments, ELMo achieves the best performance among three candidate language models (see Section 6.3 for more details). Therefore, it is adopted in the proposed method. Following the recommendation in [37], the average of all three layer embeddings in ELMo is taken as the sentence representation.

3.3. Data Imbalance

As mentioned in the previous section, data imbalance heavily exists in the depression related datasets, which introduces non-depressed preference during classification. Therefore, the sizes of two (i.e., depressed and non-depressed) classes need to be equivalent through sampling techniques to reduce prediction bias. In addition, a larger volume of signals from an individual may emphasize some characteristics that are person specific. Therefore, it is necessary to unify the length of the responses of each individual.

In order to balance the data, the approach in [32] categorized topics in interviews and labeled each participant's transcript and audio data with corresponding topics. Then, a subset of the labeled data was extracted and combined into a new training sample. This operation was done on both depressed and non-depressed participants to achieve a balanced dataset. The method in [33] performed **random over-sampling** on audio recordings to address the data imbalance issue. However, the topic modeling method is not applicable in our approach because we do not categorize questions by topics, and the random over-sampling is only feasible for the audio data, nor the text data.

Therefore, we propose a new data resampling method to balance the training samples from two classes.

For the audio features, the recordings of participants are cropped into a number of slices. Audio slices of depressed samples are resampled without redundancy until their number is equivalent to that of non-depressed ones.

For the text features, every 10 responses of a participant are grouped together. Similar to audio sampling, text samples are randomly selected from different groups of depressed patients' responses. The process is repeated until the numbers of text samples in the two classes are equivalent. In this way,

a balanced training set can be constructed. For example, a new balanced dataset consisting of 77 depressed samples and 77 non-depressed samples is constructed from the DAIC-WoZ dataset.

3.4. One-Dimensional Convolutional Neural Network

A convolutional layer in CNN convolves over the width and the height in two dimensions. However, it is not applicable when the inputs are Mel spectrograms, of which the width and the height represent the time and the frequency, respectively. Compared with two-dimensional convolution, one-dimensional convolution is more suitable when dealing with Mel spectrograms in the depression detection problem. In the proposed model, convolution over the frequency axis is used. This allows the model to generate features that capture a short-term temporal correlation. In other words, 1D convolution over the frequency axis aims to acquire the persistence of characteristics in Mel spectrograms.

Specifically, 1D convolution sets the filter size to one along the axis other than the convolution axis. For instance, if 1D convolution is along the frequency axis of Mel spectrograms, the filter size should be set to $(1, k)$, where k is the kernel size along the frequency axis.

The detailed configuration of the 1D CNN model used in the experiments is shown in Table 1. The “Out features” of FC1, FC2, and FC3 in Table 1 represents the number of nodes in the fully connected layers FC1, FC2, and FC3.

Table 1. Parameter settings of the 1D CNN model.

| Layer Name | Parameter Settings (Classification/Regression) | Activation |
|------------|---|----------------|
| Conv1 | Kernel: [1, 7] Stride: 1 Filter size: 32 | ReLU |
| Max-pool1 | Kernel: [4, 3] Stride: [1, 3] | |
| Conv2 | Kernel: [1, 7] Stride: 1/2 Filter size: 32 | ReLU |
| Max-pool2 | Kernel: [1, 3] Stride: [1, 3] | |
| FC1 | Out features: 128 | ReLU |
| FC2 | Out features: 128 | ReLU |
| dropout | 0.5 | |
| FC3 | Out features: 2/1 | Softmax/Linear |

The proposed 1D CNN model consists of two 1D convolutional layers, two pooling layers, and a three-layer FC network. The final activation function and the kernel stride in the second convolutional layer vary depending on different evaluation tasks. For the depression severity assessment task (regression), the linear activation function and a kernel stride of two are used. For the depression prediction task (classification), softmax and a kernel stride of one are used.

3.5. Bidirectional Long Short-Term Memory with Attention

The Recurrent Neural Network (RNN) was designed for processing sequence data. The strength of RNN lies in its ability to discover time relevance and logical characteristics by passing the output to the hidden layer again. As a variant of RNN, LSTM adopts the gate mechanism to solve the problem of the

vanishing gradient caused by the long-term dependency. BiLSTM enables the model to learn not only from the past time steps, but also from the future time steps. Since we wish to extract the characteristics of response sequences in the interview, we utilized BiLSTM and additionally adopted an attention layer to emphasize which sentence contributes most to the depression detection problem.

Attention is defined in Equation (2), where X is the input sentence embeddings with size $(seq_lens, embed_dim)$, \mathbb{O} is the output of BiLSTM and is of size $(seq_lens, hidden \times layers)$, $hidden$ and $layers$ are the hyperparameters of BiLSTM, \mathbb{O}_f and \mathbb{O}_b are the forward and backward outputs of BiLSTM, respectively, \mathbb{O} is the concatenation of \mathbb{O}_f and \mathbb{O}_b , \mathbf{O} is the sum of \mathbb{O}_f and \mathbb{O}_b and is of size $seq_lens \times hidden$, w is the weight vector of length $hidden$ and is generated from a fully connected layer, W and b are the parameters of the fully connected layer, c is the weighted context, and y is the final output with attention.

$$\begin{aligned}
 \mathbb{O}, H &= \text{BiLSTM}(X) \\
 \mathbb{O} &= [\mathbb{O}_f, \mathbb{O}_b] \\
 \mathbf{O} &= \mathbb{O}_f + \mathbb{O}_b \\
 w &= W \times \mathbf{O} + b \\
 c &= \tanh(\mathbf{O}) \times w \\
 y &= \mathbf{O} \times c
 \end{aligned} \tag{2}$$

Equation (2) defines a simple per-sequence attention mechanism.

The detailed configuration of the proposed BiLSTM model with an attention layer is summarized in Table 2. The “Out features” of FC1 and FC2 in Table 2 represent the number of nodes in the fully connected layers FC1 and FC2. The model consists of two BiLSTM layers, followed by an attention layer to calculate the weighted representations. The output of the attention layer is fed into a two-layer FC network to produce a prediction of whether the participant is depressive or the severity of depression.

Table 2. Parameter settings of the BiLSTM model with an attention layer.

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

3.6. Multi-Modal Fusion

To integrate different types of information extracted from audio and text content, embeddings generated from the last layer of both the BiLSTM model and the 1D CNN model are concatenated

horizontally. The concatenated embeddings are then passed to a one-layer FC network, which serves as a merge node. The concatenation of two types of embeddings is defined in Equation (3).

$$\begin{aligned} \mathbf{a} &= \text{BiLSTM}(X_{\text{embedds}}) \\ \mathbf{b} &= \text{CNN}_{1\text{D}}(X_{\text{spec}}) \\ \mathbf{x}_{\text{fuse}} &= [a_1, a_2, \dots, b_1, b_2, \dots] \end{aligned} \quad (3)$$

In the depression detection problem, there are two modalities, i.e., audio and text embeddings, that actually contribute to the final prediction. The loss function is able to measure the contributions of the two modalities separately. Similar to [32], we opted to derive a loss function that updates the parameters based on the contributions of different modalities during the back propagation process. The derived loss function is defined in Equation (4), where M is the number of modalities ($M = 2$ in this paper), ℓ is the criterion measuring the difference between the prediction and the ground-truth, x_m is the output features of the m -th modality, ω_m is the weight of the FC network with respect to the m -th modality, and y is the ground-truth.

$$\mathcal{L} = \sum_{m=1}^M \ell(x_m, \omega_m, y) \quad (4)$$

In particular, ℓ varies depending on the evaluation tasks. In the task of classifying depressed patients, Huber loss is utilized. As defined in Equation (5), Huber loss is a smooth L1 loss with great robustness to the outliers. It is a trade-off between the mean squared error and mean absolute error.

In the task of depression severity assessment, cross entropy loss, as defined in Equation (6), is chosen to be the criterion function.

$$\ell_{\text{huber}} = \begin{cases} 0.5(x - y)^2 & |x - y| < 1 \\ |x - y| - 0.5 & \text{otherwise} \end{cases} \quad (5)$$

$$\ell_{\text{ce}} = -\frac{1}{n} \sum_x [y \cdot \log x + (1 - y) \cdot (1 - x)] \quad (6)$$

3.7. Overall Pipeline

3.7.1. Training Phase

Denote \mathcal{M}_{cnn} , $\mathcal{M}_{\text{lstm}}$, and $\mathcal{M}_{\text{fuse}}$ as the trained 1D CNN model, the BiLSTM model with an attention layer, and the multi-modal fusion network, respectively. Given the recordings and transcripts of a participant in the training set (as shown in Figure 1), the recordings are first cropped into audio slices of a certain length. Mel spectrograms are then derived from the selected audio slices. For interview transcripts, the responses of the participant are extracted and encoded into a sentence embedding matrix. Each row of the matrix represents a sentence embedding of a response. If the participant belongs to the depressed class, resampling will be performed on both audio slices, and responses using these methods are illustrated in Section 3.3 to balance the numbers of depressed samples and non-depressed samples. Mel spectrograms and the sentence embedding matrix are fed into a 1D CNN model and a BiLSTM model with an attention layer separately.

After training the 1D CNN and BiLSTM models, the last layer of \mathcal{M}_{cnn} and $\mathcal{M}_{\text{lstm}}$ are concatenated to an FC network. The training process is repeated on the FC network, while the parameters of \mathcal{M}_{cnn} and $\mathcal{M}_{\text{lstm}}$ are frozen. Finally, we obtain a multi-modal fusion network $\mathcal{M}_{\text{fuse}}$.

3.7.2. Test Phase

Given the recordings and transcripts of a new participant, the feature extraction process is similar to that in the training phase. The only difference lies in that resampling is not performed on the test set.

The extracted Mel spectrograms and the sentence embedding matrix are fed into \mathcal{M}_{fuse} , which can either predict the presence of depression or assess the depression severity.

4. Dataset

4.1. Depression Datasets

There exist several datasets designed for the depression detection task [1,39–44]. Among these datasets, the most popular ones are the DAIC-WoZ and Audio-Visual Depressive Language Corpus (AViD-Corpus) datasets, because they are the only two publicly available datasets.

The DAIC-WoZ dataset contains clinical interviews designed to support the diagnosis of psychological distress conditions [39]. DAIC-WoZ involves recordings and transcriptions of 142 participants that went through clinical interviews with a computer agent. For each participant, DAIC-WoZ provides a Patient Health Questionnaire (PHQ-8) [45] score, which indicates the depression severity. In addition, a binary label of the PHQ-8 score is also provided to represent the presence of depression. A score greater than or equal to 10 indicates that the participant is suffering from depression. The DAIC-WoZ dataset is split into a training set (107 participants; 30 are depressed, and 77 are non-depressed), a development set (35 participants; 12 are depressed, and 23 are non-depressed), and a test set according to [39,46]. Labels in the training set and the development set are publicly available, while those in the test set were not given by the authors.

The second dataset is AViD-Corpus [40,47]. In AViD-Corpus, eighty-four participants were required to answer a series of queries (free-form part) in a human-computer interaction. In addition, they were requested to recite an excerpt of a fable (north wind part). The training, validation, and test sets were split according to [47], and the corresponding labels were all available. For AViD-Corpus, the model was trained on the training set and evaluated on the test set. Participants were labeled with Beck Depression Index-II (BDI-II) [48] scores. BDI-II is a self-reported 21 multiple choice depression inventory, whose scores range from 0–63. If the BDI-II score of a participant is over 29, this indicates the presence of depression.

4.2. Evaluation Task

There are two tasks when evaluating automatic depression detection methods, which include: (1) the severity of depression indicated by the PHQ-8 score or the BDI-II score; (2) the presence of depression. In the first task, a regression model was learned and used to predict the depression severity score of the patient. In the second task, a classification model was trained to predict whether the patient was depressed or not.

The DAIC-WoZ dataset provides audio recordings, videos, and transcripts of clinical interviews. Depression severity in DAIC-WoZ is measured by the PHQ-8 score. With the PHQ-8 score equal to or larger than 10, an individual is regarded to have depression symptoms. Taking the privacy problem and practical accessibility into consideration, the video data in this dataset were not utilized. For the DAIC-WoZ dataset, the model was trained on the training set and evaluated on the validation set.

AViD-Corpus provides audio recordings and video data, but no transcripts. For the same reason mentioned above, the video data were not utilized. BDI-II score is the depressive state measurement used in AViD-Corpus. A BDI-II score equal to or larger than 29 indicates that the patient is suffering from depression. The 1D CNN model for dealing with audio features was trained on the training set and evaluated on the test set.

5. Experiments

Evaluation experiments were performed on 2 datasets, namely DAIC-WoZ and AVID-Corpus, respectively.

First, a classification model predicting the presence of depression was trained using the training set of DAIC-WoZ. The trained model was evaluated on the validation set instead of the test set because the labels for the test set were not available. In the same way, a regression model predicting the severity of depression symptoms was trained and evaluated on the validation set of DAIC-WoZ.

Second, a regression model that assesses the depression severity was trained using the training set of AVID-Corpus. The classification model was not trained because AVID-Corpus does not provide binary labels. Because the transcripts of participants' responses were not available in AVID-Corpus, only audio features were extracted to train the 1D CNN model.

Third, comparison experiments for audio/text feature selection were conducted on the DAIC-WoZ dataset. The results of these experiments revealed the effectiveness of different audio/text features in depression detection.

In addition, validation experiments for the proposed data resampling method were performed on the DAIC-WoZ dataset.

The above models and experiments were implemented and conducted using Pytorch.

5.1. Data Processing

5.1.1. Audio Features

For the DAIC-WoZ dataset, the voice of each participant was first segmented from the whole recording, following the timestamps documented in the transcripts. The durations of segmented recordings ranged from 5 to 25 min. To avoid introducing some person specific characteristics, the lengths of segmented recordings were unified by cropping recordings into slices of a certain length. The length of slices was expected to be neither too long, nor too short. It should be long enough to extract subtle speech characteristics displayed by depressed individuals and short enough to ensure that the 1D CNN model is trainable. In order to discover the proper length, an audio length comparison experiment was conducted. Inspired by the exploratory trial in [49], the length of the audio clips was set to 4, 10, 15, 20, and 25 s in the experiment. The results indicate that the 1D CNN model with 15 s audio slices shows the best performance, for which the precision and recall are 0.77 and 0.83, respectively. Therefore, the segmented recordings of each participant were cropped into a number of 15 second slices.

Resampling was performed on each audio of depressed samples until the number of depressed samples was equivalent to that of the non-depressed ones. Audio slices were converted to Mel spectrograms with 80 Mel filters. After that, a min-max normalization was conducted on all of the Mel spectrograms.

Considering the question-answer nature of depression detection, the free-form part of AVID-Corpus was utilized. First, the silent part at the beginning and the end of each recording was removed. Next, the processed recordings were split into 15 second slices. The recordings in AVID-Corpus range from 6 s to 4 min. However, our method requires 15 second audio slices for audio feature extraction. Therefore, for recordings less than 15 s, silent audio segments were padded to extend the audio lengths to 15 s. The way of padding the original recording with the silent audio segment was to concatenate the original recording array with a silent segment array. The values in the silent segment array are all 0.001, and the length of the silent segment array is $sr \cdot (15 - l_{rec})$, where sr is the sample rate and l_{rec} is the length of the original recording array. The rest of the feature extracted procedure was the same as the procedure for the DAIC-WoZ dataset.

It should be noted that resampling was only performed on the training set for the depression datasets.

5.1.2. Text Features

The transcripts in the DAIC-WoZ dataset contain depressive text data. For one query (e.g., “What’s one of your most memorable experiences”), participants may have more than one response (e.g., “uh my experiences i would say in um” and “europe visiting moreso i’ve also visited”). Every response of a participant was extracted and preprocessed. An example of the preprocessing is removing insignificant meta information (such as [scrubbed_entry], which represents the removal of identifying information). Responses to the same question were concatenated to a long sentence, which was then encoded as the average of all three layer embeddings of ELMo. A response matrix of $N \times 1024$ was obtained per participant, where N is the number of queries. To solve the data imbalance problem, the response matrix was divided into m matrices of 10×1024 in size, where m is an integer of N divided by 10. Resampling was performed on the depressed individuals evenly until the numbers of samples in the two classes were equivalent.

5.2. Experiment 1: Performance Evaluation on the DAIC-WoZ Dataset

After extracting audio features and text features from the DAIC-WoZ dataset, the 1D CNN model and the BiLSTM model with an attention layer were trained separately. We connected the output layers of the two models to an FC layer, which served as a multi-modal fusion network. During the training process, the 128-dimensional embeddings generated by the BiLSTM and 1D CNN models were concatenated horizontally to produce 256-dimensional embeddings, which were then passed to the multi-modal fusion network. The multi-modal network was used to detect the depressive state of participants, including predicting PHQ-8 scores or binary labels indicating the presence of depression.

The 1D CNN model was trained with the Adadelata optimizer with a default learning rate equal to 1 and the batch size equal to 4 for 30 epochs. The BiLSTM model with an attention layer was trained with the Adam optimizer with the learning rate equal to 1×10^{-4} and the batch size equal to 8 for 100 epochs. For the multi-modal network, the derived loss function was adopted, and the optimizer was Adam with the learning rate equal to 1×10^{-4} and the batch size equal to 2. It should be noted that during the training process of the multi-modal network, only the parameters of 256 neurons were updated.

To compare with other methods, performances were evaluated on the development set. For the classification task, F1 score, recall, and precision values are reported. For the regression task, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were compared.

5.3. Experiment 2: Performance Evaluation on the AViD-Corpus Dataset

In this experiment, the performance of our method was further evaluated on the AViD-Corpus dataset. The audio features were extracted from AViD-Corpus. The training set was used for training the 1D CNN model, and the test set was used for evaluating the performance. The 1D CNN model was trained with the Huber loss and Adam optimizer with the learning rate equal to 1×10^{-3} and the batch size equal to 2. The performance on the test set is reported based on the MAE and RMSE.

5.4. Experiment 3: Audio/Text Feature Comparison Experiment

Extended experiments were performed to identify optimal audio/text features for depression detection.

For audio features, we compared the prediction accuracy using the audio features of MFCC, 79 COVAREP features, and Mel spectrograms generated from the DAIC-WoZ dataset. The feature extraction procedure of MFCC was the same as for the Mel spectrograms, with a window length of 0.25 s

and a window step of 0.1 s. As for the COVAREP features, the higher order statistics of the mean, maximum, and minimum were extracted per 0.1 s, resulting in a feature matrix of 150×237 . The classification model of 1D CNN was trained with different audio descriptors to find the optimal audio feature.

For text features, two pre-trained language models, i.e., Bert and ELMo, were compared. Concatenated sentences were encoded with different language models. Two Bert models, uncased Bert-Large with 24 layers, 1024 hidden, and uncased Bert-Base with 12 layers, 768 hidden, were tested in this experiment. The BiLSTM model with an attention layer was trained as a classification model to determine the optimal text feature.

5.5. Experiment 4: Effectiveness Evaluation of the Proposed Data Resampling Method

Additional validation experiments were conducted to verify the effectiveness of the proposed data resampling method for addressing the data imbalance issue. In this experiment, the 1D CNN model and the BiLSTM with an attention layer model were trained as classification models using the balanced training set and the unbalanced training set, respectively. The performances were further evaluated on the DAIC-WoZ dataset.

First, two audio feature sets, namely a balanced audio training set and an unbalanced audio training set, were constructed for the proposed 1D CNN model. The balanced audio training set consisted of Mel spectrograms extracted from the resampled audio clips, while the unbalanced audio training set contained Mel spectrograms extracted from audio clips without resampling. For example, the unbalanced audio training set consisted of 30 Mel spectrograms from the depressed class and 77 Mel spectrogram from the non-depressed class. As a comparison, the balanced audio training set contained 77 Mel spectrograms from the depressed class and 77 Mel spectrograms from the non-depressed class. Similarly, an unbalanced text training set and a balanced text training set consisting of ELMo embeddings were constructed for the BiLSTM with an attention layer model. Next, the 1D CNN model and the BiLSTM with an attention layer model were trained using the corresponding balanced and unbalanced audio/text training sets. Finally, the performances of four trained models evaluated on the development set are summarized and compared.

6. Results

The results of the four experiments are shown in Tables 3–6, respectively. The results of the comparative methods evaluated on the DAIC-WoZ dataset are also listed in Table 3.

6.1. Results of Experiment 1

In the first experiment, the performance of our method was compared with other methods on the DAIC-WoZ dataset. Because different methods adopt different types of features, we compared the proposed 1D CNN model with six other methods that only accept audio features. Similarly, we compared the proposed BiLSTM model with five methods that only accept text features. In addition, we compared our method with two traditional classifiers, namely SVM and decision tree. In the end, the performance of our proposed multi-model fusion network was compared with two methods that also accept both audio and text features. It should be noted that the data sampling method was only used to increase the size of the training samples in the minority class. Therefore, the different data sampling methods did not directly affect the performance of the methods to be compared. Besides, the data sampling step was only performed on the training set, not on the development set. During the evaluation, none of the methods performed any data sampling.

From Table 3, it can be seen that, for the methods utilizing only one type of features, the methods based on the text features perform better than those based on the audio features in the both depression classification task and depression severity assessment task. Compared with the methods adopting only

audio features, the proposed 1D CNN model yields the highest performance with an F1 score of 0.81 and an MAE of 4.25. Compared with the proposed 1D CNN model, the proposed BiLSTM model with an attention layer based on text features performs even better, with its F1 score equal to 0.83 and the MAE value equal to 3.88. Compared with other methods adopting only text features, the proposed BiLSTM model achieves the second best performance in the classification task, which is merely 0.01 worse than the best method in the F1 score and 0.58/0.98 worse in the MAE/RMSE values.

Table 3. Results of the experiments on the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WoZ) dataset.

| Features | Models | Classification | | | Regression | |
|--------------|-------------------------------------|----------------|-------------|-----------|-------------|-------------|
| | | F1 Score | Recall | Precision | MAE | RMSE |
| Audio | CNN-Augm [32] | 0.67 | 0.58 | 0.78 | - | - |
| | Williamson et al. [17] | 0.50 | - | - | 5.36 | 6.74 |
| | Ma et al. [33] | 0.52 | 1.00 | 0.35 | - | - |
| | Alhanai et al. [19] | 0.63 | 0.56 | 0.71 | 5.13 | 6.50 |
| | Yang et al. [50] | - | - | - | 4.63 | 5.52 |
| | Haque et al. [20] | - | - | - | 5.78 | - |
| | SVM | 0.40 | 0.50 | 0.33 | 5.93 | 6.74 |
| | Decision Tree | 0.57 | 0.50 | 0.57 | 5.70 | 6.79 |
| | Proposed 1D CNN model | 0.81 | 0.92 | 0.73 | 4.25 | 5.45 |
| Text | Trf-Augm [32] | 0.78 | 0.75 | 0.82 | - | - |
| | Haque et al. [20] | - | - | - | 6.57 | - |
| | Alhanai et al. [19] | 0.67 | 0.80 | 0.57 | 5.18 | 6.38 |
| | Sun et al. [29] | 0.55 | 0.89 | 0.40 | 3.87 | 4.98 |
| | Williamson et al. [17] | 0.84 | - | - | 3.34 | 4.46 |
| | SVM | 0.53 | 0.42 | 0.71 | 5.58 | 6.71 |
| | Decision Tree | 0.50 | 0.67 | 0.40 | 6.09 | 7.84 |
| | Proposed BiLSTM model | 0.83 | 0.83 | 0.83 | 3.88 | 5.44 |
| Audio & Text | Alhanai et al. [19] | 0.77 | 0.83 | 0.71 | 5.10 | 6.37 |
| | Trf+CNN-Augm [32] | 0.87 | 0.83 | 0.91 | - | - |
| | Proposed multi-modal fusion network | 0.85 | 0.92 | 0.79 | 3.75 | 5.44 |

Compared with the two proposed models that accept single modalities, the proposed multi-modal fusion method produces better results in both the classification and regression task (with an F1 score equal to 0.85 and an MAE equal to 3.75), which indicates that more information leads to better prediction performance. Compared with the other two methods, our method achieves the second best performance with merely 0.02 worse on the F1 score. However, our method has the best score for the recall metric, which is 0.92 and is much higher than other methods, for which the score of the recall metric is 0.83. This indicates that our method can find much more depressed patients than its counterparts with a comparative accuracy.

6.2. Results of Experiment 2

The results of the second experiment are compared with the baseline results in [47], which are shown in Table 4. Compared with the baseline performance on the test set, it can be seen that our method can effectively improve the assessment accuracy with MAE and RMSE values equal to 9.30 and 11.55, respectively. It should be noted that the language used in AVID-Corpus is German, while the language in DAIC-WoZ is English. This strongly demonstrates that our proposed 1D CNN model based on Mel spectrograms features can work properly even if the patients speak different languages.

Table 4. Results of experiments on the Audio-Visual Depressive Language Corpus (AViD-Corpus) dataset.

| Models | Regression | |
|---------------|-------------|--------------|
| | MAE | RMSE |
| Baseline [47] | 10.03 | 12.57 |
| 1D CNN | 9.30 | 11.55 |

6.3. Comparative Study on the Effectiveness of Different Audio/Text Features

The results of the comparison experiments using different features are shown in Table 5. Comparison experiments referring to different text features evaluate the two language models of Bert and ELMo in encoding sentence embeddings. Bert-768 and Bert-1024 in Table 5 refer to the Uncased Bert-Base and Uncased Bert-Large model, respectively. It can be seen that ELMo achieves the best results among the three sets of text features, with its F1 score equal to 0.83, recall equal to 0.83, and precision equal to 0.83. Three commonly used audio features (i.e., MFCC, COVAREP, and Mel spectrograms) were compared in the experiment. It is obvious that among all the types of audio features, Mel spectrograms present an overwhelming advantage, with an F1 score equal to 0.81, recall equal to 0.92, and precision equal to 0.73.

Therefore, in our proposed model, ELMo was selected to encode sentence embeddings, and Mel spectrograms were adopted as the input audio features.

Table 5. Performance comparison using different features.

| Types | Features | Classification | | |
|-------|------------------|----------------|-------------|-------------|
| | | F1 Score | Recall | Precision |
| Text | Bert-768 | 0.67 | 0.58 | 0.78 |
| | Bert-1024 | 0.78 | 0.75 | 0.82 |
| | ELMo | 0.83 | 0.83 | 0.83 |
| Audio | MFCC | 0.48 | 0.42 | 0.56 |
| | COVAREP | 0.76 | 0.79 | 0.72 |
| | Mel spectrograms | 0.81 | 0.92 | 0.73 |

6.4. Results of the Effectiveness Evaluation of the Proposed Data Resampling Method

The results of the effectiveness evaluation of the proposed data resampling method are listed in Table 6. It can be seen from Table 6 that for both the 1D CNN model and the BiLSTM with an attention layer model, the performances improve greatly after balancing the dataset, with the F1 score of the 1D CNN model increasing from 0.35 to 0.82 and the F1 score of the BiLSTM model increasing from 0.62 to 0.83. The greatly improved performances demonstrate the effectiveness of the proposed data resampling method.

Table 6. Performance of different features before and after data resampling.

| Models | Feature Set | Classification | | |
|--------------------------------|-------------|----------------|-------------|-------------|
| | | F1 Score | Recall | Precision |
| BiLSTM with An Attention Layer | Balanced | 0.83 | 0.83 | 0.83 |
| | Unbalanced | 0.62 | 0.67 | 0.57 |
| 1D CNN | Balanced | 0.81 | 0.92 | 0.73 |
| | Unbalanced | 0.35 | 0.25 | 0.50 |

7. Discussion

It is noteworthy that our method is the only method that has been tested on two different depression related datasets (i.e., DAIC-WoZ and AViD-Corpus). The superior performances on the two datasets reveal that our method has great generalization capability and is applicable in a more general situation.

Furthermore, in order to avoid relying on the clinicians' expertise, our method does not select depression related questions manually, but extracts features from responses to the universal questions. We hope that the proposed method would be able to learn representative characteristics of depressed patients from the input features. The experimental results on the two datasets demonstrate that our method can effectively extract representative characteristics that are depression related.

Psychologists usually diagnose depression by interviewing potential patients. However, there exists a situation in which psychologists may be misled by the verbal behaviors of the depressed patients when they dishonestly report their mental states unconsciously or intentionally, and consequently give an incorrect diagnosis. Different from the verbal behaviors, the nonverbal behaviors of depressed patients (i.e., audio characteristics) are affected by neurophysiological changes, which cannot be intentionally controlled by the depressed patients. Therefore, including nonverbal behaviors in depression detection can help improve diagnostic accuracy when potential patients attempt to mislead their clinicians. In practical cases, the psychologists can freely choose the prediction results generated by the fusion modalities or the audio modality according to their judgments on whether the patients are lying or not.

8. Conclusions

Automated depression detection is of practical significance in supporting the clinician's diagnosis and self-diagnosis. However, existing methods all have their own weaknesses such as requiring extra expertise or not being accurate enough. In this work, we propose a new depression detection method that adopts the audio signals and linguistic content of the interviews. In order to avoid relying on expertise and to increase the generalization capability, the proposed approach extracts features from universal questions instead of manually selected questions. To improve the prediction accuracy, the proposed approach combines information from the audio and text modalities to produce accurate prediction results. The proposed approach exploits the strong learning ability of the 1D CNN and BiLSTM model with an attention layer to summarize embeddings from the features of different modalities. These embeddings are concatenated and fed into two FC networks, which serve as multi-model fusion networks either to determine whether the interviewee has depression or to assess the depression severity of the patient. We further conduct several experiments to select the optimal audio/text features for depression detection and design a resampling method to get rid of the non-depressed preference of the proposed approach. Compared with other methods, our approach achieves the best performance. The superiority of our method indicates a promising way for automatic depression detection. In order to apply our proposed approach in practice, we intend to build an app that allows users to self-detect their depressive states based on the proposed method.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2076-3417/10/23/8701/s1>.

Author Contributions: All authors contributed to the design and implementation of the research, to the analysis of the results, and to the writing of the manuscript. All authors read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China under Grant 61972285 and in part by the Natural Science Foundation of Shanghai under Grant 19ZR1461300.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, Y.; Fairbairn, C.; Cohn, J.F. Detecting Depression Severity from Vocal Prosody. *IEEE Trans. Affect. Comput.* **2013**, *4*, 142–150. [CrossRef] [PubMed]
2. Depression Overview. Available online: <https://www.who.int/news-room/fact-sheets/detail/depression> (accessed on 30 June 2020).
3. Allen, N.B.; Hetrick, S.E.; Simmons, J.G.; Hickie, I.B. Early intervention for depressive disorders in young people: The opportunity and the (lack of) evidence. *Med. J. Aust.* **2007**, *187*, S15–S17. [CrossRef] [PubMed]
4. Craft, L.L.; Landers, D.M. The Effect of Exercise on Clinical Depression and Depression Resulting from Mental Illness: A Meta-Analysis. *J. Sport Exerc. Psychol.* **1998**, *20*, 339–357. [CrossRef]
5. Schumann, I.; Schneider, A.; Kantert, C.; Löwe, B.; Linde, K. Physicians' attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: A systematic review of qualitative studies. *Fam. Pract.* **2012**, *29*, 255–263. [CrossRef] [PubMed]
6. Wolpert, L. Stigma of depression: A biologist's view. *Lancet* **1998**, *352*, 1057. [CrossRef]
7. Yokoya, S.; Maeno, T.; Sakamoto, N.; Goto, R.; Maeno, T. A Brief Survey of Public Knowledge and Stigma Towards Depression. *J. Clin. Med. Res.* **2018**, *10*, 202–209. [CrossRef]
8. Corrigan, P. How Stigma Interferes With Mental Health Care. *Am. Psychol.* **2004**, *59*, 614–625. [CrossRef]
9. Sirey, J.A.; Bruce, M.L.; Alexopoulos, G.S.; Perlick, D.A.; Raue, P.; Friedman, S.J.; Meyers, B.S. Perceived Stigma as a Predictor of Treatment Discontinuation in Young and Older Outpatients with Depression. *Am. J. Psychiatry* **2001**, *158*, 479–481. [CrossRef]
10. Le, H.N.; Boyd, R.C. Prevention of major depression: Early detection and early intervention in the general population. *Clin. Neuropsychiatry J. Treat. Eval.* **2006**, *3*, 6–22.
11. Cummins, N.; Scherer, S.; Krajewski, J.; Schnieder, S.; Epps, J.; Quatieri, T.F. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **2015**, *71*, 10–49. [CrossRef]
12. Hönig, F.; Batliner, A.; Nöth, E.; Schnieder, S.; Krajewski, J. Automatic modelling of depressed speech: Relevant features and relevance of gender. In Proceedings of the INTERSPEECH 2014, Singapore, 14–18 September 2014.
13. Mundt, J.C.; Vogel, A.P.; Feltner, D.E.; Lenderking, W.R. Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. *Biol. Psychiatry* **2012**, *72*, 580–587. [CrossRef] [PubMed]
14. Trevino, A.C.; Quatieri, T.F.; Malyska, N. Phonologically-based biomarkers for major depressive disorder. *EURASIP J. Adv. Signal Process.* **2011**, *2011*, 42. [CrossRef]
15. Zinken, J.; Zinken, K.; Wilson, J.C.; Butler, L.; Skinner, T. Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression. *Psychiatry Res.* **2010**, *179*, 181–186. [CrossRef] [PubMed]
16. Oxman, T.E.; Rosenberg, S.D.; Schnurr, P.P.; Tucker, G.J. Diagnostic classification through content analysis of patients' speech. *Am. J. Psychiatry* **1988**, *145*, 464–468.
17. R, W.J.; Godoy, E.; Cha, M.; Schwarzentruher, A.; Khorrami, P.; Gwon, Y.; Kung, H.T.; Dagli, C.; Quatieri, T.F. Detecting Depression Using Vocal, Facial and Semantic Communication Cues. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 16 October 2016; pp. 11–18.
18. Yang, L.; Jiang, D.; Xia, X.; Pei, X.; Oveneke, M.C.; Sahli, H. Multimodal Measurement of Depression Using Deep Learning Models. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, 23–27 October 2017; pp. 53–59.
19. Al hanai, T.; Ghassemi, M.M.; Glass, J.R. Detecting Depression with Audio/Text Sequence Modeling of Interviews. In Proceedings of the INTERSPEECH 2018, Hyderabad, Indian, 2–6 September 2018; pp. 1716–1720.
20. Haque, A.; Guo, M.; Miner, A.S.; Li, F.F. Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions. *arXiv* **2018**, arXiv:1811.08592.
21. Cohn, J.F.; Krueez, T.S.; Matthews, I.; Yang, Y.; Nguyen, M.H.; Padilla, M.T.; Zhou, F.; De La Torre, F. Detecting depression from facial actions and vocal prosody. In Proceedings of the 2009 3rd International Conference on

- Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands, 10–12 September 2009; pp. 1–7.
22. Joshi, J.; Goecke, R.; Alghowinem, S.; Dhall, A.; Wagner, M.; Epps, J.; Parker, G.; Breakspear, M. Multimodal Assistive Technologies for Depression Diagnosis and Monitoring. *J. Multimodal User Interfaces* **2013**, *7*, 217–228. [\[CrossRef\]](#)
 23. Scherer, S.; Stratou, G.; Lucas, G.; Mahmoud, M.; Boberg, J.; Gratch, J.; Rizzo, A.; Morency, L.P. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image Vis. Comput.* **2014**, *32*, 648–658. [\[CrossRef\]](#)
 24. Morales, M.R.; Scherer, S.; Levitan, R. OpenMM: An Open-Source Multimodal Feature Extraction Tool. In Proceedings of the INTERSPEECH 2017, Stockholm, Sweden, 20–24 August 2017; pp. 3354–3358.
 25. Cummins, N.; Sethu, V.; Epps, J.; Schnieder, S.; Krajewski, J. Analysis of Acoustic Space Variability in Speech Affected by Depression. *Speech Commun.* **2015**, *75*, 27–49. [\[CrossRef\]](#)
 26. Meng, H.; Huang, D.; Wang, H.; Yang, H.; Al-Shuraifi, M.; Wang, Y. Depression Recognition Based on Dynamic Facial and Vocal Expression Features Using Partial Least Square Regression. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, Barcelona, Spain, 21–25 October 2013; pp. 21–30.
 27. Arroll, B.; Smith, F.G.; Kerse, N.; Fishman, T.; Gunn, J. Effect of the addition of a “help” question to two screening questions on specificity for diagnosis of depression in general practice: Diagnostic validity study. *BMJ* **2005**, *331*, 884. [\[CrossRef\]](#)
 28. Yang, L.; Jiang, D.; He, L.; Pei, E.; Oveneke, M.C.; Sahli, H. Decision Tree Based Depression Classification from Audio Video and Language Information. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 16 October 2016; pp. 89–96.
 29. Sun, B.; Zhang, Y.; He, J.; Yu, L.; Xu, Q.; Li, D.; Wang, Z. A Random Forest Regression Method With Selected-Text Feature For Depression Assessment. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, 23–27 October 2017; pp. 61–68.
 30. Gong, Y.; Poellabauer, C. Topic Modeling Based Multi-Modal Depression Detection. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, 23–27 October 2017; pp. 69–76.
 31. Mendels, G.; Levitan, S.I.; Lee, K.Z.; Hirschberg, J. Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection. In Proceedings of the INTERSPEECH 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1472–1476.
 32. Lam, G.; Huang, D.Y.; Lin, W.S. Context-aware Deep Learning for Multi-modal Depression Detection. In Proceedings of the Iccasp IEEE International Conference on Acoustics, Brighton, UK, 12–17 May 2019; pp. 3946–3950. [\[CrossRef\]](#)
 33. Ma, X.; Yang, H.; Chen, Q.; Huang, D.; Wang, Y. DepAudioNet: An Efficient Deep Model for Audio Based Depression Classification. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 16 October 2016; pp. 35–42.
 34. Dinkel, H.; Wu, M.; Yu, K. Text-based Depression Detection: What Triggers An Alert. *arXiv* **2019**, arXiv:1904.05154.
 35. Rana, R.; Latif, S.; Gururajan, R.; Gray, A.; Mackenzie, G.; Humphris, G.; Dunn, J. Automated screening for distress: A perspective for the future. *Eur. J. Cancer Care* **2019**, *28*, e13033. [\[CrossRef\]](#) [\[PubMed\]](#)

36. Valstar, M.; Gratch, J.; Schuller, B.; Ringeval, F.; Lalanne, D.; Torres, M.; Scherer, S.; Stratou, G.; Cowie, R.; Pantic, M. AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 16 October 2016; pp. 3–10.
37. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. *arXiv* **2018**, arXiv:1802.05365.
38. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
39. Gratch, J.; Arstein, R.; Lucas, G.M.; Stratou, G.; Scherer, S.; Nazarian, A.; Wood, R.; Boberg, J.; DeVault, D.; Marsella, S.; et al. The Distress Analysis Interview Corpus of human and computer interviews. In Proceedings of the LREC, Reykjavik, Iceland, 26–31 May 2014; pp. 3123–3128.
40. Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R.; Pantic, M. AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge. In Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, Barcelona, Spain, 21–25 October 2013; pp. 3–10.
41. Hollon, S.D.; DeRubeis, R.J.; Evans, M.D.; Wiemer, M.J.; Garvey, M.J.; Grove, W.M.; Tuason, V.B. Cognitive Therapy and Pharmacotherapy for Depression: Singly and in Combination. *Arch. Gen. Psychiatry* **1992**, *49*, 774–781. [[CrossRef](#)] [[PubMed](#)]
42. Alghowinem, S.; Goecke, R.; Wagner, M.; Epps, J.; Gedeon, T.; Breakspear, M.; Parker, G. A Comparative Study of Different Classifiers for Detecting Depression from Spontaneous Speech. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 8022–8026.
43. Aman, F.; Vacher, M.; Rossato, S.; Portet, F. Speech Recognition of Aged Voices in the AAL Context: Detection of Distress Sentences. In Proceedings of the 2013 7th Conference on Speech Technology and Human–Computer Dialogue (SpeD), Cluj-Napoca, Romania, 16–19 October 2013; pp. 1–8.
44. Scherer, S.; Stratou, G.; Gratch, J.; Morency, L.P. Investigating voice quality as a speaker-independent indicator of depression and PTSD. In Proceedings of the INTERSPEECH 2013, Lyon, France, 25–29 August 2013; pp. 847–851.
45. Gilbody, S.; Richards, D.; Brealey, S.; Hewitt, C. Screening for Depression in Medical Settings with the Patient Health Questionnaire (PHQ): A Diagnostic Meta-Analysis. *J. Gen. Intern. Med.* **2007**, *22*, 1596–1602. [[CrossRef](#)]
46. DeVault, D.; Arstein, R.; Benn, G.; Dey, T.; Fast, E.; Gainer, A. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, Paris, France, 5–9 May 2014; pp. 1061–1068.
47. Valstar, M.; Schuller, B.; Smith, K.; Almaev, T.; Eyben, F.; Krajewski, J.; Cowie, R.; Pantic, M. AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Orlando, FL, USA, 7 November 2014; pp. 3–10.
48. Rush, A.J.; Carmody, T.J.; Ibrahim, H.M.; Trivedi, M.H.; Biggs, M.M.; Shores-Wilson, K.; Crismon, M.L.; Toprac, M.G.; Kashner, T.M. Comparison of Self-Report and Clinician Ratings on Two Inventories of Depressive Symptomatology. *Psychiatr. Serv.* **2006**, *57*, 829–837. [[CrossRef](#)]
49. Depression Detect. Available online: <https://github.com/kykiefier/depression-detect> (accessed on 17 April 2020).
50. Yang, L.; Jiang, D.; Sahli, H. Feature Augmenting Networks for Improving Depression Severity Estimation from Speech Signals. *IEEE Access* **2020**, *8*, 24033–24045. [[CrossRef](#)]

Sample Availability: The source codes have been made publicly available at <https://github.com/linlemn/DepressionDetection>.

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).