Depression Severity Prediction by Multi-model Fusion

Bin Li, Jie Zhu, Chunpeng Wang
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
Emails: lb0621@sjtu.edu.cn, zhujie@sjtu.edu.cn, 2995078502@qq.com

***Abstract*—Depression is a common mental disorder and over 300 million people are estimated to suffer from it. Current main methods to predict the severity of depression depend on clinical interviews. Although very useful, this method lacks objectivity and efficiency. In this paper, we propose a multi-model fusion framework to detect the depression severity based on the random forest machine learning algorithms, and using features selected from different mediums. We first selected features of audio, video, and text contents of each patient with a fusion strategy using the decision-level fusion. The multi-model fusion regressors were trained with the extracted features to obtain the depression severity. To handle the imbalanced dataset problem, a sampling strategy was also conducted. Experiments were demonstrated on the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) database. The proposed framework achieved a Root Mean Square Error (RMSE) of 5.98, which is less than the baseline of 6.62 and suggest that our approach is efficient and suitable for the depression severity detection.**

Keywords–Depress detection; multi-model; fusion; machine learning.

I. Introduction

Depression is one of the most common psychological disorders, and a leading cause of disease burden [1]. According to the World Health Organization (WHO), over 300 million people are estimated to suffer from depression, i.e., $\sim$ 4.4% of the world's population [2]. As a global disease, the number of patients keeps increasing, especially in lower-income countries, where most of citizens are living to their age of high possibilities to have depression or other mental diseases. In addition to the persistent feelings of negativity, sadness, loss of interest or pleasure, and low self-esteem, depression is also corresponding to a range of physiological symptoms such as weight loss, insomnia and fatigue. It may be associated with diseases like bipolar disorder dementia [3] and even cardiovascular conditions [4]. Depression is never caused by only one thing, in the other words, a combination of factors such as biological factors, psychological factors and stressful life could be the reasons for depression. It has been identified as a burden to the economy while it affects justice and social systems [5]. Furthermore, the severe depression may lead to commit suicide and substance abuse [6]. Therefore, the detection and treatment of depression and its severity is of a high priority.

To measure the severity of depression, the Distress Analysis Interview Corpus (DAIC) is constructed, which contains clinical interviews in English for the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder [7]. The interviews were conducted by an animated virtual interviewer called Ellie (see Fig. 1), where she asked the interviewee a series of open-ended questions with intent of identifying clinical symptoms (see Fig. 2 for an example). The corpus includes audio and video recordings and extensive questionnaire responses. Each interview includes a depression score from the eight-item Patient Health Questionnaire depression scale (PHQ-8), which is established as a valid diagnostic and severity measure for depressive disorders in large clinical studies [8]. The PHQ-8 is a set of eight short multiple choices questions and every choice is corresponding to a score ranging from 0 to 3. Thus, the total point ranges from 0 to 24. A threshold of PHQ-8 score $\geq$ 10 is set to define current depression severity.

To predict the depression severity, many approaches have been proposed based on signal processing, computer vision, and machine learning algorithms. For instances, vocal utterances were taken advantaged to identify depression in [9]–[12], and intensive studies have been taken into facial expressions for the diagnostic of depression [13] [14]. The other researchers try to analyze depression from the text or language information [15] [16]. These approaches obtained results to some extent. The performance of using only one category feature is limited. They did not take full advantage of information we could get from the people we want to diagnose.

On machine learning based approaches, the decision-level fusion from multi-model strategy have been widely applied and proved efficient and accurate. Features were extracted from multiple mediums such as audio, video and text. Senoussaoui et al. proposed an i-vector based representation by extracting short term audio features for depression classification and prediction [17]. In [18], they used five acoustic feature categories, including prosodic, cepstral, spectral, glottal and, Teager Energy Operator (TEO) based features, to detect depression. Above approaches extracted features from one medium. In our view, features from multiple mediums may benefit the training of the regressors, which has been proved efficient in many works [19]. For the regressors for estimating the depression severity, some machine learning algorithms like support vector regressor (SVR), and random forest (RF) are applied. And the fusion or boosting by combining multiple models have been suggested with increase of those single model [20]. In [20], Morales and Levitan provided investigation of speech versus text features for depression detection systems, finding that multi-model system leads to the best performance. They also used automatic speech recognition to transcribe speech and found text features generated form ASR transcripts were useful for depression detection. In this paper, we propose to build a multi-model fusion framework, which fuses features extracted from audio, video, and text in the DAIC, and detect

 ISBN: 978-1-61208-675-0

Figure 1. Ellie, the virtual interviewer

```
Wizard-of-Oz
Ellie Who's someone that's been a positive
influence in your life?
Participant Uh my father.
Ellie Can you tell me about that?
Participant Yeah, he is a uh
Participant He's a very he's a man of few words
Participant And uh he's very calm
Participant Slow to anger
Participant And um very warm very loving man
Participant Responsible
Participant And uh he's a gentleman has a great
sense of style and he's a great cook.
Ellie Uh huh
Ellie What are you most proud of in your life?
```

Figure 2. Sample excerpts from virtual interviews

depression severity by predicting the PHQ-8 score of the patients.

This paper is organized as follows. In Section II we describe the selection of audio, video, and text features. The multiple-model fusion framework is explained in Section III. Experiments and results are demonstrated and discussed in Section IV. We conclude in Section V with outlooks.

II. Feature Extraction and Selection

The DAIC-WOZ dataset provides a set of samples to train the regressor, as well as features from audio, video recordings, as well as questionnaire responses in text format. However, there are two problems that should be considered before we begin our work. One is that the dataset is imbalanced, which has been found affecting the performance of machine learning algorithms [21]. In the DAIC-WOZ dataset, the number of samples is larger for the depressed than the non-depressed, which also leads an imbalanced distribution of the PHQ-8 scores. The distribution of DAIC-WOZ dataset in training and development sets is provided in TABLE I.

TABLE I. Distribution of DAIC-WOZ dataset in training and development sets

Dataset	*Depress*	*Non-depress*	*Sum*
Training	77	30	107
Development	22	13	35

TABLE II. Audio Features

Type	*Feature Name*
Prosodic	Fundamental frequency (F0), voicing (VUV)
Voice Quality	Normalised amplitude quotient (NAQ), quasi open quotient (QOQ), the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), spectral tilt/slope of wavelet responses (peak-slope), and shape parameter of the Liljencrants-Fantmodel of the glottal pulse dynamics (Rd)
Spectral	Mel cepstral coefficients (MCEP0-24), harmonic model and phase distortion mean (HMPDM0-24) and deviations (HMPDD0-12)
Formants	The first 5 formants

To handle this problem, we random oversample on the minor depressed dataset, to make the numbers of samples of every PHQ-8 score be roughly the same. The other problem is that the sample size is small, the number of samples in training set is only 107. On account of such a small sample size, the number of features should also be small to avoid dimensionality and overfitting. So, the features to be used should be filtered and selected, since many of them are with very small values approaching to zeros.

A. Audio Features

Many researches have shown that the speech production of a human is very complex and as a result slight cognitive or physiological changes can produce acoustic changes [22]. Depressed speech is always associated with a wide range of prosodic, source, formant and spectral indicators. The audios provided by DAIC-WOZ dataset consist of a series of features extracted by the Covarep toolbox [23] at 10 ms intervals over the entire recorded wav files. This toolbox can capture both prosodic characteristics of the speaker, as well as voice quality. All audio features available for this dataset are listed in the Table II. One thing should be noticed is that the VUV(voiced/unvoiced) provides a flag (0, 1) to represent whether the current segment is voiced or unvoiced. If it is unvoiced, the other features are not utilized. In order to uncorrelated with speaker, the F0 was normalized to range from 0 to 1, and the deltas and delta-deltas were extracted for F0 and MCEPs.

We use the mean and standard deviation of the raw voice features during voice time. So, there are totally 252

 ISBN: 978-1-61208-675-0

different kinds of audio features for us to choose from. The raw recorded wav files of the virtual interview are provided, so we can extract other features for further analysis.

B. Video Features

Video features have been widely used for depression analysis, including facial subtle expression, body movement, gestures and facial muscle movements. Facial expressions can be an extremely powerful medium used to convey human overt emotional feedback. Girard et al. found that people with high level depression made fewer affiliative facial expression and more non-affiliative facial expression [24]. Facial Action Coding System (FACS) is a system to taxonomize human facial movements by their appearance on the face. Movements of individual facial muscles are encoded by FACS from slight different instant changes in facial appearance. By using FACS, it is possible to code nearly any anatomically possible facial expression, deconstructing it into the specific Action Units (AU) that produced the expression. It is a common standard to objectively describe facial expressions. The DAIC-WOZ database provides different types of video features based on the OpenFace framework [25], such as facial landmarks, HOG (Histogram of Oriented gradients) features, gaze direction estimate for both eyes, head pose and AUs (Action Units). In this paper, we use the mean and standard deviation of 20 key AUs over time. Thus, we get 40 video features for use.

C. Text Features

Clinical psychologist diagnose depression is mainly base on the language. We use speech features because the cognitive and physical changes associated with depression can lead to differences in speech. Linguistic features are similar to this, psychological and sociological theories suggest that depressed language can be characterized by specific linguistic features. The DAIC-WOZ database provides us the transcript file of the interview. Since the provided transcripts are human-machine conversations, we first extract human part. Then, we use Linguistic Inquiry and Word Count (LIWC) [26] software to count the frequency of word for each interview. The LIWC software is developed for providing an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals' verbal and written speech samples. With the help of LIWC, we extracted 93 text features. In order to get more text features of value, we use the depression related words list available online and the AFINN [27] toolbox. The depression related word list contains more than 200 words, such as "Abnormal" , "Alone"and "Anger". The AFINN toolbox would represent the valence of current text by comparing it to given word list with known sentiment labels. The output scores of AFINN is between minus five (negative) and plus five (positive). The mean, median, minimum, maximum and standard deviation of the output scores were used. So, we get five additional text features.

D. Feature Selection

Now, we already get the feature we want to use. We also should take the gender into consideration. In [28], Stratou et al. showed that gender plays an important role in the assessment of psychological conditions such as depression and PTSD, and a gender dependent approach significantly improves the performance over a gender agnostic one. The TABLE III shows the dimension of each feature category in the feature vector.

TABLE III. Dimension of each feature category

Feature Name	*Dimension*
COVAREP	252
Formants	10
AUs	40
LIWC	93
AFINN	5
Gender	1
Sum	401

We totally get 401 features to use. However, there are only one score need to be predicted for each example. What' s more, the number of the train set samples is only 107, among them the number of depression samples is 30. And only a small number of features are actually useful. So, it' s essential to small the number of features to avoid the problems of dimensionality, overfitting and shorter train times. First, we discard some features that always zero, such as HMPDM_0-3, and features with too much abnormal values, such as HMPDM_4. Next, we remove the feature with low variance according to the statistics of every feature. Then, we conduct the L1-based feature selection method. Linear models penslized with the L1 norm have sparse solutions, many of their coefficients are zero.When the goal is to reduce the dimensionality of the data, they can be used to select the non-zero coefficients. Finally, we select the best feature set.

III. Multi-Model Fusion

We use the PHQ-8 score as description of the depression severity. The distribution of the depression scores based on the train and development set is shown in Figure 3 [29]. As we can see from the figure, the data are imbalanced, which mean there are much more samples with low PHQ-8 scores than those with high PHQ-8 scores. It has been reported that imbalanced classes of data will have a great influence on the performance of machine algorithms. So, we conduct a random oversampling to make the number of samples of each PHQ-8 score is roughly the same. Because the available data set are too small, a 5-fold cross-validation has been used. That mean we divided development set into 5 folds. And we use one fold for testing and another 4 folds for training each time. We conduct Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (1)$$

to evaluate the performance of the multi-model. While the RMSE score is lower, the performance of the multi-model is better.

Random Forest (RF) is an ensemble learning method which fits multiple decision tree regression by random

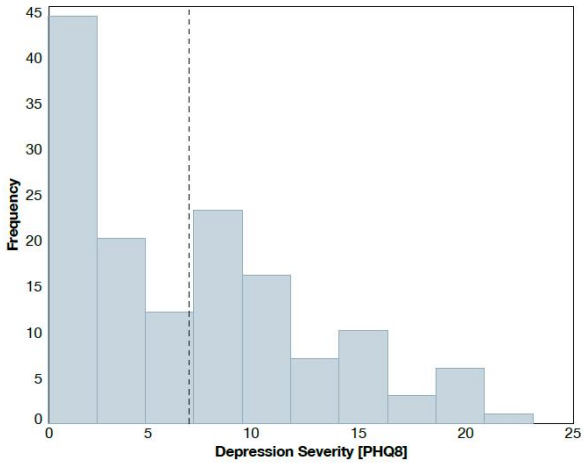

Figure 3. "Histogram of depression severity scores of training and development set"

selection of features and optimizes by bagging and aggregating the results. The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. Each base model is a sample decision tree. The main principle behind ensemble learning method is that a group of "weak learners" can come together to form a "strong learner". The random forest is of great advantages, such as fast training, high generalization capability and stable performance. Many similar researches related with depression severity prediction use random forest model. Support vector regression (SVR) uses the same principles as the support vector classification, with only a few differences. It just looks for and optimize the generalization bounds the same way as classification approach. In our work, we have tried a grid search for random forest regressor and support vector regression (linear kernel) as base regressor. However, the performance of random forest is better than support vector regression.

The framework of our proposed method is shown in Figure 4. As we mentioned above, gender is of great concern for depression severity detection. The performance of with or without gender will show in the next section. We firstly conduct a feature-level fusion, combining audio, video or text feature with gender. Then, three random forest regressors were constructed to calculate the regression value only depending on fused features. Finally, we implemented a decision-level fusion method. The linear opinion pool method was employed in the final decision-level multi-model fusion because of its simplicity. The fused result can be formulated as follow:

$$R_{final}(x) = \sum_{i=1}^{n} \alpha_i R_i(x) \tag{2}$$

where x stands for a test sample, R(x) stand for the result of ith random forest regressor, α_i is the corresponding weight which satisfies.

$$\sum_{i=1}^{n} \alpha_i = 1 \tag{3}$$

We use the optimal weights to fuse the regression value together to vote for a result that minimize error. The fusion result is shown in TABLE IV.

IV. Results and Discussion

We first needed to decide which base model to use. We have tried the random forest regressor and support vector regressor (linear kernel). The performance of this two regressor is provided in TABLE IV. In this experiment, we conducted very little optimization. Obviously, the performance of random forest is better than SVR in every feature. So, we selected the random forest regressor as base model. In the research of the other scientist found that fusing gender feature could achieve better performance. Therefore, we conducted a feature-level fusion, combining audio, video or text feature with gender. The performance of audio, video and text feature only and the performance with gender are provide in TABLE V. The weight of every base model is another problem we have to deal with. We first tried equal weight of every base model strategy. However, this strategy seemed that it could not reach the optimal result. So, we changed the weight of every base model according to their performance. Finally, we chose the weight 0.4 for audio, 0.4 for video, 0.2 for text.

TABLE IV. Performance of random forest and SVR

Features	*Base Model*	*RMSE*
Audio	Random forest	7.16
Video	Random forest	7.48
Text	Random forest	7.13
Audio	SVR	8.16
Video	SVR	7.81
Text	SVR	7.35

TABLE V. Performance include gender or not

Features	*RMSE*
Audio only	6.53
Video only	6.92
Text only	6.70
Audio&Gender	6.39
Video&Gender	6.89
Text&Gender	6.58

The AVEC 2017 organizer provided the baseline of Depression Severity Assessment Challenge (DSC). They computed the depression severity baseline using random forest regressor (number of trees: 10, 20, 50, 100, 200). In their experiment, the best performing random forest has trees = 10. Fusion of audio and video modalities was performed by averaging the regression outputs of the unimodal random forest regressors. The performance of RMSE is provided in TABLE VI.

We take the optimizing strategy mentioned above, the result of proposed multi-model fusion framework is displayed in TABLE VII. All these results are on the

 ISBN: 978-1-61208-675-0

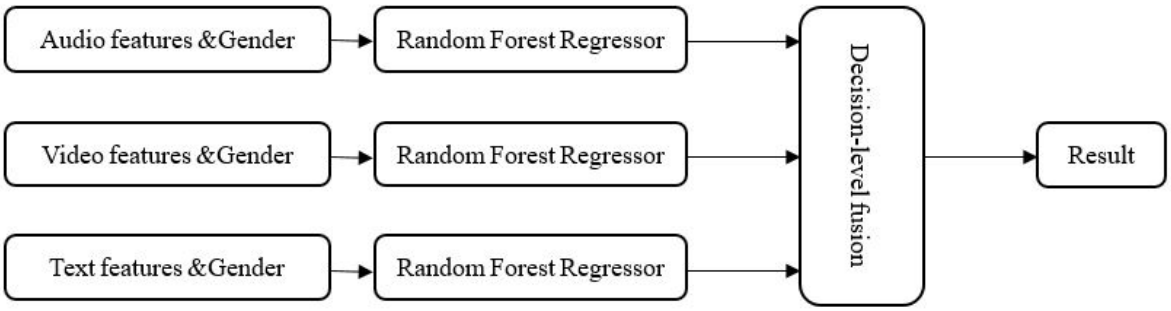

Figure 4. Framework of multi-model fusion

TABLE VI. Baseline of DSC challenge

Modality	*RMSE*
Audio	6.74
Video	7.13
Audio-Video	6.62

developing set. The result obtained were better than the baseline result.

TABLE VII. Performance of multi-model fusion framework

Features	*RMSE*
Audio&Gender	5.84
Video&Gender	5.89
Text&Gender	6.43
Multi-model fusion	5.98

V. Conclusion

Depression is a widespread mental disorder now. The accurate detection of it is still hard by manual. In this paper, we proposed a random forest regressor based multi-model fusion framework on audio, video and text features for prediction of PHQ-8 scores which ranges from 0 to 25. The weights of base models depend on their own performance. Our experiments suggested that the proposed approach performs better than DSC baseline. Since the result of multi-model fusion is very promising, future work would be devoted toward using more complex base regressor, such as neural networks. With more relevant features containing more useful information, the overall performance could also be improved. The features of body and head movement could be the next experiment we would like to try. And the more effective fusion strategy is another future work direction. The linear opinion pool method seemed could not reach the best performance of every base regressor. We may use some method to grade the performance of base model in every sample. Then, adjusting the weights dynamically to reach the best potential.

Acknowledgment

This work was supported by the National Key Research and Discovery Plan No. 2017YFF0210903, and the National Natural Science Foundation of China (grant No. 11433002).

References

[1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," PLoS medicine, vol. 3, no. 11, 2006, p. e442.

[2] WHO et al., "Depression and other common mental disorders: global health estimates," World Health Organization, 2017.

[3] V. Kral, "The relationship between senile dementia (alzheimer type) and depression," The Canadian Journal of Psychiatry, vol. 28, no. 4, 1983, pp. 304–306.

[4] R. M. Carney, K. E. Freedland, G. E. Miller, and A. S. Jaffe, "Depression as a risk factor for cardiac mortality and morbidity: a review of potential mechanisms," Journal of psychosomatic research, vol. 53, no. 4, 2002, pp. 897–902.

[5] P. E. Greenberg, A.-A. Fournier, T. Sisitsky, C. T. Pike, and R. C. Kessler, "The economic burden of adults with major depressive disorder in the united states (2005 and 2010)," The Journal of clinical psychiatry, vol. 76, no. 2, 2015, pp. 155–162.

[6] K. Hawton, C. C. i Comabella, C. Haw, and K. Saunders, "Risk factors for suicide in individuals with depression: a systematic review," Journal of affective disorders, vol. 147, no. 1, 2013, pp. 17–28.

[7] J. Gratch et al., "The distress analysis interview corpus of human and computer interviews." in LREC. Citeseer, 2014, pp. 3123–3128.

[8] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. 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, 2009, pp. 163–173.

[9] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in Twelfth Annual Conference of the International Speech Communication Association, 2011, pp. 2997–3000.

[10] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd." in Interspeech, 2013, pp. 847–851.

[11] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," IEEE Transactions on Affective Computing, vol. 4, no. 2, 2013, pp. 142–150.

[12] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," Speech Communication, vol. 75, 2015, pp. 27–49.

[13] J. F. Cohn et al., "Detecting depression from facial actions and vocal prosody," in Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. IEEE, 2009, pp. 1–7.

[14] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences," in Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE, 2013, pp. 147–152.

[15] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz,

 ISBN: 978-1-61208-675-0

"Predicting depression via social media." ICWSM, vol. 13, 2013, pp. 1–10.

[16] M. Valstar et al., "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.

[17] M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk, "Model fusion for multimodal depression classification and level detection," in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. ACM, 2014, pp. 57–63.

[18] L. Low, "Detection of clinical depression in adolescents' using acoustic speech analysis," researchbank.rmit.edu.au, 2011.

[19] S. Dham, A. Sharma, and A. Dhall, "Depression scale recognition from audio, visual and text analysis," arXiv preprint arXiv:1709.05865, 2017.

[20] M. R. Morales and R. Levitan, "Mitigating confounding factors in depression detection using an unsupervised clustering approach," in CHI 2016 Computing and Mental Health Workshop, 2016.

[21] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 39, no. 2, 2009, pp. 539–550.

[22] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: a multimodal approach," in Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. ACM, 2013, pp. 11–20.

[23] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep ¡a a collaborative voice analysis repository for speech technologies," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2014, pp. 960–964.

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

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

[26] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[27] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," arXiv preprint arXiv:1103.2903, 2011.

[28] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender," Journal on Multimodal User Interfaces, vol. 9, no. 1, 2014, pp. 17–29.

[29] F. Ringeval et al., "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, 2017, pp. 3–9.

 ISBN: 978-1-61208-675-0