Depression Severity Estimation from Multiple Modalities

Evgeny A. Stepanov*, Stéphane Lathuilière†*, Shammur Absar Chowdhury*,
Arindam Ghosh*, Radu-Laurenţiu Vieriu*, Nicu Sebe*, Giuseppe Riccardi *
*University of Trento, Italy
name.surname@unitn.it
†INRIA Grenoble, France

***Abstract*—Depression is a major debilitating disorder which can affect people from all ages. With a continuous increase in the number of annual cases of depression, there is a need to develop automatic techniques for the detection of the presence and its severity. We explore different modalities (speech, behavioral characteristics, language and visual features extracted from face) to design and develop automatic methods for the detection of depression. In psychology literature, the eight-item Patient Health Questionnaire depression scale (PHQ-8) is well established as a tool for measuring the severity of depression. In this paper we aim to automatically predict the total sum of PHQ-8 scores from features extracted from the different modalities. We demonstrate that among the considered modalities, behavioral characteristic features extracted from speech yield the lowest MAE, outperforming the best system at the Audio/Visual Emotion Challenge (AVEC) 2017 depression sub-challenge.**

***Index Terms*—Affective Computing; Depression Detection; Machine Learning; Speech; Natural Language Processing; Facial Expressions**

I. Introduction

According to the World Health Organization (WHO), depression is a major mental disorder with about 300 million people of all ages affected worldwide. According to the Global Burden of Disease Study [1], depression is the second leading cause of disability worldwide and is on the rise. If left untreated, it can lead to complications such as reductions in the volume of the hippocampus [2]. There is a growing evidence that depression can cause impairment of the immune function by affecting different immunological pathways such as the central nervous system (CNS), the endocrine system, and the cardiovascular system. This can lead to the development or aggravation of co-morbidities and worsen health conditions in other diseases [3].

Current diagnosis of depression is dependent on an evaluation by a psychiatrist, supported by standard questionnaires to screen for depression. The Personal Health Questionnaire Depression Scale (PHQ-8) Scoring and the Hamilton Depression Rating Scale are two well established tools for the diagnosis of depression. However, the stigma around the disease and lack of understanding often prevents patients from seeking early psychiatric help. Depression often comes with side effects such as social anxiety, and decreased social contact, making it often go unnoticed by friends and family for a long time.

The growing burden of this disease suggests that there is a need to develop technologies, which can aid in automatic detection and effective care of patients suffering from depression. The recent development in the field of affective computing, focuses on the sensing, detection, and interpretation of affective states of people from their interactions with computers or machines. Affective computing methodologies use various modalities ranging from overt signals such as speech, language and video to covert signals such as heart rate, skin temperature, and galvanic skin response to understand the mental and affective states of humans. Such techniques can be used for the automatic detection of psychological states and mental health, including conditions such as Post-Traumatic Stress Disorder (PTSD) and depression.

The motivation of this study is to explore different sources of information, such as audio, video, language and behavioral cues, to predict the severity of depression. While doing so, we also investigate different feature representations and modeling techniques corresponding to each modality to improve the automatic prediction.

The paper is organized as follows. In Section II, we present relevant works in the literature for the detection of depression and affective disorders from speech, language, and facial expressions. This is followed by a brief description of the multi-modal data used for the study in Section III. An overview of the features and experimental methodology used in this study are given in Section IV. Section V provides concluding remarks.

II. State of the Art - Speech, language and facial expressions

Speech, language and facial expressions are three of the major overt signals which have been widely used for interpreting human psychological states. Automatic analysis of speech has been used for emotion recognition [4], [5], stress detection [6], [7], and mood state characterisation [8], [9]. Natural language and speech processing from diaries and recordings have been used to detect the onset of dementia, alzheimer's, and aphasia [10], [11]. Analysis of facial expressions have shown to be highly effective in tracking the progressive degeneration of cognitive health in patients suffering from schizophrenia and bipolar disorder [12].

A. Speech and Language

Several psychological conditions clearly manifest themselves through changes in speech patterns and language use. Computational and automatic screening methods have the power to detect micro-changes in speech and language patterns which would otherwise have gone unnoticed.

978-1-5386-4294-8/18/$31.00 ©2018 IEEE

Properties such as speech rate, pause duration and usage of fillers can be indicative of cognitive decline in individuals. Changes in prosody and fluency can also be useful in detecting mental health changes of depressive patients.

The utility of speech and language for the diagnosis of mental health is well established [13], [14]. The speech features such as tempo, prosodic pauses, absence of glottal rasping are associated with patients suffering from affective disorders. Formant and power spectral density (PSD) based features, on the other hand, have been demonstrated to have the highest discriminative power for classification in both genders [15]. Speech disturbances and resistivity have been found to have a positive correlation with anxiety, whereas silent pauses with depression [16]

Natural Language Processing (NLP) techniques such as topic modeling and sentiment analysis have been applied to detection of depression and prediction of psychiatric disorders [17], [18].

B. Face Analysis

Facial expressions can be an extremely powerful medium used to convey human overt emotional feedback. In recent times, there has been significant progress in developing methods for facial feature tracking for the analysis of facial expressions and the detection of emotions. Studies have shown that it is possible to effectively detect the presence of pain shown on faces.

Machine learning techniques have been shown to be effective for the automatic detection of pain and mental state from facial expressions [19], [20]. One of the most popular technique used for capturing the subtlety and fine-grained variations in facial expression is the Facial Action Coding System (FACS) developed by [21]. The FACS is based on the consensus of the judgment of human experts who observe pre-recorded facial expressions and perform manual annotation of FACS codes for each frame. These annotations, which are called action units (AUs), can belong to one of 44 different classes. FACS has been widely used in the field of psychology for measuring emotions, affect, and behavior [22], [23], [24]. More recently [25], FACS has been shown to be correlated with depression severity. Specifically, [25] found that severely depressed subjects are more likely to show fewer affiliative facial action units (AU12 and AU15) and more non-affiliative ones (AU14).

Head pose and eye gaze have also been shown to encode information about depression. For instance, [25] observes that an increase in the severity of depression comes with a diminished head motion. Other works such as [26], [27], [28] have also investigated the link between head pose, eye gaze and depression, providing evidence that such a link exists and it is all worth considering.

III. AVEC Audio Video Database

The 2017 Audio/Video Emotion Challenge and Workshop (AVEC 2017) "Real-life depression" provides a corpus comprising of audio and video recordings and transcribed speech from the Distress Analysis Interview Corpus (DAIC) [29].

TABLE I: Distribution of the AVEC data set into training and development sets for depressed (**D**) and non-depressed (**ND**) classes, and overall (**ALL**).

	ND		**D**		**ALL**
Training	77	(72%)	30	(28%)	107
Development	23	(66%)	13	(34%)	35

The dataset comprises of recordings from 189 sessions of human agent interaction, where each subject was interviewed by a virtual psychologist (see Table I for the distribution of labels in the training and development sets). The audio files, transcripts and continuous facial features of the human subject are provided as part of the challenge. The Personal Health Questionnaire Depression Scale (PHQ-8) score of the subjects is also provided in the dataset. The PHQ-8 [30] is a set of 8 short multiple choice questions which has been established as a diagnostic tool for the measurement of the severity of depressive disorders. Automatic estimation of the total sum of PHQ-8 scores from different modalities, such as speech and video, can aid in the early detection of depression and monitoring of depressive states. In the AVEC challenge, the goal is to look at different streams of data recorded during a session with the subject to predict the total sum of PHQ-8 scores, and to classify the subject as depressed or not.

IV. Experiments

In this section we describe the experiments conducted for the feature extraction and regression experiments conducted on the speech, behavioral, language and visual (facial) modalities.

A. Speech and Behavioral Characteristic Features

1) Acoustic Features: To understand the predictive characteristics of low-level acoustic feature groups to assess the depression severity of the participant, we extracted low-level descriptors (LLDs). These features are extracted using openSMILE [31], from the participant's turns in each conversation, motivated by the studies in [32], [33], [34], [35]. To extract the acoustic features we used approximately 100 overlapping frames per second and with 25 milliseconds window. The low-level features are extracted as three groups including:

- Spectral features (**S**) such as energy in spectral bands (0-250Hz, 0-650Hz, 250-650Hz, 1-4kHz), centroid, flux, max and min-position and roll-off points (25%, 50%, 70%, 90%).
- Prosodic features (**P**) such as pitch (Fundamental frequency f0, f0-envelope), voice-probability, loudness .
- Voice Quality features (**VQ**) such as jitter, shimmer, logarithmic harmonics-to-noise ratio (logHNR).

Additionally, we extract delta and acceleration coefficients of the above features and then projected onto statistical functionals used in [36] .

2) Behavioral Characteristics Features: Apart from extracting low-level features from raw speech signals, we also explore the transcriptions. We crafted features that can capture information regarding the participant's non-vocal behavior (NB) along with their turn-taking behaviors (TB) and participants' Previous Diagnosed Information (PDI) features. The non-vocal behavior ($|NB| = 3$) includes:

- frequency of laughter in participant's turns;
- percentage of disfluencies in the participant's turns, which might indicate hesitations;
- counts of cues that might suggest inconvenience like whistling, mumbling, whispering or taking deep breaths among others.

The features that are used to describe the turn-taking behaviors, ($|TB| = 6$) are the first and third quartiles and the median duration of response time (in seconds) of the participants. Similarly, we also extract statistics for the with-in speaker silence (pause). The response time represents how long the participants took to respond to the previous turn of the agent.

The PDI feature set ($|PDI| = 3$) contains numerical representations of the response of the participants to queries such as having any Post-traumatic Stress Disorder (PTSD), ptsd, depression dep, even having any military backgrounds mb. Each individual feature is encoded into three values (-1,0,1) where -1 represents the query is not present in the session, 0 presents a dis-confirmation (e.g ptsd=0 means the participant responded as "no" to the previous turn query) and 1 presents confirmation of the query.

For the regression task, we study the performance of acoustic and behavioral characteristic features. For modeling individual acoustic feature groups and their linear combination we use support vector machine (SVM) for regression, implemented in weka [37]. SVMs use Radial Basis Function (RBF) kernel with $\gamma = 0.01$ and $C = 1.0$.

As for the linear combination of different acoustic feature groups, we first concatenate all the acoustic feature vectors – including prosody (P), spectral (S) and voice quality (VQ), linearly to obtain vector M.

For predictor using behavioral characteristic feature group, we use Reduced Error Pruning Tree ("REPT") implemented in weka [37], which is a fast regression tree learner that uses information of variance reduction and prunes the tree using reduced error pruning.

The results are presented in Table II for individual feature sets and their combination. The results indicate that spectral features are a good predictor of the total PHQ score, compared to all other acoustic features. Using behavioral characteristic features, however, we obtain a decrease in both MAE and RMSE by a magnitude of 0.63 and 1.20 respectively compared to all the baselines.

TABLE II: Root mean square error (RMSE) and mean absolute error (MAE) for depression severity regression using *acoustic*, *behavioral*, *lexical*, and *visual* features on the development set. We also provide the audio and audio-video feature-based baselines (*BL: Audio* and *BL: Audio-Video*) using Random Forests.

Features	RMSE	MAE
BL: mean	6.57	5.50
BL: Audio	6.74	5.36
BL: Audio-Video	6.62	5.52
Acoustic Features		
Spectral	**6.32**	**4.96**
Voice Quality	7.05	5.70
Prosody	7.10	5.75
Merged	6.43	5.40
Behavioral Characteristics Features		
Behavioral characteristic	***5.54***	***4.73***
Language Features		
Lexical: BOOL	**6.31**	**5.17**
Lexical: TFIDF	6.78	5.40
Lexical: WE	6.84	5.41
Visual Features		
Visual	6.09	4.66

B. Language

Additional to the speech-based features, we explore text-based representations to predict depression severity estimates. The widely used representation of a document in NLP is bag-of-words, where a document is represented by word occurrences ignoring the order in which they appear. We experiment both with binary (BOOL) and tf-idf (TFIDF) weighted representations. While the binary representation encodes words that are present in the document regardless of their frequency, tf-idf weighted representation considers both the frequency of the term (*tf*) in a document and the inverse document frequency (*idf*) – which lowers the weight of the very frequent terms in a collection and increases the weight of the rare terms with respect to the equations 1-2.

$$tf - idf(t,d) = tf(t,d) * idf(t) \quad (1)$$

$$idf(t) = log\frac{n_d}{df(d,t)} + 1 \quad (2)$$

Where $tf(t,d)$ is the term frequency, n_d is the total number of documents, and $df(d,t)$ is the frequency of documents containing the term.

Besides bag-of-words representation, we also experiment with the word embedding representation (WE) [38], where pre-trained per-word embedding vectors are averaged for a document. We make use of the SKIPGRAM embedding vectors pre-trained on GoogleNews with a embedding dimension 300 and window 10.

Since the provided speech transcripts are of human-machine conversations, we first extract human turns and convert them into bag-of-words representation. The transcripts contain annotations for the speech phenomena such as laughter, sigh, etc., which were treated as any other token. Thus, the representation implicitly encodes the presence of these phenomena in the conversation; and also its frequency, in the case of tf-idf based representations. For the word embedding representation, however, this is not the case, as there are no pre-trained vectors for these.

The algorithm of our choice for text-based representations is Support Vector Regression (SVR) with linear kernel, implemented in scikit-learn [39]. The regression results for each of the document representations are given in Table II in terms of RMSE and MAE. As it can be observed, the only representation that outperforms all the baselines is the binary bag-of-word representation that yields RMSE=6.31 and MAE=5.17.

C. Visual Features

Inspired by [40] and the success reported in [41], we use the 68 3D facial keypoints and compute geometric features as follows: for every facial representation, we first remove the 3D bias (equal to a translation in the Euclidean space by subtracting the mean value in 3D), then we normalize the resulting representation so that the average distance to the center (origin) is equal to 1. Finally, we compute Euclidean distances between all possible pairs of 3D normalized points and add them to the normalized representation. This results in a feature vector of size $2,482$. Consequently, we reduce this dimension by applying Principal Component Analysis (PCA) and keeping over 99.5% of variance, resulting in a feature vector of size 33.

Since we are dealing with video sequences, we propose to regress depression using models naturally designed for temporal data. Specifically, we propose the use of Long Short-Term Memory (LSTM) neural networks [42] for this task. LSTMs have emerged as an effective and scalable model for several learning problems related to sequential data, such as handwriting recognition [43], [44], generation of handwritten characters [45], language modeling and translation [46], [47], audio [48] and video [49] signal analysis, acoustic speech modeling [50] and others. They are successful at capturing long-term temporal dependencies while being robust against the optimization problems faced by simple recurrent neural networks (RNNs).

In order to build our training set, we apply a sliding window approach to the video sequences, using windows of size W, overlapped by O samples. We use the *success* flag provided by the dataset creators which models the tracking confidence for each frame. We adopt a 0-tolerance strategy and discard all windows for which at least one failed tracking is present. We do this to exclude the risk of introducing artifacts into the feature space, that the model might misleadingly exploit for solving the task. We set the values for W and O empirically to 60 and 30, respectively. We down-sample the data to 1 second, which makes our windows 1 minute long, with an overlap of 30 seconds. During testing, we apply the same windowing scheme and average the window-level predictions over the length of the test sequence.

Next, we train a double layered LSTM model on regressing depression at window level on the training set. The model is composed of two stacked layers of size 16, followed by a $Dense$ layer with a $linear$ activation function. We use dropout [51] equal to 0.5 to control over-fitting and batch normalization [52] to limit internal covariance shift. As loss function, we use the mean squared error. In order to validate our LSTM model, we perform a leave-one-sequence-out cross-validation scheme on the training set. After 100 epochs, our models achieve an MAE of 4.97 and an RMSE of 6.26, which we find encouraging. We further retrain the model on the full training set and monitor the performance on the development partition.

Figure 1 shows the learning plots of the loss function during training for both training (black) and validation (red) sets. We observe a monotonic decrease of the loss function on the training set, while on the validation, the behavior is a typical decrease, followed by an increase of the same loss. We use the validation set to early stop the training, thus resulting in a model ($lstm\_opt$) with the best performance on this set. As can be observed from Table II, our LSTM model manages to score a promising MAE on the development partition, better than all baselines, as well as in the last year's winning paper [41].

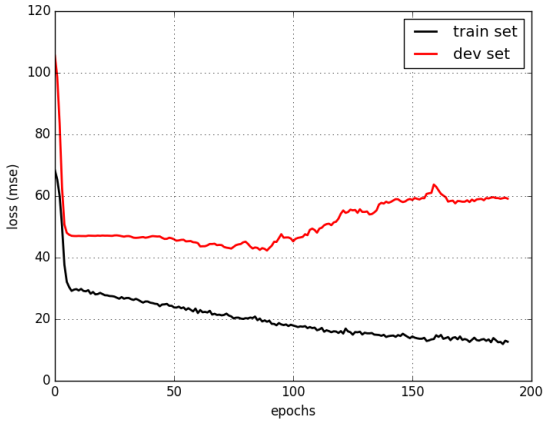

Fig. 1: LSTM learning curves: training set (black) and development set (red). We note the existence of a turning point in the validation loss, typically used as a good compromise between underfitting and overfitting

D. Final evaluation on the test set

In the context of the AVEC challenge, methods are compared on a specific test set. Results on the test set are presented in Tab. III. The behavioral characteristic features extracted from audio transcriptions achieve the lowest errors on the test partition, which is unsurprising considering the promising results obtained on the the development set (*i.e.* RMSE of 5.54 and MAE of 4.73). The winners [53] of the AVEC 2017 on depression sub-task have achieved RMSE of 4.99, which is outperformed by our behavioral characteristics model with RMSE of 4.94.

V. Conclusions

In this paper we address the depression sub-challenge problem formulated in AVEC2017, *i.e.* regressing the total sum of PHQ-8 depression scores from multi-modal data. We process different modalities (audio, language, visual)

TABLE III: Root mean square error (RMSE) and mean absolute error (MAE) for depression severity regression using *acoustic*, *behavioral*, *lexical*, and *visual* features on the AVEC 2017 test set.

	RMSE	MAE
Spectral features (speech)	6.63	5.08
Behavioral features (speech)	**4.94**	**4.11**
Language features (text)	5.83	4.88
Video features	6.72	5.36

accompanying the corpus and developed regression systems separately. In the audio domain, we find the spectral features to be most suited for this task, achieving an MAE score of 4.96 on the development set (RMSE = 6.32) while lexical features score no lower than 5.17 (MAE) and 6.31 (RMSE). Despite being the worst performing modality in the baseline manuscript, visual features achieve the smallest errors on the development set in our experiments. Using a sliding window approach and temporal modeling, we obtain an MAE of 4.66 (RMSE = 6.09). We also observed that behavioral cues extracted from transcripts achieve smaller errors (MAE = 4.73, RMSE = 5.54) compared to audio and language features and are good predictor of the depression severity scores. When studied further, we found that previous diagnosed information cues, participants' response time to the agent among others are one of the most informed feature to predict the depression PHQ-8 scores. This is indeed confirmed by the results obtained on the test set, where behavioral cues scored the smallest MAE values among all other feature sets.

In this paper, we have studied each modality individually to understand its strength in estimating the depression severity. In future work, we plan investigating how we can combine individual modalities to improve the overall performance.

REFERENCES

1 Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J., Vos, T., and Whiteford, H. A., "Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010," *PLoS medicine*, vol. 10, no. 11, 2013.

2 Sheline, Y. I., Gado, M. H., and Kraemer, H. C., "Untreated depression and hippocampal volume loss," *American Journal of Psychiatry*, vol. 160, no. 8, pp. 1516–1518, 2003.

3 Nicholson, A., Kuper, H., and Hemingway, H., "Depression as an aetiologic and prognostic factor in coronary heart disease: a meta-analysis of 6362 events among 146 538 participants in 54 observational studies," *European heart journal*, vol. 27, no. 23, pp. 2763–2774, 2006.

4 Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

5 Patel, P., Chaudhari, A., Kale, R., and Pund, M., "Emotion recognition from speech with gaussian mixture models & via boosted gmm," *International Journal of Research In Science & Engineering*, vol. 3, 2017.

6 Hansen, J. H., "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech communication*, vol. 20, no. 1-2, pp. 151–173, 1996.

7 Yogesh, C., Hariharan, M., Ngadiran, R., Adom, A. H., Yaacob, S., Berkai, C., and Polat, K., "A new hybrid pso assisted biogeography-based optimization for emotion and stress recognition from speech signal," *Expert Systems with Applications*, vol. 69, pp. 149–158, 2017.

8 Vanello, N., Guidi, A., Gentili, C., Werner, S., Bertschy, G., Valenza, G., Lanata, A., and Scilingo, E. P., "Speech analysis for mood state characterization in bipolar patients," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 2104–2107.

9 Chang, K.-h., Fisher, D., Canny, J., and Hartmann, B., "How's my mood and stress?: an efficient speech analysis library for unobtrusive monitoring on mobile phones," in *Proceedings of the 6th International Conference on Body Area Networks*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011, pp. 71–77.

10 Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E., "Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech," in *Mechatronics and Automation, 2005 IEEE International Conference*, vol. 3. IEEE, 2005, pp. 1569–1574.

11 Fraser, K., Rudzicz, F., Graham, N., and Rochon, E., "Automatic speech recognition in the diagnosis of primary progressive aphasia," in *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, 2013, pp. 47–54.

12 Bersani, G., Polli, E., Valeriani, G., Zullo, D., Melcore, C., Capra, E., Quartini, A., Marino, P., Minichino, A., Bernabei, L. *et al.*, "Facial expression in patients with bipolar disorder and schizophrenia in response to emotional stimuli: a partially shared cognitive and social deficit of the two disorders," *Neuropsychiatric disease and treatment*, vol. 9, p. 1137, 2013.

13 Zwirner, E., "Contribution to the speech of depressives. phonometry iii. special applications i," 1930.

14 Newman, S. and Mather, V. G., "Analysis of spoken language of patients with affective disorders," *American journal of psychiatry*, vol. 94, no. 4, pp. 913–942, 1938.

15 France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., and Wilkes, M., "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

16 Pope, B., Blass, T., Siegman, A. W., and Raher, J., "Anxiety and depression in speech." *Journal of Consulting and Clinical Psychology*, vol. 35, no. 1p1, p. 128, 1970.

17 Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., and Bao, Z., "A depression detection model based on sentiment analysis in micro-blog social network," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013, pp. 201–213.

18 Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V., McCoy, T., and Perlis, R., "Predicting early psychiatric readmission with natural language processing of narrative discharge summaries," *Translational psychiatry*, vol. 6, no. 10, p. e921, 2016.

19 Littlewort, G. C., Bartlett, M. S., and Lee, K., "Automatic coding of facial expressions displayed during posed and genuine pain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1797–1803, 2009.

20 Ambadar, Z., Cohn, J. F., and Reed, L. I., "All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous," *Journal of nonverbal behavior*, vol. 33, no. 1, pp. 17–34, 2009.

21 Ekman, P. and Friesen, W. V., "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

22 Craig, S. D., D'Mello, S., Witherspoon, A., and Graesser, A., "Emote aloud during learning with autotutor: Applying the facial action coding system to cognitive–affective states during learning," *Cognition and Emotion*, vol. 22, no. 5, pp. 777–788, 2008.

23 Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J., "Fully automatic facial action recognition in spontaneous behavior," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*. IEEE, 2006, pp. 223–230.

24 Russell, J. A., "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies." *Psychological bulletin*, vol. 115, no. 1, p. 102, 1994.

25 Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., and Rosenwald, D. P., "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image and vision computing*, vol. 32, no. 10, pp. 641–647, 2014.

26 Alghowinem, S., Goecke, R., Wagner, M., Parkerx, G., and Breakspear, M., "Head pose and movement analysis as an indicator of depression," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 283–288.

27 Joshi, J., Goecke, R., Parker, G., and Breakspear, M., "Can body expressions contribute to automatic depression analysis?" in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–7.

28 Scherer, S., Stratou, G., Lucas, G., Mahmoud, M., Boberg, J., Gratch, J., Morency, L.-P. *et al.*, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014.

29 Gratch, J., Artstein, 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 *LREC*, 2014, pp. 3123–3128.

30 Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., and Mokdad, A. H., "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.

31 Eyben, F., Weninger, F., Gross, F., and Schuller, B., "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

32 Schuller, B., Batliner, A., Steidl, S., and Seppi, D., "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

33 Alam, F. and Riccardi, G., "Comparative study of speaker personality traits recognition in conversational and broadcast news speech." in *INTERSPEECH*, 2013, pp. 2851–2855.

34 Chowdhury, S. A. and Riccardi, G., "A deep learning approach to modeling competitiveness in spoken conversation," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

35 Chowdhury, S. A., Danieli, M., and Riccardi, G., "Annotating and categorizing competition in overlap speech," in *Proc. of ICASSP*. IEEE, 2015.

36 Chowdhury, S. A., Riccardi, G., and Alam, F., "Unsupervised recognition and clustering of speech overlaps in spoken conversations," in *Proc. of Workshop on Speech, Language and Audio in Multimedia*, 2014.

37 Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

38 Mikolov, T., Chen, K., Corrado, G., and Dean, J., "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

39 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 2011.

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

41 Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M. C., and 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*. ACM, 2016, pp. 89–96.

42 Hochreiter, S. and Schmidhuber, J., "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

43 Pham, V., Bluche, T., Kermorvant, C., and Louradour, J., "Dropout improves recurrent neural networks for handwriting recognition," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014, pp. 285–290.

44 Doetsch, P., Kozielski, M., and Ney, H., "Fast and robust training of recurrent neural networks for offline handwriting recognition," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014, pp. 279–284.

45 Graves, A., Mohamed, A.-r., and Hinton, G., "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

46 Zaremba, W., Sutskever, I., and Vinyals, O., "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

47 Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., and Zaremba, W., "Addressing the rare word problem in neural machine translation," *arXiv preprint arXiv:1410.8206*, 2014.

48 Marchi, E., Ferroni, G., Eyben, F., Gabrielli, L., Squartini, S., and Schuller, B., "Multi-resolution linear prediction based features for audio onset detection with bidirectional lstm neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2164–2168.

49 Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

50 Sak, H., Senior, A., and Beaufays, F., "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

51 Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

52 Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

53 Gong, Y. and Poellabauer, C., "Topic modeling based multi-modal depression detection," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '17. New York, NY, USA: ACM, 2017, pp. 69–76. [Online]. Available: http://doi.acm.org/10.1145/3133944.3133945