A Topic-Attentive Transformer-based Model For Multimodal Depression Detection

Yanrong Guo, Chenyang Zhu, Shijie Hao, and Richang Hong, Senior Member, IEEE

Abstract—Depression is one of the most common mental disorders, which imposes heavy negative impacts on one's daily life. Diagnosing depression based on the interview is usually in the form of questions and answers. In this process, the audio signals and their text transcripts of a subject are correlated to depression cues and easily recorded. Therefore, it is feasible to build an Automatic Depression Detection (ADD) model based on the data of these modalities in practice. However, there are two major challenges that should be addressed for constructing an effective ADD model. The first challenge is the organization of the textual and audio data, which can be of various contents and lengths for different subjects. The second challenge is the lack of training samples due to the privacy concern. Targeting to these two challenges, we propose the TOpic ATtentive transformer-based ADD model, abbreviated as TOAT. To address the first challenge, in the TOAT model, topic is taken as the basic unit of the textual and audio data according to the question-answer form in a typical interviewing process. Based on that, a topic attention module is designed to learn the importance of each topic, which helps the model better retrieve the depressed samples. To solve the issue of data scarcity, we introduce large pre-trained models, and the fine-tuning strategy is adopted based on the small-scale ADD training data. We also design a two-branch architecture with a late-fusion strategy for building the TOAT model, in which the textual and audio data are encoded independently. We evaluate our model on the multimodal DAIC-WOZ dataset specifically designed for the ADD task. Experimental results show the superiority of our method. More importantly, the ablation studies demonstrate the effectiveness of the key elements in the TOAT model.

Index Terms—Multimodal depression detection, transformer, topic attention

I. INTRODUCTION

DEPRESSION is one of the most common mental disorders in nowadays society. Patients suffering from depression may have difficulty in focusing attention, keeping a good mood, or have suicidal tendency in extreme conditions [1]. According to [2] 4.4% people of the world are now suffering from depression, and the prevalence is still on the rise. The good news is that depression can be effectively alleviated or be cured with an in-time diagnosis. However, unlike other physical disease, depression symptoms are usually obscure, making the depression detection difficult. Currently, the diagnosis of depression relies commonly on subjective rating scales, such as the Beck Depression Inventory (BDI) and the Eight-item Patient Health Questionnaire depression scale (PHQ-8) [3]. The accurate diagnosis based on these kinds of information usually depends on the participation of experienced psychologists, which can be labor-intensive for many occasions. Thus, the urgent demand for a practical and accurate method for detecting depression gives rise to a new task based on machine learning techniques: automatic depression detection (ADD) [4]. ADD systems typically work through recording and analyzing data from interviewees, such as speech or facial expression, which is more objective and requires fewer expert interventions.

Prior ADD works [5-7] have shown that information from one modality is limited, while different modalities such as audio, text, and video are complementary and have the potential of improving the ADD performance. Therefore, multimodal depression detection has been attracting more and more attention in recent years. As the cornerstone, the DAIC-WOZ dataset [8] is designed and built as a multimodal depression dataset, which becomes the most popular platform for conducting the ADD research. During the data collection, an AI is adopted to ask participants fundamental questions and collects audio, video, and depth sensor recordings during the interviews. Among the modalities, we can see that speech and its text transcripts are easy to obtain in daily life. Therefore, it is more practical to build an ADD system that is based on the audio and textual modality data. For example, the data can be simply collected through phone calls. In this context, we plan to construct an effective ADD model by fully exploring the data from these two modalities.

To achieve this goal, we have two major challenges to address. The first one is the reasonable organization of the speech data. For different subjects, their interviewing contents may be diverse. For example, the question list can be different among subjects, and the answers of a same question can be varied from one to another in terms of the contents and the duration time. In this context, it is important to organize the raw data into a suitable form that facilitates the full exploration of ADD-specific features. The second challenge is the data scarcity due to the privacy concern. Despite of the prevalence of depression, it is not easy to collect a large-scale dataset for ADD research, as most people are less willing to authorize the use of their data even if the research ethics can be guaranteed. This issue poses difficulty to the learning-based ADD task, which usually requires sufficient training data for fitting the complex mapping between the multimodal data and the semantic label.

In our research, we propose the TOpic ATtentive transformer-based ADD model to address the above challenges. First, we organize the audio and textual data by taking...
topic as the basic unit. In a typical data collection scenario, a subject answers pre-defined questions one by one during the interview. Therefore, the organization of the original data is to take each question as a single topic, which is the basic unit for building the ADD model. The consideration behind this idea are two-fold. On the one hand, the topic is at the mesoscopic level that stays between the sentence level and the subject level. A topic contains semantic information that is more complete than a single sentence. In the meanwhile, the information from multiple topics of a subject can be more comprehensive for evaluating his/her mental state. On the other hand, more importantly, we can introduce the topic attention mechanism, which enables our model to estimate the importance of each topic for accurately inferring the subject’s mental state.

Second, we relieve the data scarcity issue by introducing and fine-tuning pre-trained models. Recently, transformer-based models based on self-supervised learning (SSL) have achieved great success in natural language processing, computer vision, and automatic speech recognition. These SSL models are pre-trained on a large scale of unlabelled datasets for learning feature representations, and they can be effectively applied on downstream tasks. Several pioneering works have demonstrated the superiority of fine-tuning these SSL models on downstream tasks, which sheds lights on solving the data scarcity issue in our application.

Based on the above considerations, we design the proposed TOAT model as a two-branch structure, in which two pre-trained models are introduced. As for the textual modality, we introduce and fine-tune the RoBERTa model to encode the texts of each topic into the depression-related features. As for the audio modality, we introduce and fine-tune wav2vec 2.0 to encode the raw audio signals into the features that can be used for the ADD task. The complementarity of the two modalities is utilized by gathering the two branches in a simple fusion module, which also helps to relieve the data insufficiency issue to some extent.

The technical contributions of this research can be summarized as follows:

- We propose a novel architecture for multimodal depression detection. The introduction of pre-trained models, as well as the fusion mechanism, addresses the data scarcity issue in the ADD task.
- We design the topic attention module to learn the importance of the questions for a participant during the interview. From the experiments, we empirically find that the attention mechanism boosts the system by enhancing its ability of retrieving more positive samples from the participants.

The rest of the paper is organized as follows. Section II introduces the related works. Section III introduce the overall framework of our TOAT model and its details. In Section IV, we provide the experimental results and their analysis. Section V finally concludes the paper.

II. RELATED WORK

Multimodal depression detection infers one’s mental state by jointly exploring data of multiple modalities, such as visual, acoustic, or textual modalities. In this section, we discuss the multimodal fusion and topic-based methods in the depression detection task.

Fusion-based modeling is a common technical roadmap, in which intelligible and appropriate features are firstly extracted from different modalities, and the fusion of these features is then conducted at different stages of the method. Valstar et al. propose a baseline fusion method at the inference level, in which a logic AND operation for the audio-based inference and the video-based inference is made. Yang et al. propose a gender-specific decision tree that utilizes three modalities, of which the statistical features cooperate with each other along the built trees. Qureshi et al. propose the gender-aware model, in which the multi-modality features are concatenated, and a fusing vector is learned to assign different weights for them. Ray et al. propose a novel framework that builds the two-level fusion mechanism. The first-level fusion is conducted within each modality, and the second-level fusion is conducted among the different modalities. As pre-trained models have become more and more effective to extract features, Makiuchi et al. introduce BERT to extract text-modality features. For the audio modality, deep spectrum features are extracted by using a pre-trained VGG-16 network, and a gated CNN followed by a LSTM layer. In the fusion part, these two kinds of features are concatenated directly. From the above works, we can see that the late-fusion strategy is mainly adopted in the multimodal depression detection task, in which the main focus is either on the effective feature extraction, or the accurate fusion-weight learning. Despite the subtle modeling, a main challenge for these methods lies in the small dataset, which tends to bring in the over-fitting issue, especially for the models with large solution space. It is noted that the well pre-trained models have the potential of solving this issue, in which the fine-tuning strategy can be applied based on the small-scale dataset for depression detection.

From another perspective, the multimodal depression detection models are built upon the collected data at different scales. Specifically, scale means the observation range of the multi-modality data. For example, Rohanian et al. perform fusion after each word in an utterance, in which fusion gates are adopted to control the contributions from audio and video modalities. The observation range of this model is at the word level, which is very local. In contrast, many above-mentioned methods directly extract and fuse features at the global level, i.e. the whole input data is taken as the input of the feature extractor. Apart from the methods based on the microscopic scale or the macroscopic scale, modeling the multi-modality data at the mesoscopic scale is also useful for the depression detection task. As the interview for depression detection typically has a question-answer form, the structured data facilitates the construction of a topic-based model. It is noted that it is different from the traditional topic model aiming at discovering topics from documents. Instead, the
topic-based models here treat a question as a topic directly. Gong et al. [29] propose a topic-modeling-based method to perform context-aware analysis. They concatenate visual, textual and audio features based on topics to build input data. The research in [30] also considers the topic as the basic unit during the modeling.

Our model extracts the features of different modalities at the topic level as well, and simply fuses them before sending them into a linear classifier. As for the topic-based modeling, our model is different from [29] and [30], as we construct a topic attention model to enable the model to learn the topic importance. Of note, in [30], the whole dataset containing multiple topics is re-organized by splitting it into topic-wise subsets, while our model is still built upon the whole dataset. Therefore, the difference of the data organization makes the two methods fundamentally different in terms of the model construction and performance evaluation.

III. METHOD

A. Overview

Figure 1 shows the framework of our TOAT model. The model input includes the textual data and the corresponding audio data of $N$ topics, which is prepared for the two branches in our model. One branch aims to explore the depression-related semantic information from the textual modality data. The other branch aims to explore the acoustic information that is correlated with depression from the audio modality data. Then, the learned features from the two branches are directly combined and sent into a linear classifier. In this way, the

depression detection is formulated as a binary classification problem (being positive or negative).

B. Learning Textual Modality

As for the textual branch, we detail the process ahead of the topic attention. It has been proved that fine-tuning a BERT-like model to obtain textual features can be effective for the depression detection task [30]. Besides, considering the data deficiency, a complex model can be easily overfit if trained from scratch. Therefore, we fine-tune a pre-trained improved version of BERT, i.e. RoBERTa, to extract textual features.

Firstly, for every participant, we build a group that consists of $N$ frequently-asked questions and their replies. We represent each topic as $S_i (i = 1, ..., N)$, which contains a topic index $i$, the corresponding question and reply $S_i = \{q_i, r_i\}$. Therefore, the textual data of $N$ topics is represented as $X_t = \{S_1, S_2, ..., S_N\}$.

To extract semantic information, we introduce RoBERTa [19] as the backbone, which is an upgraded BERT via training on larger datasets containing longer sentences with larger batches. Besides, it removes the next sentence prediction objective in BERT [10]. RoBERTa has two versions, i.e. RoBERTa-large and RoBERTa-base, of which the architecture contains 24 transformer blocks and 12 transformer blocks, respectively. In our application, we simply choose the RoBERTa-base version for the balance between performance and time complexity.

To extract semantic information, we introduce RoBERTa [19] as the backbone, which is an upgraded BERT via training on larger datasets containing longer sentences with larger batches. Besides, it removes the next sentence prediction objective in BERT [10]. RoBERTa has two versions, i.e. RoBERTa-large and RoBERTa-base, of which the architecture contains 24 transformer blocks and 12 transformer blocks, respectively. In our application, we simply choose the RoBERTa-base version for the balance between performance and time complexity.

\{S_i\} are sent into RoBERTa one by one, in which the tokenizer encodes them as

\[
\text{tokenizer}(S_i) = \{[CLS], s_{i,1}, s_{i,2}, ..., s_{i,L}, [SEP]\}
\]

(1)
where $s$ is a byte-level token. CLS and SEP tokens are inserted at the beginning and the end of $S_t$. The tokens are further sent into RoBERTa-base’s main body, which contains 12 transformer layers and has the hidden dimension of $D = 768$. As we are doing a classification task, the CLS token is taken as the output of RoBERTa, which is denoted as $h_{ti} \in \mathbb{R}^{1 \times D}$. For all the $N$ topics, we concatenate their CLS heads $H_t = [h_{t1}; h_{t2}; \cdots; h_{tN}]$.

Of note, it is common that not all the questions are asked for each participant in our applications. To solve this practical issue, the missing ones are marked as None. For those None parts, we directly fill the corresponding $h_{ti}$ with numbers approaching $-\infty$ (-1e9 in our implementation). To an extent, it can be regarded as a mask operation.

### C. Topic Attention Module

![Diagram of topic attention module]

It is plausible to assume that the topics are of varying significance for the depression detection task. For example, it seems that the question like “How easy is it for you to get a good night’s sleep?” is more helpful than “What’d you study at school?”, as reply of the former question possibly contains important information about the mental state of the participant.

To approach this issue, the missing ones are marked as None. For those None parts, we directly fill the corresponding $h_{ti}$ with numbers approaching $-\infty$. Based on this consideration, we introduce the topic attention mechanism, and design the topic attention module to learn the importance of all the adopted topics. As shown in Figure 2, we feed the learned representation $H_t \in \mathbb{R}^{D \times N}$ into a linear layer to reduce its dimension into $1 \times N$. Aiming to help the linear layer adjust the topic score, an element-wise product is imposed by a learnable parameter $w \in \mathbb{R}^{1 \times N}$. This process can be expressed as:

$$g = w \odot \text{Linear}(H_t)$$  \hspace{1cm} (2)

Then, we apply a softmax layer $g_i^* = \frac{\exp(g_i)}{\sum_{j=1}^{N} \exp(g_j)}$ to normalize the topic scores. Of note, the masked vectors do not have influence on this process as $\exp(-\infty) = 0$. Based on the topic score estimation, the model is able to learn the importance of each topic with respect to the depression detection. We set a threshold $\alpha = 1/N$ to penalize these normalized topic scores, and obtain the refined scores $\hat{g}_i$ as:

$$\hat{g}_i = \begin{cases} 0 & \text{if } g_i^* < \alpha \\ g_i^* & \text{if } g_i^* >= \alpha \end{cases}$$  \hspace{1cm} (3)

This non-linear threshold operation intentionally highlights the important topics while ignoring the less important ones. Finally, we multiply scores $\hat{g}_i$ and matrix $H_t$ followed by a sum operation to obtain the weighted representations $\hat{h}_t$:

$$\hat{h}_t = \sum_{i=1}^{N} \hat{g}_i H_{ti}$$  \hspace{1cm} (4)

### D. Learning Audio Modality

In our application, we aim to extract depression-related features from audio data, such as discriminating cues in pronunciation and intonation, to assist the textual information. To this end, we introduce the pre-trained Wav2vec 2.0 model as the backbone of our second branch, which is trained on the Librispeech dataset [31] and achieves the SOTA performance on several audio-related tasks. Different from some previous works [32, 33] using low-level acoustic features as inputs, we directly use the raw speech signals of the replies from the participants as the input of the pre-trained model.

The architecture of Wav2vec 2.0 consists of a feature encoder composed of several CNNs, and a transformer module that is similar to RoBERTa. The difference between them is that Wav2vec 2.0 replaces the fixed position embedding with a convolutional layer that works as relative positional embedding. As there is no CLS head for Wav2vec 2.0, we use a projector layer instead, which reduces the dimension with a linear layer, and calculates the average along the timeline:

$$\hat{H}_a = \text{Linear}(H_a)$$  \hspace{1cm} (5)

$$\hat{h}_a = \frac{1}{L} \sum_{i=0}^{L} \hat{H}_{a,i}$$

Here $H_a \in \mathbb{R}^{L \times D}$ is the output of the last transformer layer, $L$ is the timestep. In our implement, $D_a$ is set as 256. $\hat{h}_a$ is the obtained feature from the audio branch. The abbreviation of audio.

In the implementation, we ignore the quantization module in a typical Wav2vec 2.0 model, which is not applicable in our application. In addition, considering the computational efficiency, we extract the features by choosing only a part of interviewees’ replies, and truncate them into a fixed size.

### E. Multi-modal Fusion

In the fusion stage, to demonstrate the usefulness of the feature extraction stage, we directly concatenate the two features $\hat{X} = \text{concat}(\hat{h}_t, \hat{h}_a)$, followed by layer normalization [34] and a dropout layer, aiming to mitigate the gap between the two kinds of features. Finally, a linear classification layer is adopted to obtain the prediction score $y$. 
F. Optimization Objectives

To measure the difference between the ground-truth and prediction, we choose the cross entropy loss as the optimization objective for the classification task. The loss function is shown as follows:

\[
L = \frac{1}{N} \sum_{i=1}^{N} -y^{(i)} \log y^{(i)} - (1 - y^{(i)}) \log (1 - y^{(i)})
\]  

(6)

where \( y \) is the prediction and \( \hat{y} \) is the ground-truth label.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

(7)

IV. EXPERIMENTS

A. Dataset and Implementation Details

We evaluate our method on a public dataset for multimodal depression detection named Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) [8]. In the interviews, an artificial intelligence named Ellie is used to ask participants some task-related questions and collect the recordings from the participants in the meanwhile.

The DAIC-WOZ dataset includes three modalities: video, audio, and text. The visual features extracted from the recordings are preprocessed with the OpenFace toolkit [35] for the sake of privacy protection. For the audio modality, besides the raw speech audio, the dataset also provides low-level features that are preprocessed by COVAREP toolbox [36]. For the textual modality, the text transcripts are provided. In our work, we use the raw speech audio and the text transcripts. The DAIC-WOZ dataset collects the interviews from 189 participants. Since a few samples (their IDs are 451, 458, and 480) have no question information, we ignore these samples in all our experiments. For the left 186 samples, the division in the experiments for training/validation/testing is 107/33/46. In addition, as the DAIC-WOZ dataset is somewhat imbalanced, we apply the strategy of random-oversampling [29, 30] for the depressed subjects during the training process, aiming to alleviate the imbalance issue.

We choose ten mostly-asked questions (Table II) in our research (\( N = 10 \)). Each of them represents a topic. We implement TOAT with Pytorch [37] and HuggingFace library [38]. We use the AdamW optimizer [39] and set the learning rate as 4e-6. We freeze the parameters of the feature encoder as the DAIC-WOZ dataset is somewhat imbalanced, as the RoBERTa is able to process two sentences at most. From Table III, we can see the performance without the TA module has an obvious drop shows for both two versions of our model. For example, the F1-score drops from 64.7 (TOAT, \( \alpha = 0.1 \)) to...
TABLE I
Comparison with several methods based on the DAIC-WOZ validation and test set. ‘A’, ‘V’, ‘T’ denotes the audio, visual and textual modality, respectively. Best results are highlighted.

| Model          | Modality | Validation Set | Test Set |
|----------------|----------|----------------|----------|
|                | Accuracy | Recall | Precision | F1-score | Accuracy | Recall | Precision | F1-score |
| Decision Tree  | AVT      | 85.7   | 85.7       | 85.7     | -        | 66.7   | 50.0      | 57.1     |
| SVM            | AVT      | -      | 42.8       | 60.0     | 50.0     | -      | 77.8      | 46.7     | 58.3     |
| Topic Model    | AVT      | -      | -          | 70.0     | -        | -      | -         | 60.0     |
| Random Forest  | T        | -      | 67         | 57       | 62       | -      | 89.0      | 40.0     | 55.0     |
| TDCN           | V        | 85.7   | 83.3       | 77.0     | 80.0     | 66.0   | 64.3      | 45.0     | 53.0     |
| TOAT           | AT       | 78.8   | 83.3       | 66.7     | 74.1     | 73.9   | 78.6      | 55.0     | 64.7     |

TABLE II
The list of ten frequently asked questions and their sample numbers.

| Index | Question Description | Samples |
|-------|----------------------|---------|
| Q1    | How are you doing today? | 184     |
| Q2    | How are you at controlling your temper? | 176     |
| Q3    | What’d you study at school? | 162     |
| Q4    | Is there anything you regret? | 152     |
| Q5    | Have you been diagnosed with depression? | 167     |
| Q6    | When was the last time you argued with someone and what was it about? | 182     |
| Q7    | What advice would you give to yourself ten or twenty years ago? | 177     |
| Q8    | What are you most proud of in your life? | 171     |
| Q9    | When was the last time you felt really happy? | 179     |
| Q10   | How easy is it for you to get a good night’s sleep? | 173     |

TABLE III
Results of ablation studies on the DAIC-WOZ test set. $\alpha$ is the threshold of the topic attention. When it’s set as 0, it means there is no threshold. ‘-’ represents there is no topic attention.

| Multimodal | Description | $\alpha$ | Accuracy | Recall | Precision | F1-score |
|------------|-------------|----------|----------|--------|-----------|----------|
| TOAT       | 0.1         | 73.9     | 78.6     | 55.0   | 64.7      |
|            | 0.2         | 69.5     | 57.1     | 50.0   | 53.3      |
|            | -           | 69.6     | 57.1     | 50.0   | 53.3      |
| Unimodal   | Text        | 0.1      | 73.9     | 50.0   | 53.9      | 51.9     |
|            | 0           | 69.6     | 50.0     | 50.0   | 50.0      |
|            | -           | 67.4     | 42.9     | 46.2   | 44.4      |
|            | Audio       | -        | 71.7     | 42.9   | 54.6      | 48.0     |

53.3 (w/o TA). As for the unimodal-based version, the F1-score drops from 51.9 (text, $\alpha = 0.1$) to 44.4 (text w/o TA) by removing the TA module. The reasons are two-fold. First, as the tokenizer in [19] receives a sentence whose length is 512 at most, we have to drop some information when we aggregate all questions and replies together. Second, the language model has difficulty in learning information from multiple Question+Answer pairs when we put them in one sentence, which are commonly separated by a $SEP$ token.

The threshold $\alpha$ also has large impact on the performance. In our experiments, we achieve the best results by empirically setting $\alpha$ as $1/N = 0.1$. We observe that the overall performance is very sensitive to the threshold setting. On the one hand, if we set $\alpha$ as 0, or keep originally learned weights for each topic in another word, the overall performance has an obvious drop in both two versions, which again shows the effectiveness of the topic attention mechanism. Surprisingly, when we increase the threshold $\alpha$ into 0.2, the TOAT model totally fails to detect the depressed samples from the test set. In this threshold setting, we find that all the learned topic weights $G_i$ are suppressed as zero, as no $G_i$ is larger than $\alpha = 0.2$. In this context, the so-called textual feature $H_t$ collapses into a zero vector, and therefore disables the whole model.

2) Unimodal Setting: In the ablation study, we also construct two incomplete versions of our model by removing the one of the branches. From the results in the Unimodal part of Table III, the audio branch itself achieves the F1-score of 48.0, while the textual branch itself obtains the F1-score of 51.9. The performance gap between the two incomplete versions and the full TOAT model ($\alpha = 0.1$) demonstrate the necessity of combing multiple modalities. As the textual modality contains semantic information and the audio modality contains pronunciation and intonation information, they are highly complementary to each other, therefore forming the a more comprehensive representation.

3) Topic Analysis: We design another experiment to study the individual usefulness of each topic. We conduct the following experimental setting. First, we reorganize the whole dataset into $N$ topic-wise subsets. For the $i$-th topic, we build the $i$-th subset by selecting the samples containing this topic, in which only the textual and audio data of this topic are retained. Then, we re-train the multiple versions of our model and test their performance based on these subsets. It is noted that we do not
include the topic attention module in this experiment, as the data in these subsets only have one single topic. The F1-scores of the ten topics are shown in Figure 3. In addition, we obtain the topic-related statistics from the experiment conducted on the whole DAIC-WOZ dataset. For each topic during the test, we accumulate its chosen time if the topic score of a sample is larger than the threshold $\alpha$. By dividing the chosen times by the number of samples (Table II), we obtain the usage rates of all the topics, as shown in Figure 3. From the figure, despite the F1-scores of these single-topic models stay at a low-level in general, they vary a lot across the different topics. In addition, we can see that the F1-scores of the topics are generally positively correlated with their usage rates, except the 5th topic. The results from the two experiments empirically present the different importance existed among the topics. In another word, if the performance of a certain topic is relatively better, it is more likely to be selected by the topic attention model. The exceptional case occurs in the 5th topic, i.e. ‘Have you been diagnosed with depression?’. In this case, the trained model solely based on this topic is highly sensitive to its answer (yes or no). On the other hand, the results shown in Figure 3 demonstrate the necessity of utilizing multiple topics during the depression detection, as the result based on the whole dataset (F1-score of 64.7) is better than those of the ten models trained with their corresponding subsets.

As is shown in Table IV, we also provide two typical examples in the DAIC-WOZ test set. In the table, the replies of the topics chosen by the topic attention module ($G^i_\alpha \geq \alpha$) are highlighted in bold. From the answering details in Table IV, we can see that the first participant is in a relatively good mood. He seldom feels about something, and feels easy to sleep. The second participant talks a lot, so we omit some words due to the limited space. We can see that this participant is not in a good mood. For example, she has difficulty in falling asleep, as she tends to think depressing things when she tries to sleep. She seldom feels happy, feeling miserable most time instead. As for these two cases, our model assigns the large weights to the replies that are highly related with the mental state inference, which is also generally in accordance with the subjective evaluation.

V. CONCLUSION

In this paper, we propose the TOAT model for the multi-modal depression detection task. To alleviate the overfitting issue caused by small-dataset training, we construct the two-branch framework, in which two models pre-trained on large corpus are utilized, and fine-tunings are made based on the depression detection dataset. We also build the topic attention model for the textual modality branch, which fully explores the importance of each topic, and effectively enhances the detection accuracy. Experimental results on the DAIC-WOZ dataset validates the effectiveness of our model. In the future, we plan to build a more comprehensive depression detection framework by introducing the other modalities such as visual cues.

Acknowledgements This work was supported by National Key Research and Development Program (Grant No. 2019YFA0706200), the National Nature Science Foundation of China under Grant No. 62072152, 62172137, and the Fundamental Research Funds for the Central Universities under Grant No. PA2020GDKC0023.

REFERENCES

[1] F. Hao, G. Pang, Y. Wu, Z. Pi, L. Xia, and G. Min, “Providing appropriate social support to prevention of
depression for highly anxious sufferers,” *IEEE Transactions on Computational Social Systems*, pp. 1–9, 2019.

[2] M. Friedrich, “Depression Is the Leading Cause of Disability Around the World,” *JAMA*, vol. 317, no. 15, p. 1517, 2017. [Online]. Available: [http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.3826](http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.3826)

[3] K. Kroenke, T. W. Strine, J. B. Williams, J. T. Berry, and A. H. Mokdad, “The phq-8 as a measure of current depression in the general population,” *Journal of Affective Disorders*, vol. 114, no. 1, pp. 163–173, 2009.

[4] P. Wang, F. Barrett, E. Martin, M. Milonova, R. E. Gur, R. C. Gur, C. Kohler, and R. Verma, “Automated video-based facial expression analysis of neuropsychiatric disorders,” *Journal of Neuroscience Methods*, vol. 168, no. 1, pp. 224–238, 2008.

[5] M. Niu, K. Chen, Q. Chen, and L. Yang, “Hcag: A hierarchical context-aware graph attention model for depression detection,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[6] T. A. Hanai, M. M. Ghassemi, and J. R. Glass, “Detecting depression with audio/text sequence modeling of interviews,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 1716–1720.

[7] M. Rohanian, J. Hough, and M. Purver, “Detecting depression with word-level multimodal fusion,” in *Inter-speech 2019*, 2019.

[8] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberk, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, “The distress analysis interview corpus of human and computer interviews,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, pp. 3123–3128. [Online]. Available: [http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf)

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008. [Online]. Available: [https://proceedings.neurips.cc/paper/2017/hash/3f5ec243547dee91fbd053c1e4a845a-Abstract.html](https://proceedings.neurips.cc/paper/2017/hash/3f5ec243547dee91fbd053c1e4a845a-Abstract.html)

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: [https://aclanthology.org/N19-1423](https://aclanthology.org/N19-1423)

[11] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 5754–5764. [Online]. Available: [https://proceedings.neurips.cc/paper/2019/hash/1c6a7e655d7e5840e66733e9e67cc69-Abstract.html](https://proceedings.neurips.cc/paper/2019/hash/1c6a7e655d7e5840e66733e9e67cc69-Abstract.html)

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” 2020.

[13] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, “ibot: Image bert pre-training with online tokenzier,” 2021.

[14] H. Bao, L. Dong, and F. Wei, “Beit: Bert pre-training of image transformers,” 2021.

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” 2021.

[16] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: [https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd69f9ba3227870bb6f707-Abstract.html](https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd69f9ba3227870bb6f707-Abstract.html)

[17] A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec: Self-supervised learning of discrete speech representations,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia*, April 26-30, 2020. OpenReview.net, 2020. [Online]. Available: [https://openreview.net/forum?id=rlywJxrYDS](https://openreview.net/forum?id=rlywJxrYDS)

[18] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in *Inter-speech 2019*, 2019, pp. 3465–3469.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019.

[20] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, D. DeVault, M. T. Torres, S. Scherer, R. Cowie, and M. Pantic, “Avec 2016: Depression, mood, and emotion recognition workshop and challenge,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3–10.

[21] L. Yang, D. Jiang, L. He, E. Pei, M. C. Ovoneke, and H. Sahli, “Decision tree based depression classification from audio video and language information,” in *Proceedings of the 6th International Workshop on
[22] S. A. Oureshi, G. Dias, S. Saha, and M. Hasanuzzaman, “Gender-aware estimation of depression severity level in a multimodal setting,” in 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021, pp. 1–8.

[23] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, “Multi-level attention network using text, audio and video for depression prediction,” in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 81–88.

[24] M. R. Makiuchi, T. Warnita, K. Uto, and K. Shinoda, “Multimodal fusion of bert-cnn and gated cnn representations for depression detection,” in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 55–63.

[25] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[26] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] M. Rohanian, J. Hough, and M. Purver, “Detecting depression with word-level multimodal fusion,” in Interspeech 2019, 2019, pp. 1443–1447.

[28] L. Hong and B. D. Davison, “Empirical study of topic modeling in twitter,” in Proceedings of the First Workshop on Social Media Analytics, 2010, pp. 80–88.

[29] Y. Gong and C. Poellabauer, “Topic modeling based multi-modal depression detection,” in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017, pp. 69–76.

[30] E. Toto, M. Tlachac, and E. A. Rundensteiner, “Audibert: A deep transfer learning multimodal classification framework for depression screening,” in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 4145–4154.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015. IEEE, 2015, pp. 5206–5210. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178964

[32] Z. Dai, H. Zhou, Q. Ba, Y. Zhou, L. Wang, and G. Li, “Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis,” Journal of Affective Disorders, vol. 295, pp. 1040–1048, 2021.

[33] M. Niu, K. Chen, Q. Chen, and L. Yang, “Hcag: A hierarchical context-aware graph attention model for depression detection,” in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 4235–4239.

[34] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” arXiv preprint arXiv:1607.06450, 2016. [Online]. Available: https://arxiv.org/abs/1607.06450

[35] T. Baltrusaitis, P. Robinson, and L.-P. Morency, “Openface: An open source facial behavior analysis toolkit,” in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1–10.

[36] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, “Covarep — a collaborative voice analysis repository for speech technologies,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014), 2014, pp. 960–964.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alch´e-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/bdbca2881ee7f921f4a9701272740-Abstract.html

[38] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Huggingface’s transformers: State-of-the-art natural language processing,” 2020.

[39] I. Loschchivol and F. Hutter, “Decoupled weight decay regularization,” in 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg66RuCqY7

[40] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, “A random forest regression method with selected-text feature for depression assessment,” in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017, pp. 61–68.

[41] Y. Guo, C. Zhu, S. Hao, and R. Hong, “Automatic depression detection via learning and fusing features from visual cues,” 2022.
Yanrong Guo Yanrong Guo is an associate professor at School of Computer and Information, Hefei University of Technology (HFUT). She is also with Key Laboratory of Knowledge Engineering with Big Data (Hefei University of technology), Ministry of Education. She received her Ph.D. degree at HFUT in 2013. She was a postdoc researcher at University of North Carolina at Chapel Hill (UNC) from 2013 to 2016. Her research interests include pattern recognition and medical image analysis.

Chenyang Zhu Chenyang Zhu received his B.E. from Hefei University of Technology, Hefei, in 2020. Now he is a master student at School of Computer Science and Information Engineering, Hefei University of Technology (HFUT). His current research interest is pattern recognition.

Shijie Hao Shijie Hao is an associate professor at School of Computer Science and Information Engineering, Hefei University of Technology (HFUT). He is also with Key Laboratory of Knowledge Engineering with Big Data (Hefei University of technology), Ministry of Education. He received his Ph.D. degree at HFUT in 2012. His research interests include image processing and pattern recognition.

Richang Hong Richang Hong received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008. He was a Research Fellow of the School of Computing with the National University of Singapore, from 2008 to 2010. He is currently a Professor with the Hefei University of Technology, Hefei. He is also with Key Laboratory of Knowledge Engineering with Big Data (Hefei University of technology), Ministry of Education. He has coauthored over 70 publications in the areas of his research interests, which include multimedia content analysis and social media. He is a member of the ACM and the Executive Committee Member of the ACM SIGMM China Chapter. He was a recipient of the Best Paper Award from the ACM Multimedia 2010, the Best Paper Award from the ACM ICMR 2015, and the Honorable Mention of the IEEE Transactions on Multimedia Best Paper Award. He has served as the Technical Program Chair of the MMM 2016. He has served as an Associate Editor of IEEE Multimedia Magazine, Neural Processing Letter (Springer) Information Sciences (Elsevier) and Signal Processing (Elsevier).