Detecting Depression with Audio/Text Sequence Modeling of Interviews

Tuka Alhanai[1]*, Mohammad Ghassemi*[2]*, and James Glass*[1]

[1]Computer Science and Artificial Intelligence Laboratory
[2]Institute for Medical Engineering and Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{tuka,ghassemi,glass}@mit.edu

Abstract

Medical professionals diagnose depression by interpreting the responses of individuals to a variety of questions, probing lifestyle changes and ongoing thoughts. Like professionals, an effective automated agent must understand that responses to queries have varying prognostic value. In this study we demonstrate an automated depression-detection algorithm that models interviews between an individual and agent and learns from sequences of questions and answers without the need to perform explicit topic modeling of the content. We utilized data of 142 individuals undergoing depression screening, and modeled the interactions with audio and text features in a Long-Short Term Memory (LSTM) neural network model to detect depression. Our results were comparable to methods that explicitly modeled the topics of the questions and answers which suggests that depression can be detected through sequential modeling of an interaction, with minimal information on the structure of the interview.

Index Terms: medical speech signal processing, depression, neural networks, computational paralinguistics, question answering

1. Introduction

Individuals suffering from depression are beset by debilitating sadness for weeks to years on end [1]. To treat depressed individuals, they must first be diagnosed. To obtain a diagnosis, depressed individuals must actively reach out to mental health professionals. In reality, it can be difficult for the depressed to attain professional attention due to constraints of mobility, cost, and motivation. Passive automated monitoring of human communication may address these constraints and provide better screening for depression [2].

The standard method for screening and diagnosing depression is the Patient Health Questionnaire (PHQ), which has been designed by the American Psychological Association [3]. The questionnaire packages the DSM-IV depression criteria into a brief self-report instrument which asks whether the individual finds pleasure in doing things, feels down, tired, has a poor appetite, trouble concentrating, is slow or fidgety, and/or struggles to sleep [4].

Methods to automate this screening process have pursued two approaches. The first approach models a subject's outcome based on responses to *specific* questions (e.g. 'Do you have a history of depression?'), while the second approach models outcomes based on responses that are *independent* of the question asked (e.g. speaking rate). In the first approach, Arroll *et al.* explored asking key sets of questions that optimized for increased accuracy and minimized time spent screening [5]. Another example of this approach, manually selects questions of which responses are most predictive of a subject's state and assign weights according to the text-based illicited response [6]. In a similar spirit, Yang *et al.* (2016) and Sun *et al.* modeled depression by structuring questions and responses in the form a decision tree [7, 8], while Gong *et al.* developed an ensemble of audio, text, and video features as a function of the question type asked [9]. Utilizing a deep learning framework, Yang *et al.* (2017) combined multiple modalities conditioned on manually selected questions [10].

The second approach to modeling depression attempts to exploit global and/or time varying statistics, independent of the question that prompted the response. Williamson *et al.* utilized correlations of formants and spectral information across different time scales [11], Syed *et al.* developed audio and video features to capture temporal variations [12], while Pampouchidou *et al.* and Nasir *et al.* fused low and high-level features [13, 14]. Utilizing emerging techniques, Ma *et al.* used audio to model depression by allowing deep neural networks to learn such associations rather than perform feature engineering [15].

Given the question-answer nature of depression screening tests, we were interested in modeling depression via sequences of responses, without the need to formally condition on the type of questions being asked. Such a system has the advantage of being data-driven with minimal need for a-priori knowledge of the structure of an interview or interaction. Furthermore, for a model to be truly data-driven, it needs to have minimal feature engineering. Current developments due to more affordable computational and storage infrastructure, as well as increased data streams, have allowed deep learning based methods to become accessible [16]. Their strengths lie in their ability to represent information through non-linear transforms, at varying spatial and temporal resolution, and from multiple modalities [17, 18]. While work in the domain of detecting depression has looked at fusing features from multiple modalities together [9, 13, 14, 19], and utilizing neural networks to model single sequences [10, 15], there remains to explore the sequence modeling of depression that utilizes deep learning approaches.

To this end we conducted our work on the same dataset (the distress analysis and interview corpus) that most of the above mentioned studies utilized, which allowed us to compare methods and performance results [20].

2. Objective

In this study our objective was to detect depression by modeling audio and text sequences of an interaction between a human subject and a virtual agent. We were motivated to perform this modeling in a data-driven manner, without the need to formally condition on the question being asked, given the potential utility of such techniques.

3. Data

3.1. Audio and Text

We utilized the audio and text transcriptions of 142 individuals undergoing depression screening through a human-controlled virtual agent. The virtual agent prompted each individual with a subset of 170 possible queries that included direct questions (e.g. 'How are you?', 'Do you consider yourself to be an introvert?'), and dialogic feedback (e.g. 'I see', 'that sounds great'). The data was from the publicly available distress analysis and interview corpus (DAIC) and contains audio and text transcriptions of the spoken interactions [21, 20]. The data was split into a training (57%, 107 subjects), development (19%, 35 subjects), and test (25%, 47 subjects) sets as specified by [21, 20]. The test set annotations were not provided in the DAIC public release so all models were evaluated on the development set.

3.2. Outcomes of Interest

We were interested in modeling (1) the binary state of a subject (depressed or not), as well as (2) the severity of their depression. The severity of depression ranged from 0 to 27 with a score from 0-4 considered none or minimal, 5-9 mild, 10-14 moderate, 15-19 moderately severe, and 20-27 severe. A soft cutoff within the moderate and moderately severe range resulted in binary outcomes. These outcomes were pre-defined in the DAIC dataset, and were derived from the PHQ-8 depression questionnaire screening the subjects underwent [22]. Within the dataset, 28 out of 142 subjects (20%) were labeled as depressed.

4. Experimental Approach

In this study, we sought to model sequences of 142 interactions in order to detect whether individuals were depressed. We conducted three sets of experiments using features extracted from the audio and text data to predict depression.

Exp 1 A regularized logistic regression model *without* conditioning on the type of questions asked.

Exp 2 A regularized logistic regression model *with* conditioning on the type of questions asked.

Exp 3 An LSTM model using the *sequences* of responses, and *without* knowledge of the type of questions that prompted the response.

Details of the experiments are outlined below. Our code is available in an online repository [1]

4.1. Experiment 1: Context-free Modeling

4.1.1. Model

We were interested in assessing the predictive performance of several audio and text features (described below), when considered independently of the *type* of question asked, and *time* it was asked during the interview session (i.e. what we term as 'context-free' modeling). For this analysis, we provided 279 audio and 100 text features to a logistic regression model with L1 regularization.

4.1.2. Text Features

Using Doc2Vec of the Python Gensim library, we generated embeddings of individual responses to all queries and the queries themselves, for a total of 8,050 training examples, 272,418 words, and a vocabulary size of 7,411 [23]. These embeddings were trained with the following explored hyperparameters: minimum word `count` of {1, 2, 3, 4, 5, 7, 10}, a context `window` of {3, 5, 7, 10, 12, 15} words, `dimensions` of {50, 75, 100, 125, 150, 175, 200, 225, 250, 300, 350, 400}, downsampling `threshold` of 1e-04, `hierarchical sampling`, using the `distributed memory` training algorithm (akin to Continuous Bag of Words in Word2Vec), random generator `seed` of 1, and training `epochs` of {15, 25, 35, 50}. We selected the value of the hyperparameters that optimized the model's performance on the training set. The optimum embedding dimension was found to be 100, with a minimum word count of 3, context window size of 3, and 50 training epochs.

4.1.3. Audio Features

We extracted an initial set of 553 features representing each subject response. The features were higher-order statistics (mean, maximum, minimum, median, standard deviation, skew, and kurtosis) of 79 COVAREP features provided with the DAIC dataset [20]. The COVAREP features were frame-level (20 ms window, 10 ms shift) features composed of spectral (Mel-frequency cepstral coefficients 0-24, harmonic model and phase distortion mean 0-24 and deviations 0-12) prosodic (pitch, voicing probability, formants 1-5), and voice quality (normalized amplitude quotient, quasi open quotient, difference in amplitude of the first two harmonics of the differentiated glottal source spectrum, parabolic spectral parameter, maxima dispersion quotient, spectral tilt/slope of wavelet responses, and shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics) features. Zero-mean and variance normalization was applied to all features, and any segments without audio information were set to zero. From the initial set of 553 features, we excluded all features without a statistically significant univariate correlation with outcomes on the training set ($|\rho| <$ 1e-01, $p >$ 1e-02) nor a significant L1 regularized logistic regression model coefficient ($|\beta| <$ 1e-04), thus resulting in a subset of 279 features and 8,050 examples (responses).

4.2. Experiment 2: Weighted Modeling

4.2.1. Model

We were interested in assessing the predictive performance of several audio and text features (described previously), when conditioning on the *type* of question asked, and independent of the *time* it was asked during the interview session (i.e. what we term as 'weighted' modeling). For this analysis, we provided 279 audio and 100 text features to a logistic regression model with L1 regularization, and weighted the model probabilities based on the predicitive power of the question found in the training set.

4.2.2. Assigning Question Value

In our dataset, each subject i, was asked a subset of queries q_i from a set of Q possible queries. We represented the responses to the queries as a matrix $V_i \in \mathbb{R}^{q_i \times m}$, where m was the number of features (m = 100 for text, and 279 for audio). Each response matrix V_i, had a corresponding binary outcome vector indicating depression $y_i \in \mathbb{R}^{q_i \times 1}$.

To train the model, we horizontally concatenated subject response matrices into a training and development matrix, $A_{train} \in \mathbb{R}^{n \times m}$ and $A_{dev} \in \mathbb{R}^{d \times m}$ where n was the num-

[1] `https://github.com/talhanai/redbud-tree-depression`

ber of training examples and d was the number of development examples. A_{train} was then used to train the model.

Next, let $c(j)$ represent the performance of the trained model when evaluated using only the rows of A_{train} that corresponded to a specific query, $j \in \{1 : Q\}$. We identified the set of k informative queries with predictive performance above a particular threshold θ on the training set: $k = \{j \mid c(j) >= \theta\}$.

For the subset of queries in k, we assigned query weights equal to the training set performance $w(j) = c(j)$. These weights were used in conjunction with the logistic regression model f to provide a question-weighted probability p of depression:

$$p = \frac{1}{|k|} \sum_{j=1}^{|k|} f * w(j)$$

4.3. Experiment 3: Sequence Modeling

4.3.1. LSTM Model

The strength of neural networks lies in their ability to extract feature representations through non-linear transforms of the input data, yielding stronger discriminative power than classical models. Since we were interested in modeling temporal changes of the interview, we utilized a bi-directional Long Short-Term Memory (LSTM) neural network since it has the additional advantage of modeling sequential data. To find the optimum topology of the LSTM model we explored the following hyperparameter space: number of `layers` {1, 2, 3, 4}, number of hidden `nodes` in each layer {4, 8, 16, ..., 256}, 'tanh' `activation` function, and hard sigmoid recurrent activation function, input and recurrent `dropout` rates of {0, 0.1, 0.2, 0.4, 0.6, 0.8, 0.8}, `merge` mode of {sum, mul, concat, ave}, and `batch` size of {32, 64, 128, ..., 4096}. For the `loss` function we used 'binary crossentropy' to model binary outcomes, and {'categorical crossentropy', 'mean squared error', 'mean absolute error'} for multi-class outcomes. The

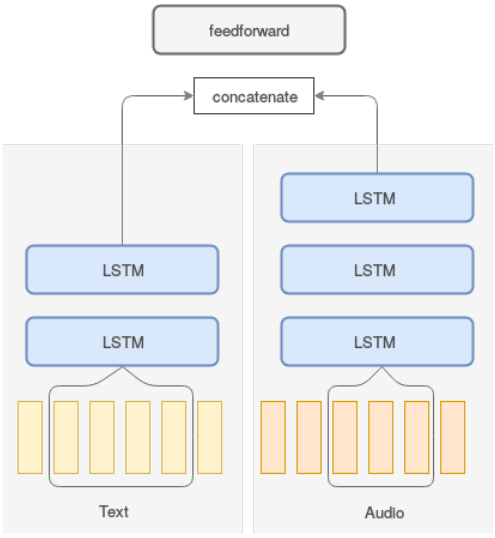

Figure 1: ***Diagram of network topology.*** *Each modality (audio and text) were trained separately as bi-directional LSTMs with differing hyperparameters capturing the characteristics of each feature set. A multi-modal model that combined both audio and text was also trained through concatenation into feedforward layers.*

`optimizer` algorithm was stochastic gradient descent with `learning rates` of {1e-01, 1e-02, ..., 1e-06}, `momentum` {0, 0.1, 0.2, 0.4, 0.6, 0.8, 0.85, 0.9, 0.95, 0.99 }, and step `decay` rates {0.9, 0.99, 1}. The training early stopping criterion was set for a minimum loss change of 1e-4 for a duration of 25 epochs. For the input data we explored `timesteps` (responses) of [2, 30], and `strides` (contiguous responses) of [1, 3]. We trained an LSTM model for each of the audio and text modalities separately, utilizing the response-level features of Experiment 1.

The audio-based LSTM model had 3 bi-directional LSTM layers, each with 128 hidden nodes, multiplicative merge mode, and input and recurrent dropout rates of 0.2. The inputs to the model had a timestep of 20, and stride of 1. The learning rate was 1e-06, the momentum was 0.8, decay rate was 0.99, with a batch size of 128. The text-based LSTM model had 2 bi-directional LSTM layers, each with 4 hidden nodes, concatenated merge mode, and input and recurrent dropout rates of 0.1 and 0.8 respectively. The inputs to the model had a timestep of 7, and stride of 3. The learning rate was 1e-01, the momentum was 0.85, no decay, with a batch size of 64.

4.3.2. Multi-modal LSTM Model

Audio and text features may contain not only discriminative and temporally varying information about a subject's state, but also *complementary* information. Therefore we trained a model that combined these two modalities in the form of a multi-modal model. The model was composed of two LSTM branches, one for each of the modalities, with their outputs merged into a final feedforward network. The branches were composed of different topologies, and were optimized with respect to the characteristics and information content of each modality. The weights of the branches were fixed, their outputs were concatenated, and the final feedforward network topology and weights were trained according to the explored hyperparameter space defined earlier, with an additional hyperparameter of `activations` {'tanh', 'sigmoid', 'relu'}. The optimum feedforward network was composed of 2 layers, with 128 hidden nodes, and'tanh' activations. The learning rate was 1e-05, with momentum 0.8, no decay, and a batch size of 32.

The audio and text inputs for each LSTM branch had different strides and timesteps yielding a different number of training (and development) examples, therefore we needed to equalize the number of examples (Audio was 30 timesteps, with stride 1. Text was 7 timesteps, and stride 3). This step was performed by padding the number of training examples in the smaller set (text) to match that larger set (audio) by mapping examples together that appeared in the same window of the interview. For this experiment we utilized the Keras library and Tensorflow back-end [24, 25].

4.4. Baselines

We compared our methods with the DAIC dataset baseline [20] as well as three other reported methods that utilized the same dataset, and were either the best performing or whose experimental approach most closely related to our work. To summarize, the DAIC baseline utilized an ensemble of features (audio, text, and video) in a Support Vector Machine (SVM) model, Ma *et al.* applied a convolutional neural network followed by an LSTM on the audio in a given sequence segment [15], Williamson *et al.* applied both unsupervised word representation techniques as well as context/topic modeling (weighted Glove embeddings) on the text [6], while Gong *et al.* modeled

the ensemble of features while conditioning on the topic [9].

4.5. Evaluation Metrics

Following the structure of the DAIC baseline [20], we reported the results of the development set for the binary classification task using F1 score, precision, and recall. For the multi-class classification task (categorical range of 0-27) we reported the subject-level mean absolute error (MAE) and subject-level root mean squared error (RMSE). We also performed an evaluation according to the 'fusion scoring' metric utilized by Williamson *et al.* [6], whereby the top N (in our case $N = 3$) most predicted depression class(es) among all the segments of a subject were selected as the subject's predicted outcome.

5. Results

The results of the experiments, as well as the baselines, are displayed in Table 1.

Table 1: ***Results***. *Baselines and our approach. Best in* ***bold***.

Model	Features	F1	Prec.	Rec.	MAE	RMSE
Baseline Approaches						
Baseline [20]	(Ensemble)	.50	.60	.43	6.62	5.52
Williamson *et al.* [6]	(Audio)	.50	/	/	5.36	6.74
Ma *et al.* [15]	(Audio)	.52	.35	1.00	/	/
Gong *et al.* [9]	(Ensemble)	.70	/	/	2.77	3.54
Williamson *et al.* [6]	(Text)	.76	/	/	/	/
†Williamson *et al.* [6]	(Text)	.84	/	/	3.34	4.46
Our Approach						
Context-free	(Audio)	.50	.71	.38	5.31	6.94
Context-free	(Text)	.59	.71	.50	7.02	9.43
Weighted	(Audio)	.67	**1.00**	.50	7.60	10.03
Weighted	(Text)	.44	**1.00**	.29	7.32	8.85
Sequence	(Audio)	.63	.71	.56	5.13	6.50
Sequence	(Text)	.67	.57	.80	5.18	6.38
Multi-modal	(Audio+Text)	**.77**	.71	**.83**	5.10	6.37
†Multi-modal	(Audio+Text)	.43	.43	.43	**4.97**	**6.27**

†Fusion scoring.

5.1. Experiment 1: Context-free Modeling

When conducting context-free modeling of the interviews, text features performed better than audio features when classifying for a binary outcome (F1 0.59 vs. 0.50), with a higher recall rate (0.50 vs. 0.38), and equivalent precision (0.71). However, audio features were more accurate in determining the multi-class depression score (MAE 5.01 vs. 7.02).

5.2. Experiment 2: Weighted Modeling

When weighting the model according to the questions asked of the subject, audio features performed better than text features (F1 0.67 vs. 0.44) with perfect rates of precision (1.00), but at the cost of recall (0.50 and 0.29 for audio and text). The overall performance of audio improved when conditioning on the question being asked (F1 0.67 vs. 0.50). The overall MAE and RMSE for both modalities decreased compared to the previous experiment.

5.3. Experiment 3: Sequence Modeling

Sequence models utilizing text features performed better (F1 0.67) than the context-free and were on par with the weighted models, while sequence models utilizing audio features performed better (F1 0.63) than the context-free model. The recall rates of the sequence models were higher than the previous models (0.56 and 0.80 for audio and text). Combining both modalities into a multi-modal model yielded the highest performance (F1 0.77, recall 0.83). Sequence models also displayed the best multi-class classification performance. Conducting fusion scoring on the multi-modal model, resulted in the best multi-class classification score (MAE 4.97, RMSE 6.27). A majority of our models out-performed the baseline results, and our multi-modal model performed better than previous work with respect to F1 score (sans the fusion scoring baseline).

6. Summary

In this study, we sought to model sequences of 142 interactions in order to detect whether individuals was depressed during the course of their interview. We conducted three sets of experiments where audio and text modalities were modeled (1) *without* the question that prompted the response, (2) *with* the context by conditioning on the question asked, and (3) with respect to the *sequence* of the responses (and *without* conditioning on the question asked).

6.1. Information over Time and Across Modalities

We observed that while context-free modeling does provide some discriminative power, sequence modeling is more accurate (highest binary F1 score) and/or robust (lowest multi-class MAE, RSME) for predicting depression. This indicates that the model was capturing information across sequences. We also observed that the optimum input parameters for each modality were different. Text was provided to the model in timesteps of 7 and a stride of 3, while the audio was provided in timesteps of 30 and stride 1. This indicates that temporally varying and discriminative information of the way a depressed person may speak as contained in the audio, exists at longer time intervals relative to the syntactic and semantic information contained in the text. The multi-modal model yielded the best performance which shows that not only did a combination of modalities provide additional discriminative power, but that they contained *complementary* information.

6.2. Model Calibration

If a model is to be deployed as a screening tool, then how well a model is calibrated becomes important for its utility [26]. In this study, we defined a model to be well calibrated if it had a relatively low MAE and RMSE in its multi-class performance. We noticed that the text-based weighted model performed worse than the context-free model (0.44 from 0.59 F1, albeit with precision). This may be because the text-based context-free model was not as well calibrated, as indicated by its relatively high error in the multi-class result (MAE 7.02), as a result, the discriminative power of conditioning on the question may have been too noisy. Further evidence to this behavior can be observed in the audio-based model, whereby the context-free model that was relatively better calibrated at the multi-class level (MAE 5.31), yielded a weighted model with improved performance (0.50 to 0.67 F1). Moreover, applying fusion scoring improved multi-class performance at the cost of binary classification performance. We also hypothesize that this was due to relatively poor calibration. While some of the baselines performed better in the classification task, it is important to note that they both applied model calibration techniques. Gong *et al.* had performed cross-validation on both the training and development set for hyperparameter optimization [9], while Williamson *et al.* - who explicitly modeled the the topic of the question, whereas we did not - applied probability scaling to their model predictions [6].

7. References

[1] A. J. Ferrari, F. J. Charlson, R. E. Norman, S. B. Patten, G. Freedman, C. J. Murray, T. Vos, and H. A. Whiteford, "Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010," *PLoS Med*, vol. 10, no. 11, p. e1001547, 2013.

[2] A. Pinto-Meza, A. Serrano-Blanco, M. T. Penarrubia, E. Blanco, and J. M. Haro, "Assessing depression in primary care with the phq-9: can it be carried out over the telephone?" *Journal of general internal medicine*, vol. 20, no. 8, pp. 738–742, 2005.

[3] B. Arroll, F. Goodyear-Smith, S. Crengle, J. Gunn, N. Kerse, T. Fishman, K. Falloon, and S. Hatcher, "Validation of phq-2 and phq-9 to screen for major depression in the primary care population," *The Annals of Family Medicine*, vol. 8, no. 4, pp. 348–353, 2010.

[4] 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, pp. 163–173, 2009.

[5] B. Arroll, F. G. Smith, N. Kerse, T. Fishman, and J. Gunn, "Effect of the addition of a help question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study," *Bmj*, vol. 331, no. 7521, p. 884, 2005.

[6] 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 *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 11–18.

[7] 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 *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 89–96.

[8] 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*. ACM, 2017, pp. 61–68.

[9] Y. Gong and C. Poellabauer, "Topic modeling based multimodal 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

[10] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 53–59.

[11] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 41–48.

[12] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression severity prediction based on biomarkers of psychomotor retardation," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 37–43.

[13] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias *et al.*, "Depression assessment by fusing high and low level features from audio, video, and text," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 27–34.

[14] 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 *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 43–50.

[15] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 35–42.

[16] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[17] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[18] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 966–973.

[19] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 45–51.

[20] 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 *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[21] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *LREC*, 2014, pp. 3123–3128.

[22] S. Gilbody, D. Richards, S. Brealey, and C. Hewitt, "Screening for depression in medical settings with the patient health questionnaire (phq): a diagnostic meta-analysis," *Journal of general internal medicine*, vol. 22, no. 11, pp. 1596–1602, 2007.

[23] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in *ICML*, vol. 14, 2014, pp. 1188–1196.

[24] F. Chollet *et al.*, "Keras," 2015.

[25] M. Abadi, "Tensorflow: A system for large-scale machine learning."

[26] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General*, vol. 144, no. 1, p. 114, 2015.