INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews

Adria Mallol-Ragolta[1], *Ziping Zhao*[1,2], *Lukas Stappen*[1], *Nicholas Cummins*[1], *Björn Schuller*[1,3]

[1] ZD.B Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[2] College of Computer and Information Engineering, Tianjin Normal University, China
[3] GLAM – Group on Language, Audio & Music, Imperial College London, UK

adria.mallol-ragolta@informatik.uni-augsburg.de

Abstract

The high prevalence of depression in society has given rise to a need for new digital tools that can aid its early detection. Among other effects, depression impacts the use of language. Seeking to exploit this, this work focuses on the detection of depressed and non-depressed individuals through the analysis of linguistic information extracted from transcripts of clinical interviews with a virtual agent. Specifically, we investigated the advantages of employing hierarchical attention-based networks for this task. Using Global Vectors (GloVe) pretrained word embedding models to extract low-level representations of the words, we compared hierarchical local-global attention networks and hierarchical contextual attention networks. We performed our experiments on the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WoZ) dataset, which contains audio, visual, and linguistic information acquired from participants during a clinical session. Our results using the DAIC-WoZ test set indicate that hierarchical contextual attention networks are the most suitable configuration to detect depression from transcripts. The configuration achieves an Unweighted Average Recall (UAR) of .66 using the test set, surpassing our baseline, a Recurrent Neural Network that does not use attention.

Index Terms: natural language processing, depression detection, hierarchical networks, attention mechanisms.

1. Introduction

Mental disorders are present in the society and they represent a major public health concern, as they affect approximately 25 % of the population in the European Region [1]. In this work, we focus on the automatic detection of *major depression disorder* (MDD) because of its prevalence in society; more than 300 million people globally in 2015 were affected by the disorder [2]. The diagnosis of MDD in primary health care settings is a well-known challenging task [3], which highlights the need for objective diagnostic aids.

Psychological studies have observed differences in the use of language between depressed and non-depressed individuals [4, 5, 6]. The present work, therefore, aims to automatically identify depressed and non-depressed participants through the analysis of linguistic information extracted from transcribed clinical interviews. To identify the most relevant linguistic information, our approach implements a hierarchical attention-based network that works at both word- and sentence-level. In machine learning, attention is a technique that can be used to weight the importance of certain input information with respect to its context. An advantage of attention, motivating this work, is that it allows the automatic identification of words in a sentence, and sentences in the interviews, relevant for depression detection.

Recently, attention mechanisms have been employed in a broad range of applications, such as acoustic scene classification [7], speaker diarisation [8], speech emotion recognition [9], image classification [10], video classification [11], and video description [12]. Attention mechanisms with linguistic information have also been used in document classification [13] and sentiment and self-assessed emotion detection [14] problems. Nevertheless, the use of attention in the depression detection problem has not been studied deeply, although depression detection has been a problem widely investigated among the research community. Previous works investigated the use of different machine learning techniques with audio, video, and linguistic data, employed in isolation [15, 16, 17, 6] or multimodally [18, 19, 20]. The techniques included decision trees [21], support vector machine [22], support vector regression [23], and convolutional neural networks combined with recurrent neural networks [24].

This work presents a novel approach to tackle the depression detection problem using hierarchical attention-based networks and transcribed clinical interviews. We employed transcriptions in the *Distress Analysis Interview Corpus - Wizard of Oz* (DAIC-WoZ) database [25, 22]. We processed the transcriptions and used *Global Vectors* (GloVe) pretrained word embedding models [26] to map words onto real-valued feature vectors. The vectors are then fed into a hierarchical attention-based network to extract high-level representations from the interviews. We compared three different hierarchical networks: (i) a naïve hierarchical network, (ii) a hierarchical local-global attention network [27], and (iii) a hierarchical contextual attention network [13]. The choice of (ii) and (iii) was motivated by their successful performance in document classification tasks [13, 27]. In our experiments, we defined the performance of the naïve hierarchical network as a baseline to assess the benefits of using attention mechanisms in hierarchical networks. Finally, we used a fully-connected layer to perform the classification.

The rest of the paper is laid out as follows: Section 2 presents the dataset employed, and Section 3 describes the methodology followed in this work. Section 4 details the experiments performed and analyses the results obtained, while Section 5 concludes the paper and suggests some future work directions.

2. DAIC-WoZ Dataset

The DAIC-WoZ depression database employed in this work was used in AVEC 2017 [28] (cf. Table 1) and is a subset of the *Distress Analysis Interview Corpus* (DAIC) [25]. It consists of clinical interviews conducted to aid the diagnosis of psychological distress disorders. The data was collected for the development of a virtual therapist capable of detecting verbal and nonverbal signs of an individual's mental illness [29]. The DAIC-WoZ database contains the Wizard-of-Oz interviews, which were conducted

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2036

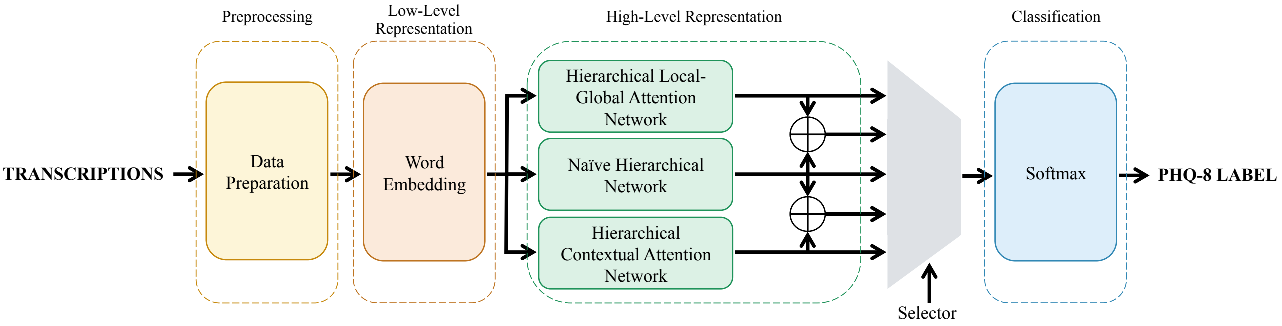
Preprocessing
Low-Level Representation
High-Level Representation
Classification
TRANSCRIPTIONS
Data Preparation
Word Embedding
Hierarchical Local-Global Attention Network
Naïve Hierarchical Network
Hierarchical Contextual Attention Network
Selector
Softmax
PHQ-8 LABEL

Figure 1: *Pipeline of the implemented system, which receives transcriptions from interviews and identifies whether they belong to depressed or non-depressed interviewees. Following preprocessing, low- and high-level representations of the interviews are extracted from the transcripts. The learnt high-level representations are then fed into a classifier.*

Table 1: *DAIC-WoZ dataset summary with the total number of participants gathered, in addition to the number of participants categorized as depressed.*

	Train Set	**Devel. Set**	**Test Set**	$\sum$
Total participants	107	35	47	189
Depressed participants	30	12	9	51

by Ellie, an animated virtual interviewer controlled by a human interviewer in another location. Audio and video recordings and questionnaire responses were collected, and transcriptions were made of the interviews. Depression severity was assessed in the interviewed individuals using the eight-item *Patient Health Questionnaire* (PHQ-8) depression scale, a valid diagnostic technique to measure the severity of depression-related disorders, which can be employed to discriminate depressed (PHQ-8 score $\geq$ 10) from non-depressed (PHQ-8 score $<$ 10) individuals [30]. Although the studied task was presented in AVEC 2016 [22], we used the updated version of the dataset, which was released in AVEC 2017 [28] edition. In both cases, train, development, and test data partitions were unchanged.

3. Methodology

In this section, we present our system (cf. Figure 1), which was designed to predict whether transcriptions from clinical interviews belong to a depressed or non-depressed participant through the analysis of linguistic information. The different stages of the system are described below.

3.1. Preprocessing

The objective of this stage is to prepare the transcriptions for further processing. To focus our investigation on the interventions by the study participants, we used the supplied timings to delete all interventions by Ellie, the virtual agent. Analysing the raw transcriptions, we observed that, in some cases, they included additional, non-verbal communication, such as when the participants sighed or breathed deeply. These were also removed from the transcriptions, as these are not easily accessible from automatic systems. The last stage of preprocessing is the *tokenisation* of the transcriptions by splitting interviewees' interventions, obtained as a string of characters, into meaningful semantic units, i. e., space-separated sequences of characters. Once defined, these tokens are mapped onto numerical identifiers for further processing. To model the flow of linguistic information over the course of the interviews, we converted the data in the train, development and test sets into three-dimensional tensors with dimensions (I, S, W). The first dimension, (I)ndividual, contains information from each participant, while the second dimension, (S)entence, contains sentence-level information, and the third dimension, (W)ord, contains word-level information. Thus, in each position, the identifier of the corresponding token associated with a particular word in a particular sentence coming from a specific participant's interview is stored. Since the number of sentences per interview and their length is participant-dependent, we identified the longest interview and the longest sentence over the entire set and used it to generate the aforementioned tensors. Shorter interviews and sentences are zero-padded.

3.2. Low-Level Representation

Following preprocessing, low-level representations are generated by extracting features from the transcriptions. The process of feature extraction in natural language processing might not be as intuitive as in the fields of speech or image processing, as there is a need to map words onto real-valued vectors. In this regard, we employed pretrained word embedding models, specifically GloVe pretrained models [26], to extract vector representations of the tokens contained in our data. These pretrained word embedding models were chosen for three reasons. Firstly, they are trained in the interview language. Secondly, much larger corpora (with 6, 27, 42, and 840 billion tokens, respectively) than DAIC-WoZ are used to train the word embedding models, making them more stable. Thirdly, these models output feature vectors with 25, 50, 100, 200, and 300 dimensions, which contributes to the analysis of performance differences. The use of GloVe pretrained word embedding models allows every token in our data to be mapped to a real-valued vector representation. To perform this mapping, a matrix with vector representations corresponding to all tokens in our data is built. We assigned pseudorandom feature vector representations to those tokens that are not present in the pretrained word embedding models. Once built, we z-normalised the embedding matrix. Finally, we represented each token in our three-dimensional tensor with their corresponding feature vector.

3.3. High-Level Representation

Once the low-level representations are extracted, we then computed the high-level features using a two-staged hierarchical network which operates at the word- and sentence-level [14]. Herein, we use $s_i \in i = 1, \ldots, S$ to represent the i-th sentence in the transcription, $w_{ij} \in j = 1, \ldots, W$ to denote the j-th word in the i-th sentence, and $\boldsymbol{x}_{ij}$ to indicate the feature vector representation of the j-th word in the i-th sentence. First, at the word-level, feature vectors $\boldsymbol{x}_{ij} \in j = 1, \ldots, W$ are fed into a bidirectional *Gated Recurrent Unit* (GRU) with 25 units (50 units in total) to learn a representation of the sequence of words:

$$\boldsymbol{h}_{ij} = \left[\overrightarrow{GRU}\left(\boldsymbol{x}_{ij}\right), \overleftarrow{GRU}\left(\boldsymbol{x}_{ij}\right) \right]. \tag{1}$$

This network configuration was chosen, as it gave the strongest results in our initial experiments. The resulting word sequences representations $\boldsymbol{h}_{ij}$ are then fused in order to generate a single representation of all words in the sentence i-th, denoted as $\tilde{\boldsymbol{h}}_i$. This fusion is performed using one of three different strategies: a naïve approach described in Equation (3), a local attention approach defined in Equations (5) to (7), and a contextual attention approach detailed in Equations (12) to (14). The sentence-level representations, $\tilde{\boldsymbol{h}}_i \in i = 1, \ldots, S$, are then fed into their own bidirectional GRU, with the same configuration used previously, to extract a representation of the sequence of sentences, i. e.,

$$\boldsymbol{h}_{i} = \left[\overrightarrow{GRU}\left(\tilde{\boldsymbol{h}}_{i}\right), \overleftarrow{GRU}\left(\tilde{\boldsymbol{h}}_{i}\right) \right]. \tag{2}$$

Similar to the word-level representations, the sequential sentences representations are fused using either a naïve approach as described in Equation (4), a global attention approach as defined in Equations (8) to (11), or a contextual attention approach as detailed in Equations (15) to (17). This generates a single high-level representation $\tilde{\boldsymbol{h}}$ of all sentences in the whole interview. This high-level representation $\tilde{\boldsymbol{h}}$ is suitable for classification.

3.3.1. Hierarchical Naïve Network

Naïve fusion at word- and sentence-level is performed by averaging the representations $\boldsymbol{h}_{ij}$ over all words, and $\boldsymbol{h}_i$ over all sentences, respectively:

$$\tilde{\boldsymbol{h}}_i = \frac{\sum_{j=1}^{W} \boldsymbol{h}_{ij}}{W}, \tag{3}$$

and

$$\tilde{\boldsymbol{h}} = \frac{\sum_{i=1}^{S} \boldsymbol{h}_{i}}{S}. \tag{4}$$

As this is our simplest fusion strategy, i. e., no learnable parameters, we employed it as a baseline.

3.3.2. Hierarchical Local-Global Attention Network

Based on the approach presented in [27], we implemented a word-level information fusion approach based on a local attention mechanism:

$$\boldsymbol{u}_{ij} = \tanh\left(\mathbf{W}_w \boldsymbol{h}_{ij} + \mathbf{b}_w\right), \tag{5}$$

$$\boldsymbol{\alpha}_{ij} = \frac{\exp\left(\boldsymbol{u}_{ij}\right)}{\sum_{j=1}^{W} \exp\left(\boldsymbol{u}_{ij}\right)}, \tag{6}$$

$$\tilde{\boldsymbol{h}}_i = \sum_{j=1}^{W} \boldsymbol{\alpha}_{ij} \boldsymbol{h}_{ij}, \tag{7}$$

and a sentence-level information fusion approach based on a global attention mechanism:

$$\boldsymbol{p}_i = \text{softmax}\left(\mathbf{W}_{s_1} \boldsymbol{h}_i + \mathbf{b}_{s_1}\right), \tag{8}$$

$$\boldsymbol{o}_i = \tanh\left(\mathbf{W}_{s_2} \boldsymbol{p}_i + \mathbf{b}_{s_2}\right), \tag{9}$$

$$\boldsymbol{\alpha}_i = \frac{\exp\left(\boldsymbol{o}_i^T \mathbf{o}_s\right)}{\sum_{i=1}^{S} \exp\left(\boldsymbol{o}_i^T \mathbf{o}_s\right)}, \tag{10}$$

$$\tilde{\boldsymbol{h}} = \sum_{i=1}^{S} \boldsymbol{\alpha}_i \boldsymbol{p}_i. \tag{11}$$

In this approach, $\mathbf{W}_w$, $\mathbf{W}_{s_1}$, $\mathbf{W}_{s_2}$, $\mathbf{b}_w$, $\mathbf{b}_{s_1}$, $\mathbf{b}_{s_2}$, and $\mathbf{o}_s$ are the parameters to be learnt by the network.

3.3.3. Hierarchical Contextual Attention Network

Based on the methodology presented in [13], we implemented a word- and sentence-level fusion approach based on a contextual attention mechanism. At the word-level, the fusion can be defined as follows:

$$\boldsymbol{u}_{ij} = \tanh\left(\mathbf{W}_w \boldsymbol{h}_{ij} + \mathbf{b}_w\right), \tag{12}$$

$$\boldsymbol{\alpha}_{ij} = \frac{\exp\left(\boldsymbol{u}_{ij}^T \mathbf{u}_w\right)}{\sum_{j=1}^{W} \exp\left(\boldsymbol{u}_{ij}^T \mathbf{u}_w\right)}, \tag{13}$$

$$\tilde{\boldsymbol{h}}_i = \sum_{j=1}^{W} \boldsymbol{\alpha}_{ij} \boldsymbol{h}_{ij}, \tag{14}$$

while at the sentence-level, the information fusion can be defined as:

$$\boldsymbol{u}_{i} = \tanh\left(\mathbf{W}_s \boldsymbol{h}_{i} + \mathbf{b}_s\right), \tag{15}$$

$$\boldsymbol{\alpha}_{i} = \frac{\exp\left(\boldsymbol{u}_{i}^T \mathbf{u}_s\right)}{\sum_{i=1}^{S} \exp\left(\boldsymbol{u}_{i}^T \mathbf{u}_s\right)}, \tag{16}$$

$$\tilde{\boldsymbol{h}} = \sum_{i=1}^{S} \boldsymbol{\alpha}_{i} \boldsymbol{h}_{i}. \tag{17}$$

In this approach, $\mathbf{W}_w$, $\mathbf{W}_s$, $\mathbf{b}_w$, $\mathbf{b}_s$, $\mathbf{u}_w$, and $\mathbf{u}_s$ are the parameters to be learnt by the network. Furthermore, $\mathbf{u}_w$ and $\mathbf{u}_s$ can be interpreted as context vectors, which contribute to the identification of relevant words and sentences.

3.4. Classification

The last stage of our pipeline classifies the high-level representations extracted from the interviews. To this end, we employed a fully connected layer with two output neurons, which use a *softmax* function as activation. The first neuron models the depressed class, while the second neuron models the non-depressed class. The label indicating whether a participant is depressed is determined by the neuron with the highest output.

4. Experimental Results

The purpose of this work is to analyse the impact of hierarchical attention-based networks on the identification of depressed and non-depressed participants from linguistic information. To study the feasibility of this approach, we considered the task as a classification problem. We compared five hierarchical networks: (i) a *Naïve Hierarchical Network* (NHN), (ii) a *Hierarchical Local-Global Attention Network* (HLGAN), (iii) a *Hierarchical Contextual Attention Network* (HCAN), (iv) a *Naïve Hierarchical Network* concatenated with a *Hierarchical Local-Global Attention Network* (NHN+HLGAN), and (v) a *Naïve Hierarchical Network* concatenated with a *Hierarchical Contextual Attention Network* (NHN+HCAN). Network configurations (iv) and (v) are built by concatenating the high-level representations learnt from participants' interviews with the merged configurations. We also compared the performance of all the available GloVe word embedding models.

Table 2: *UAR measurements computed using the development set of the DAIC-WoZ dataset for different GloVe word embedding models, αB.βd, and hierarchical network configurations; α indicates the number of billion tokens used to train the model, while β indicates the dimension of the output word embedding vector. Networks with an UAR of .50 are underfitted, so we highlighted those above.*

Embedding Models	**NHN**	**HLGAN**	**HCAN**	**NHN + HLGAN**	**NHN + HCAN**
6B.50d	.50	.50	**.75**	.50	**.73**
6B.100d	.50	.50	.46	.50	**.52**
6B.200d	.50	.50	.47	.50	**.71**
6B.300d	.50	**.58**	**.71**	.46	**.56**
27B.25d	.50	.50	.48	.50	.46
27B.50d	.50	.50	**.79**	.50	**.67**
27B.100d	.50	.48	.41	.50	**.52**
27B.200d	.50	.50	**.58**	.50	**.62**
42B.300d	.50	**.54**	**.62**	**.52**	.43
840B.300d	.50	**.60**	**.54**	**.52**	**.54**

The networks, implemented in Keras with TensorFlow at the backend, use *Categorical Cross-Entropy* as the loss function to minimise, and *Adam* as the optimiser. We set the parameter corresponding to the batch size for training the networks to 8. As the data is unbalanced (cf. Table 1), we assessed the performance of the trained models by computing the *Unweighted Average Recall* (UAR) between the true and predicted labels associated to each participant. In initial experiments, we observed the behaviour of loss and accuracy from both train and development sets over 75 epochs computed from all hierarchical networks and word embedding models. Analysis of the generated curves, which are not reported here, revealed that the number of epochs used to train the networks is an important parameter to prevent trained models to suffer from either underfitting or overfitting. Despite behaviour differences among the evaluated scenarios, we fixed the maximum number of epochs for training to 30 to provide a fair comparison between network configurations and word embedding models.

Analysing the results computed using the development set (cf. Table 2), we observe that 23 out of 50 models achieve an UAR of .50. This result likely indicates that, in these cases, the models are underfitted, as they predict the same label to all samples. Thus, to exclude underfitted models, we focus on model performances with an UAR greater than .50. No clear patterns in model performance are observed when different word embeddings are employed. However, in the case of GloVe '*840B.300d*' word embedding, all network configurations surpass the UAR of .50, with the exception of the NHN configuration (cf. Table 2). Hence, the GloVe '*840B.300d*' word embedding model was selected to assess the performance of the implemented hierarchical networks using the test set. At this point, the development phase is over and we move to the testing phase, in which models are trained with data from both train and development sets, and tested with data reserved exclusively for testing. Additionally, the performances of the trained models in the testing phase are assessed by computing the F1 score between the true and predicted labels, enabling comparison with other works in the literature.

The results obtained using the test set (cf. Table 3) allow us to state that the NHN models are underfitted. Furthermore, the measured UAR using the test set for HLGAN and NHN+HLGAN configurations are .33 and .45, versus .60 and .52, respectively, using the development set. The deterioration in performance upon testing casts doubt on the benefits of using HLGAN for depression detection with linguistic information. On the other hand, configurations based on HCAN improve their performances using the test set, with the HCAN configuration achieving the best UAR of .66. For the NHN+HCAN configuration, there is a subtle decline from an UAR of .54 using the development set to .52 using the test set. Thus, we conclude that HCAN is the most suitable configuration to perform depression detection using linguistic data, and that its combination with NHN provides low benefits in terms of system performance.

Table 3: *UAR and F1 score measurements computed using the development and test sets of the DAIC-WoZ dataset for different hierarchical network configurations with the GloVe '840B.300d' word embedding model. Those configurations that obtained an UAR greater than .50 using the test set are highlighted. The best performances are obtained from network configurations using hierarchical contextual attention.*

		NHN	**HLGAN**	**HCAN**	**NHN + HLGAN**	**NHN + HCAN**
UAR	Devel. Set	.50	.60	.54	.52	.54
	Test Set	.50	.33	**.66**	.45	**.52**
F1 score	Devel. Set	.40	.60	.51	.46	.53
	Test Set	.45	.35	.63	.42	.51

5. Conclusions

In this work, we developed and analysed the performance of hierarchical attention-based networks to identify depressed and non-depressed participants from linguistic information, which was acquired from interviews with a virtual agent. Our results on the DAIC-WoZ test set with the GloVe '*840B.300d*' pretrained word embedding model indicate that a hierarchical contextual attention network is the most suitable configuration for the task, achieving an UAR of .66. Furthermore, the performance of this configuration surpasses the baseline of a hierarchical network without attention. Thus, this result supports the benefits of employing attention mechanisms in the problem of depression detection. Nonetheless, our analyses might be biased, as the baseline models are underfitted, which does not permit a proper evaluation of the baseline architecture. As future work, we aim to continue this investigation using regression techniques. In addition, we plan to explore the benefits of using attention mechanisms in both audio and video modalities, so we can investigate the use of multimodal attention-based approaches on depression detection and recognition systems.

6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE), and it has been supported by the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300). Funding has also been received from BMW Group Research. Further funding has been received from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 115902, which receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. We thank Dr. Judith Dineley for her proofreading work.

7. References

[1] World Health Organization, “The European Mental Health Action Plan 2013–2020,” 2015. [Online]. Available: https://bit.ly/2UvIQi6

[2] ——, “Depression and Other Common Mental Disorders: Global Health Estimates,” 2017, licence: CC BY-NC-SA 3.0 IGO. [Online]. Available: https://bit.ly/2YPy1qi

[3] H. Lester and A. Howe, “Depression in primary care: three key challenges,” *Postgraduate Medical Journal*, vol. 84, no. 996, pp. 545–548, 2008.

[4] S. Rude, E.-M. Gortner, and J. Pennebaker, “Language use of depressed and depression-vulnerable college students,” *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[5] J. D. Bernard, J. L. Baddeley, B. F. Rodriguez, and P. A. Burke, “Depression, Language, and Affect: An Examination of the Influence of Baseline Depression and Affect Induction on Language,” *Journal of Language and Social Psychology*, vol. 35, no. 3, pp. 317–326, 2016.

[6] E.-M. Rathner, J. Djamali, Y. Terhorst, B. Schuller, N. Cummins, G. Salamon, C. Hunger-Schoppe, and H. Baumeister, “How Did You like 2017? Detection of Language Markers of Depression and Narcissism in Personal Narratives,” in *Proc. of Interspeech*, Hyderabad, India, 2018, pp. 3388–3392.

[7] T. Zhang, K. Zhang, and J. Wu, “Multi-modal Attention Mechanisms in LSTM and Its Application to Acoustic Scene Classification,” in *Proc. of Interspeech*, Hyderabad, India, 2018, pp. 3328–3332.

[8] H. Song, M. Willi, J. J. Thiagarajan, V. Berisha, and A. Spanias, “Triplet Network with Attention for Speaker Diarization,” in *Proc. of Interspeech*, Hyderabad, India, 2018, pp. 3608–3612.

[9] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, “An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition,” in *Proc. of Interspeech*, Hyderabad, India, 2018, pp. 3087–3091.

[10] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual Attention Network for Image Classification,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, 2017, pp. 3156–3164.

[11] X. Long, C. Gan, G. de Melo, X. Liu, Y. Li, F. Li, and S. Wen, “Multimodal Keyless Attention Fusion for Video Classification,” in *Proc. of the AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, 2018, pp. 7202 – 7209.

[12] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, “Attention-Based Multimodal Fusion for Video Description,” in *Proc. of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 4193 – 4202.

[13] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical Attention Networks for Document Classification,” in *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, 2016, pp. 1480–1489.

[14] L. Stappen, N. Cummins, E.-M. Meßner, H. Baumeister, J. Dineley, and B. Schuller, “Context Modelling Using Hierarchical Attention Networks for Sentiment and Self-assessed Emotion Detection in Spoken Narratives,” in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, 2019, pp. 6680–6684.

[15] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Communication*, vol. 71, pp. 10 – 49, 2015.

[16] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller, “Enhancing Speech-Based Depression Detection Through Gender Dependent Vowel-Level Formant Features,” in *Proc. of Conference on Artificial Intelligence in Medicine*, A. ten Teije, C. Popow, J. H. Holmes, and L. Sacchi, Eds., Vienna, Austria, 2017, pp. 209–214.

[17] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, “Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics,” *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 578–584, 2018.

[18] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, “Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features,” in *Proc. of the International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 43–50.

[19] Z. Huang, B. Stasak, T. Dang, K. Wataraka Gamage, P. Le, V. Sethu, and J. Epps, “Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016,” in *Proc. of the International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 19–26.

[20] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias, F. Yang, and M. Tsiknakis, “Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text,” in *Proc. of the International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 27–34.

[21] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, “Decision Tree Based Depression Classification from Audio Video and Language Information,” in *Proc. of the International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 89–96.

[22] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, “AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge,” in *Proc. of the International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 3–10.

[23] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, “Detecting Depression Using Vocal, Facial and Semantic Communication Cues,” in *Proc. of the International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 11–18.

[24] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, “DepAudioNet: An Efficient Deep Model for Audio Based Depression Classification,” in *Proc. of the International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 35–42.

[25] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, “The Distress Analysis Interview Corpus of human and computer interviews,” in *Proc. of the International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014, pp. 3123 – 3128.

[26] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1532–1543.

[27] X. Niu and Y. Hou, “Hierarchical Attention BLSTM for Modeling Sentences and Documents,” in *Proc. of the International Conference on Neural Information Processing*, D. Liu, S. Xie, Y. Li, D. Zhao, and E.-S. M. El-Alfy, Eds., vol. 2, Guangzhou, China, 2017, pp. 167–177.

[28] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, “AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge,” in *Proc. of the Annual Workshop on Audio/Visual Emotion Challenge*, Mountain View, California, USA, 2017, pp. 3–9.

[29] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, “SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support,” in *Proc. of the International Conference on Autonomous Agents and Multi-agent Systems*, Paris, France, 2014, pp. 1061–1068.

[30] K. Kroenke, T. W. Strine, R. L. Spitzer, 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 – 3, pp. 163 – 173, 2009.