Gender-aware Estimation of Depression Severity Level in a Multimodal Setting

Syed Arbaaz Qureshi
Research Fellow
Microsoft Research India
arbaaz.qureshi29@gmail.com

Gaël Dias
Department of Computer Science
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
14000, Caen, France
gael.dias@unicaen.fr

Sriparna Saha
Department of Computer Science and Engineering
Indian Institute of Technology, Patna
Patna, India
sriparna@iitp.ac.in

Mohammed Hasanuzzaman
Department of Computer Science
Cork Institute of Technology
Cork, Ireland
mohammed.hasanuzzaman@cit.ie

***Abstract*—Depression is a severe psychological disorder that is experienced by a significant number of individuals across the globe. It greatly changes the way one thinks, triggering a constant decline in mood. Studies have shown that gender can act as a good indicator of depression. In this paper, we analyse the effects of gender information in the estimation of depression. We have carried out different experiments on the benchmark data set named Distress Analysis Interview Corpus - a Wizard of Oz (DAIC-WOZ). Concretely, we discovered that a) gender information substantially improves the performance of depression severity estimation, and b) adversarially learning to predict the depression score distributed by gender improves the performance of depression severity estimation.**

***Index Terms*—Depression, Multitask learning, Gender**

Depression is a widespread and severe psychiatric disorder that has a negative effect on how one perceives things. It is marked by constant sorrow, lack of motivation and an inability to pursue tasks that are typically loved by one. It is one of the largest worldwide causes of ill health. Over 300 million people are reportedly suffering from depression, a rise of over 18% from 2005 to 2015. [1]

On an average, depression lasts from 4 to 8 months. Insomnia, emaciation, exhaustion, feelings of worthlessness, opioid or alcohol misuse, and diminished capacity to remember, focus and undertake decisions are some of the signs and side-effects of depression. It may also be characterized by feelings of death, suicide and attempted suicide in serious situations. The annual number of deaths due to depression is, unfortunately, on the rise. [2]

It is not fully clear what causes depression, and there may not be one particular source. It is possible that major depressive disorders are due to complex variations of variables such as the sufferer's genes, psychology, and social climate. People who have witnessed major traumatizing events like a family member's or friend's death, those with personality disorders like the incapability to cope with rejection and disappointment, patients with past history of severe depression, and people with a history of child abuse are at increased risk of depression [1].

The diagnosis of depression is a complex topic since many of its manifestations are covert. As depressed individuals don't socialize often, it becomes even harder to identify. The sufferer is evaluated on regular questionnaires, for the proper diagnosis of depression. Various methods for diagnosing depression have been researched on, in the literature. A few of them are the Hamilton Depression Rating Scale (HDRS), the Personal Health Questionnaire Depression Scale (PHQ), the Center for Epidemiologic Studies Depression Scale (CES-D), the Montgomery and Asberg Depression Rating Scale (MADRS) [3], the Beck Depression Inventory (BDI), and the Hospital Anxiety and Depression Scale (HADS). In particular, the eight-item PHQ-8 [2] is agreed upon as a valid diagnostic and severity measure for depressive disorders in many clinical studies [3].

The increasingly rising global prevalence of depression and mental disease serves as a catalyst for the development of more complex, customized and automatic solutions that can diagnose it. Affective computing is one branch of study that relies on extracting data to quantify human emotions via facial expressions, vocal sounds and body gestures. A significant business purpose of affective computing is to develop human-computer interfaces that can perceive and respond effectively to the emotional state of the participant. As a result, affective computation methods have been used for predictive depression diagnosis [4], [5]. However, seldom initiatives have attempted to tackle the impact of gender on the automatic diagnosis of depression [7], [8]. In this paper, we propose to study different

[1]A figure published by the World Health Organisation, available at https://bit.ly/2rsqQoP.

[2]A finding by Max Roser and Hannah Ritchie, available at https://bit.ly/2mnyVZ6.

[3]Recommandation of the French Haute Autorité de la Santé available at https://bit.ly/2EaOs92.

gender-aware models in multimodal settings, as clinical studies have evidenced the differences in experiencing depression between male and female [9].

I. Related work

A significant number of computer science research projects have been proposed over the past few years to deal with mental health problems [10], [11], and there has been a resurgence of activity in the automated diagnosis of depressive disorders. A few initial efforts have focused on the use of appropriate descriptors and features, that could be leveraged using learning systems. Scherer et al [12] explore the capacity of descriptors from non-verbal behaviour, to point out indicators of psychiatric conditions like depression. In particular, the four suggested descriptors that can be derived spontaneously via visual cues are: downward angling of the head, smile length and strength, eye gaze direction, and self-touching. Chatterjee et al [13] study the role of various context-based descriptors of heart rate variability to assess the psychological health of an individual. Cummins et al [14] concentrate on how distress and suicidal behavior are indicated by certain paralinguistic speech characteristics (prosodic characteristics, source characteristics, formant characteristics, spectral characteristics) and on how to use this knowledge in prediction and classification systems. Morales et al [15] believe that by designing features that capture syntactic structure and semantic information, researchers can look beyond the acoustic properties of expression. Within this context, Wolohan et al [16] state that the overall performance of the classification indicates linguistic model versions are relatively stable and also reliable for the diagnosis and monitoring of depression. Another interesting branch of study using textual features include the analysis of information from social media [17], [18]. Some of the works use specific corpora fine-tuned for these tasks [19].

Another potentially fruitful trend seeks to leverage all modalities into a single learning paradigm and is often referred to as multimodal detection of depression [33]. Within this context, many active research studies have been performed. Dibeklioglu et al. [21] compare, individually and in combination, the facial movement patterns, head movement nuances, and voice prosody, and show that techniques using multimodal machine learning result in the most effective detection. Yang et al. [22], in the DAIC-WOZ (Distress Analysis Interview Corpus - a Wizard of Oz) benchmark dataset, achieve acceptable performance to predict the PHQ-8 rating by integrating acoustic, visual and textual features, using decision-tree classifiers. More recently, Morales et al. [5], [33] worked on a detailed review of fusion techniques using SVM (early, late and hybrid) for the detection of depression integrating acoustic, visual and textual characteristics (especially syntactic). Concretely, they demonstrate that the syntax-informed fusion strategy is capable of exploiting syntactic data to target insightful nuances of speech data, but the overall findings appear to imply that this finding is not statistically confirmed. Qureshi et al. [34] undertook a multimodal-multitask approach to simultaneously predict the depression level, and classify the severity of depression, and found that fusion of modalities, and the concurrent learning that happens over the other related task, both improve the estimation of depression severity.

Not a lot of work has been done on how depression is dependent on gender, and how it is different in males and females. Within this context, Conklin et al. [23] analyse the correlation between chronic sleep deprivation and the symptoms of depression among adolescents, from a gender perspective, where they find that persistent sleep deprivation is correlated with steady increase in depression scores among young women, but no such pattern was found in young men. More recently, Lopez et al. [24] find that in general, automatic depression detection, when done for each gender separately, leads to its better estimation. They even de-identify speech, and find that the difference is not a lot for original speech, but it is considerable when using with de-identified speech. [6] reviews several works in psychological research on the difference in gender, in depression. They state that by the middle of adolescence, females are about twice as likely to be diagnosed with depression and exhibit twice as many depressive symptoms as males, and this trend may continue till they are atleast 55 years old.

II. DAIC-WOZ depression dataset

The depression dataset of DAIC-WOZ [4] is part of a broader corpus, the Distress Analysis Interview Corpus [25], which involves psychiatric interviews meant to benefit the assessment of psychological distress disorders like anxiety, depression, and PTSD (Post Traumatic Stress Disorder). These recordings were gathered as part of a broader initiative to build a software agent which interviews individuals and recognizes verbal and non-verbal mental disorder indicators. The information collected includes video and audio clips, and detailed questionnaire answers from the interviews done by a virtual interviewer named Ellie, which is controlled by a human. For a range of verbal and non-verbal attributes, the data was transcribed and annotated.

The dataset contains 189 interview sessions. Out of these sessions, out of which 102 are males, and 87 are females. In the training split, 63 out of 107 are males, and 44 are females; in the validation split, 16 out of 35 are males, and 19 are females; and in the test split, 23 out of 47 are males, and 24 are females. We discard a few interviews, since a few were not complete, and had discontinuities. A unique ID is assigned to every interview, allowing it to be identified. Each instance of the dataset is comprised of the session's audio clip, the interviewee's cartesian coordinates of sixty eight **facial landmarks**, the HoG (Histogram of oriented Gradients) features of his/her face, the **eye gaze** and **head pose** features extracted using OpenFace [26], through out the whole length of the interview, the continuous **facial action units** of the face, extracted using CERT [27], the facial action coding software, the **formant** and **COVAREP** features of the voice, extracted using COVAREP [28], and the whole **transcript** of

[4] http://dcapswoz.ict.usc.edu/.

the session. All the extracted features are temporal by nature. For the ease of discussion, we cluster the modalities into three major groups: a) the visual or ocular modalities, which is comprised of the 68 landmarks on the face, the eye-gaze and head-pose features, and action units of the face, b) the auditory or the acoustic modalities, comprised of the formant and the COVAREP features, and c) the linguistic or the textual modality, consisting of the transcript of the interview session. This grouping is done for a better explanation of our analysis, and holds no relevance in the experiments described in this paper.

We begin with the ocular modalities. Every time-step of the facial landmark modality is comprised of the *time-stamp*, *reliance*, *detection success indicator*, and the (X, Y, Z) cartesian coordinates of all the above mentioned facial landmarks. Every time-step of the head-pose data is made up of *time stamp*, *reliance*, *detection success indicator*, (R_x, R_y, R_z), and (T_x, T_y, T_z). Here, (R_x, R_y, R_z) are the coordinates of head rotation in Radians, and (T_x, T_y, T_z) are the coordinates of head position in millimetres. The eye-gaze data time-steps is made up of *time stamp*, *reliance*, *detection success indicator*, (x_0, y_0, z_0), (x_1, y_1, z_1), (x_{h0}, y_{h0}, z_{h0}), (x_{h1}, y_{h1} and z_{h1}). Four vectors are used to represent the eye-gaze. The direction of the eye-gaze is described by two vectors (x_0, y_0, z_0) and (x_1, y_1, z_1). The gaze is also described in the coordinate space of the head, using the remaining vectors (x_{h0}, y_{h0}, z_{h0}) and (x_{h1}, y_{h1}, z_{h1}) (when the eyes roll upwards, these two vectors specify 'up', even though the head is not facing the camera). Each time step of the action units modality is made up of *time stamp*, *reliance*, *detection success indicator*, and some numbers specifying the appropriate action unit of the face. The frequency at which these features were recorded is 30 Hertz.

In the auditory group of modalities, every time-step of the formant and COVAREP data contains 74 and 5 features respectively. These are different features of the interviewee's or the virtual assistant's voice. The frequency at which both these data are recorded is 100 Hertz. There's one common attribure in both these modalities, an indicator named *VUV* (Voiced/Unvoiced), which specifies if that particular bit is voiced or not. According to the manual of the DAIC-WOZ dataset, we are advised to not use those time steps where the *VUV* indicator is 0.

The interview-transcript, that happens to be the only modality of the linguistic group, is made up of the words uttered by the virtual assistant Ellie, and the interviewee. Each time-step is made up of *start time*, the starting time of each sentence of the speaker, *stop time*, the time-stamp of the instant when the speaker finishes uttering, *speaker*, an indicator specifying who the speaker is, whether it is the virtual assistant, or the interviewee, and *value*, the words uttered by the speaker, verbatim.

We use the training, validation and test split specifications provided with the dataset. Both the training and validation splits have the unique identifications of the interviews, the binary values of PHQ-8 (which is 1 when it is greater than 10, 0 otherwise), the PHQ-8 score values, the gender of the interviewee, and individual answers to each of the questions of the PHQ-8 questionnaire. The test split is comprised of the interview identification numbers, and the gender of the interviewee.

III. Data pre-processing

We have used different pre-processing techniques for different modalities of data. They are listed in the subsections to come.

A. Visual modalities pre-processing

In the coordinates of the 68 landmarks of face, we scale the Z-coordinate by first removing its average value (calculated over all the time steps), from all the time-steps. This ensures that there's no bias along the Z axis, since the person can be sitting far away from ,or near to the camera. We then scale the coordinates in a way that the mean distance of each of the point from the origin of the coordinate system is 1. Next, we compute the Euclidean distance between all the possible 2278 point pairs, and append this to the scaled coordinates of the facial landmarks. This gives us a vector of length 2482, for every time step.

When it comes to the head-pose modality, we re-size (T_x, T_y, T_z) by dividing all of them by 100. Since 30 Hertz makes it a lot of time steps, we down-size all the visual data modalities to 5Hz. We have adopted a zero-tolerance strategy towards time steps with no tracking success (where the success indicator is 0), and discard all such time steps. We take this decision, as we are wary of introducing any artefacts into the attribute space of the modality. Since different interviews go on for different time lengths, we pad each of the ocular modality time series data with zeroes along the time axis, to have a uniform length of 10000 time steps for each of the data point. We don't perform any further pre-processing on the eye-gaze data, as all the values fall between -1 and 1. We don't perform pre-processing on the facial action units too, since scaling has already been done to these features, and they fall between 0 and 1.

B. Acoustic modalities pre-processing

We follow the same zero-tolerance strategy, and throw away all the time steps where the Voiced/Unvoiced indicator values are 0. Like in any typical interview, the virtual assistant and the interviewee speak in turns. Since we look for the formant and COVAREP features of the interviewee only (and not of the virtual assistant), we segregate them using the start and stop time values specified in the interview transcription, and use those of the interviewee, to make predictions. Since speaking anything meaningful in less than a second is seldom, we discard those time steps where the interviewee has spoken for less than one second. We pad these features with zero vectors, along the temporal axis, to get a uniform length of 80000 and 120000 time steps respectively for COVAREP and formant modalities.

C. Text modality pre-processing

We only collect the utterances made by the interviewee, and analyse them in an increasing order, sorted with respect to their start times. Since colloquial language has been used by many of the interviewees, we make the English utterances formal by substituting the contractions (like "isn't", "ain't") with the corresponding full words ("is not", and "are not"). Then, we encode each sentence of the interviewee into a 512 dimensional vector, which are extracted from a pre-trained Universal Sentence Encoder [29], after feeding the sentences to it. Again, We pad these time steps with zero vectors on the time axis, to get a uniform length of 400 time-steps for all the interviewees.

IV. Methodology

To test our hypothesis, we designed and experimented with 5 different networks. They are:

- Depression estimation without gender information (***Gen**$_{less}$*)
- Depression estimation with concatenated gender information (***Gen**$_{concat}$*)
- Multitask prediction of depression level and gender (***Gen**$_{pred}$*)
- Multitask prediction of depression level in males and females separately, using shared-private multitask network [35] (***Gen**$_{SP}$*)
- Multitask prediction of depression level in males and females separately, using adversarial shared-private multitask network [35] (***Gen**$_{ASP}$*)

We group the explanation of these 5 networks into 2 groups. We proceed by explaining ***Gen**$_{less}$*, ***Gen**$_{concat}$* and ***Gen**$_{pred}$* first, then ***Gen**$_{SP}$* and ***Gen**$_{ASP}$* next.

A. *Gen*$_{less}$, *Gen*$_{concat}$ and *Gen*$_{pred}$

The ***Gen**$_{concat}$* and ***Gen**$_{pred}$* networks derive their architecture from ***Gen**$_{less}$* with a few modifications. So we describe the architecture of ***Gen**$_{less}$* in detail, and state the necessary modifications in order to derive the architectures of ***Gen**$_{concat}$* and ***Gen**$_{pred}$*. The ***Gen**$_{less}$* consists of three main sub-networks. They are 1) modality encoding sub-network, 2) the modality fusion sub-network and 3) the PHQ-8 score estimator.

1) Modality encoding sub-network: Since all the modalities of the dataset are time-series data, we use LSTMs to process them to predict the PHQ-8 score. Concretely, we pass the features of all the time-steps to a standard LSTM network (`LSTM`) [36]. We take the state vector from the last time step, and pass it to a fully connected layer (`FC`).

In the case of ***Gen**$_{concat}$*, we concatenate the gender binary value (a binary value representing male/female) before passing it to the fully connected layer. We then feed the output of this fully connected layer to a linear regression unit, to get the PHQ-8 score. There are 7 modalities in our dataset. We train 7 different modality encoders separately to predict the PHQ-8 score.

In the case of ***Gen**$_{pred}$*, along with PHQ-8 score regression, we predict the gender of the participant using the output of the fully connected layer `FC`, by passing it to a binary classification unit (i.e. a fully-shared multitask architecture).

We train these networks using Adam optimizer [37] on mean squared error, and use network which performs the best on the validation set. For ***Gen**$_{pred}$*, we take a weighted average of mean squared error (for PHQ-8 prediction) and the binary cross-entropy error (for gender classification) as the loss function, and optimize on this using the same optimizer.

2) Modality fusion sub-network: In addition to using individual modalities to predict the PHQ-8 score, we employ a multi-modal fusion technique to use information from all modalities, and predict the PHQ-8 score. Our technique for multi-modal fusion resembles with the one used by Qureshi et al [38]. We call the output from the fully connected layer `FC` of a modality encoder as the modality encoding of that modality. Note that the modality encodings are vectors of different lengths, as we use different hyper-parameters (number of units in `LSTM` and in `FC`) for the encoders of different modalities. In order to make them of the same length (which is 128, in our setup), we pass these encodings to 7 fully-connected layers (all these fully connected layers have the same number of units). Now, we take the weighted average of these 7 modified modality encodings (as a dot product of the horizontally-concatenated modality encodings $\mathbf{H}$ and the weight vector α), to get the fused encoding $\mathbf{F}$. The weight vector α is learnable parameters. For learning it, we first concatenate the 7 modified modality encodings, and pass it to a 1-layer neural network whose output layer is a 7-unit softmax layer. We use softmax layer as its outputs. We use this specific activation function because the outputs are values which sum to 1, which in our use case represent the weightage or importance of a given modality, in the fused representation. We observe that the textual modality is usually given high weight, which is suggestive of the fact that it plays the most important role among all the modalities, for estimating the PHQ-8 score. See Figure 1 for a visual representation of the modality fusion network. Note that this sub-network is the same for all of the ***Gen**$_{less}$*, ***Gen**$_{concat}$* and ***Gen**$_{pred}$* networks. Also note that we have fixed the hyperparameter values after performing an extensive grid search over the hyperparameter space, and chose those which gave the best performance on the validation set.

3) PHQ-8 score estimator: This is a relatively straight forward neural network with 1 hidden layer, conditioned on the fused encoding $\mathbf{F}$. We feed the fused encoding $\mathbf{F}$ from the modality fusion sub-network to a fully connected layer. The output of this layer is then fed to a linear regression unit, which outputs the predicted PHQ-8 score. This sub-network too is the same for all three ***Gen**$_{less}$*, ***Gen**$_{concat}$* and ***Gen**$_{pred}$* networks.

B. *Gen*$_{SP}$ and *Gen*$_{ASP}$

We use multitask learning approach at individual modality level only. We combine the individual modalities to estimate the PHQ-8 score, and for realising this, we use the same modality fusion sub-network and PHQ-8 score estimator as described in the previous sections.

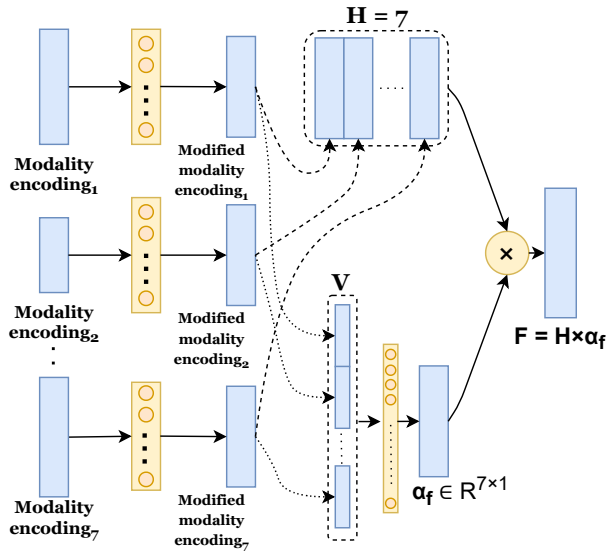

Fig. 1. Modality fusion sub-network. The blue boxes represent vectors (either the input to, or the output from the neural network layers), and the yellow boxes represent the neural network layers.

Gen$_{SP}$ and ***Gen$_{ASP}$*** are multitask learning networks. The 2 tasks in our context are depression detection for males and females. The intuition behind the experiments with these two architectures is that although there might be many commonalities in detecting depression in males and females, we believe that there may be some differences, and that simultaneously learning to detect depression for these two genders may improve the performance on these two tasks.

Gen$_{SP}$ is a general shared-private multitask learning architecture. We describe it briefly in this section. The ideas behind ***Gen$_{ASP}$*** have been derived from Liu et al [35], and have been used by Qureshi et al [39] for concurrently estimating depression and emotion intensity.

1) ***Gen$_{SP}$****:* The ***Gen$_{SP}$*** multitask neural network is made up of 3 LSTM layers - 2 networks specific to the tasks (the tasks are: depression detection in males, and in females), and 1 network for the combined (shared) task. Each of these layers have the same number of neurons. In particular, the input belonging to a particular task is passed on to the shared LSTM layer, and the LSTM layer corresponding to that task. The output vectors from these layers are combined to get a vector. This combination is done using a technique similar to the modality fusion sub-network which is discussed briefly in the following paragraph. The combined vector is then passed on to a dense layer, whose output is in turn passed on to the output layer specific to the task.

For fusing the outputs from the shared and task-specific LSTMS in the fusion sub-network, we append them vertically, and feed this concatenation to a dense layer, whose output is in turn passed on to a softmax layer. This softmax layer gives two numbers: α_{shared} and α_{task}. These numbers are the weights for the outputs of the shared LSTM layer and the task-specific LSTM layer respectively, in calculating the final output, where we take the weighted average of the two LSTM layer outputs.

We were particular about including the attention mechanism, so that we could better comprehend the importance of the shared and task-specific features, while making a prediction. For predicting the PHQ-8 score, if it is the case that task-specific embeddings provide less information than the shared embeddings, then α_{task} would be lesser than 0.5, and α_{shared} would have a greater value than 0.5. This mechanism allows the neural network to learn the contributions of the task-specific and shared embeddings, for the final task of estimating depression level. In addition, it is a well founded fact that neural networks with attention usually perform well, when compared to their counterpart with no attention [30].

The network architecture described above, which is also shown in Figure 2 lays out an infrastructure which has distinct vector spaces for shared features and the features specific to the task. But it is found that this architecture too has it's own lacunae. The feature space of the shared task could have some needless features specific to the task, and some shared features may creep into the feature space specific to the task. Such a network is prone to suffer from feature redundancy, as depicted on the right half of figure 2. To circumvent this issue, we test the ***Gen$_{ASP}$***. It is discussed in the coming section.

2) ***Gen$_{ASP}$****:* We take inspiration from the findings of [31], [32] and develop a modified architecture with two changes. Just like in ***Gen$_{SP}$***, there are 3 LSTM layers in the ***Gen$_{ASP}$*** neural network, 2 specific to the task and 1 shared. And just like the case in ***Gen$_{SP}$***, all the three layers have the same number of neurons. A task's input vector is passed on to both the shared and the task-specific LSTM layers. Using the same attention fusion mechanism as described in the previous section, the output vectors from the shared and the task-specific layers are fused.

However, unlike ***Gen$_{SP}$***, before the fusion takes place, the output of the shared LSTM layer is fed to a dense layer with softmax activation. The output of this dense layer is the task label prediction (if t_1 and t_2 are two tasks at hand, the label for t_1 is [1, 0], and the label for t_2 is [0, 1]). Observe that the shared-LSTM layer and dense layer with softmax activation act as the generator-discriminator networks of an adversarial network, where the dense softmax layer is the discriminator and shared-LSTM layer, the generator. This makes sure that the shared space contains only the shared features.

But how do we make sure that the shared features don't creep into the task-specific vector spaces? We have the L_{diff} to handle this. The value of L_{diff} objective function is computed using the output vectors of the shared and the task-specific LSTM layers. This acts as an orthogonality constraint, making sure that the output from the task-specific layer is as orthogonal to the shared LSTM layer output as possible. We have used a slightly different definition for L_{diff}; it is not the one used by [31] or [32]. L_{diff} is defined in Equation 1. Here, $\|v\|_1$ is the L_1 norm of v, H and S are two matrices whose rows are the output vectors from all the time-steps of the task-specific and the shared LSTM layers respectively. m, n are the row and column dimensions of $H^\top S$ respectively. We chose this

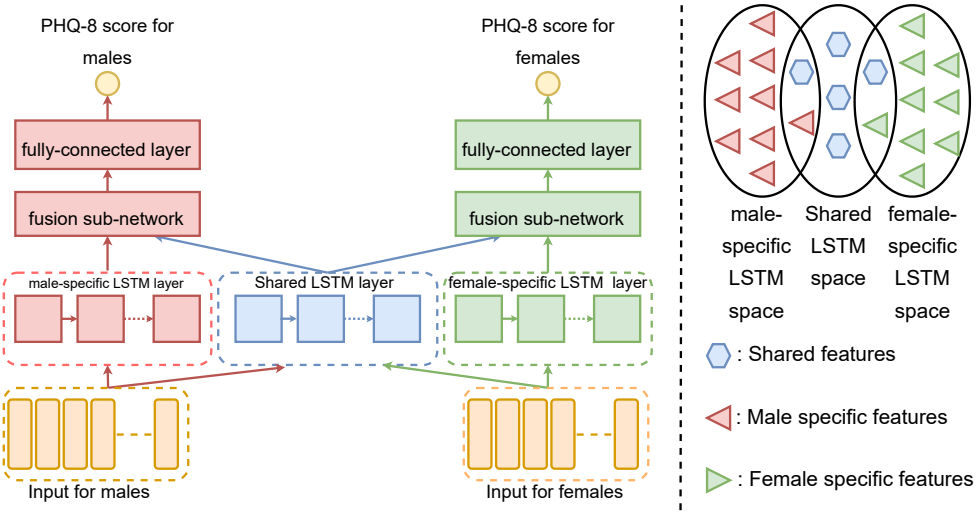

Fig. 2. ***Gen**$_{SP}$*. In the network, the boxes in yellow represent either inputs or outputs. The boxes in red, blue and green represent layers of the network.

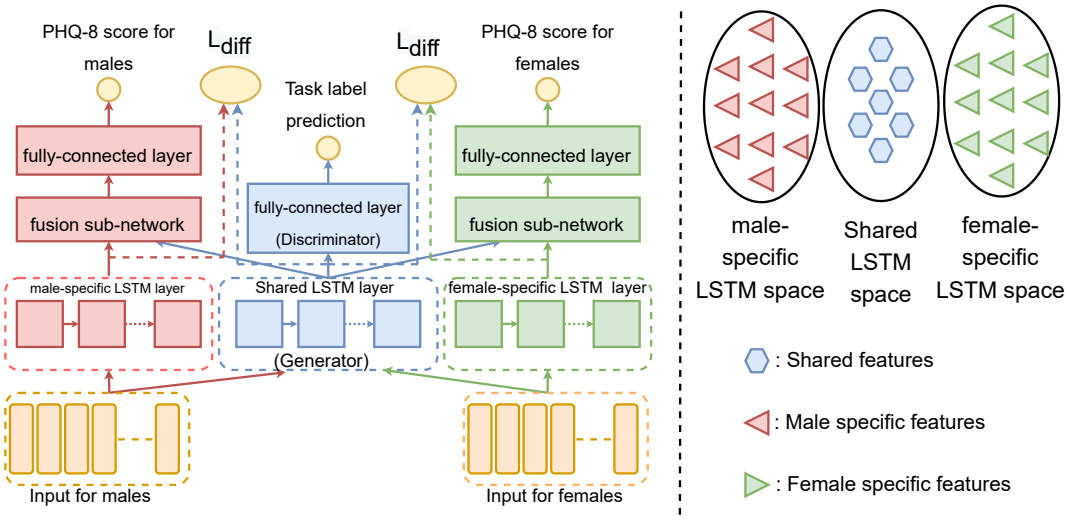

Fig. 3. ***Gen**$_{ASP}$*. In the network, the boxes in yellow represent either inputs or outputs. The boxes in red, blue and green represent layers of the network.

variation of L_{diff} after thoroughly experimenting with different alternatives. The full network diagram is shown in figure 3.

$$L_{\text{diff}} = \frac{\|H^\top S\|_1}{m \times n} \tag{1}$$

This architecture makes sure that the shared and the task-specific spaces are as distinct as possible, as seen in the right half of Figure 3. As mentioned earlier, the addition of the adversarial network (shared LSTM layer - dense softmax layer pair) discards the chance of features specific to a task crawling into the shared vector space. And the orthogonality constraint makes sure that these two spaces are as orthogonal as possible (they are fully orthogonal when L_{diff} is 0). This implies that the task-specific vector space would not have any of the shared features, as these two spaces should be orthogonal. Keep in mind that if the two tasks are very similar and correlated, ***Gen**$_{ASP}$* is bound to perform badly, as it would be challenging for the shared-LSTM layer (the generator) to generate an encoding that can truly fool the discriminator and make it classify poorly.

V. Results and Discussion

Results are presented in Table I. The first observation that we make from the results is that gender-aware models (***Gen**$_{concat}$*, ***Gen**$_{pred}$*, ***Gen**$_{SP}$* and ***Gen**$_{ASP}$*) tend to perform better on estimation of depression that the gender-unaware model (***Gen**$_{less}$*). For each of the 7 modalities, we obtain better mean squared error (MSE) values by using one of ***Gen**$_{concat}$*, ***Gen**$_{pred}$*, ***Gen**$_{SP}$* and ***Gen**$_{ASP}$*. This is a strong corroboration to the hypothesis that gender information substantially improves the performance of depression severity estimation.

TABLE I
RESULTS OF ALL THE 5 MODELS. MSE: MEAN SQUARED ERROR. MAE: MEAN ABSOLUTE ERROR.

Models	**Evaluation Metrics**									
	Gen_{less}		Gen_{concat}		Gen_{pred}		Gen_{SP}		Gen_{ASP}	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
COVAREP	44.98	5.32	44.59	5.16	**43.05**	5.14	43.70	**5.11**	44.13	5.14
Formant	43.11	5.48	42.21	5.50	42.29	5.54	42.53	5.56	**41.96**	**5.19**
Facial action units	42.32	5.51	41.97	**5.13**	**41.06**	5.47	41.90	5.15	41.97	5.21
Eye gaze	47.26	5.57	47.04	5.62	46.05	5.75	48.01	5.72	**44.41**	**5.23**
Facial landmarks	52.82	6.21	50.72	6.06	52.45	5.93	47.16	5.87	**45.13**	**5.51**
Head pose	48.99	5.78	47.31	5.74	46.92	5.76	46.56	5.54	**44.29**	**5.40**
Text	23.82	**3.78**	23.28	3.87	**23.12**	3.87	24.12	4.10	24.02	4.09
Multimodal	24.12	3.74	**20.06**	3.50	20.56	3.50	21.01	3.51	22.25	**3.49**

A simple concatenation of gender binary value does not seem to help in improving the estimation of depression at individual modalities level. Indeed, Gen_{concat} does not have the best MSE values for any of the 7 individual modalities, and only shows improved results for the mean average error (MAE) within a unique case. However, Gen_{concat} seems to be the best performing model in terms of MSE when all modalities are merged together (MSE: 20.06) with very close values of MSE, compared to the second best (Gen_{pred}, MSE: 20.56). This is an anomaly, and we attribute this deviation in trend to the less amount of data that we have. However, the general trend that follows in all other modalities is that concatenation of gender binary can only help so much.

We also observe that a multitask learning approach on depression estimation and gender detection helps in improving the performance on the first task. For COVAREP, facial action units, and text modalities, Gen_{pred} has the lowest mean squared error (43.05, 41.06, and 23.12 respectively), and for the multimodal model, it has the second lowest mean squared error. We do not analyse if depression estimation helps the performance on gender classification; it is out of the scope of our work.

Among Gen_{SP} and Gen_{ASP}, the latter has the lowest mean squared error for 4 out of 7 modalities. It is a strong indication that detecting depression in men is different than doing the same for women, when using Formant, eye gaze, facial landmarks or head pose features; women tend to display a different head movement, eye gaze turns and facial expression changes than men, in the context of depression. No strong conclusion can be made for the other three modalities, where Gen_{pred} shows the least mean squared error.

The text modality is the best marker of depression, as the mean squared and mean absolute errors are consistently lower than all other modalities. However, we observe that Gen_{SP} and Gen_{ASP} both have higher mean squared error values (24.12 and 24.02 respectively), when compared to Gen_{pred} (23.28), hinting that both the genders may use sentences with similar meanings, in the context of depression. We also find that the multimodal combination of all the modalities gives the best mean squared error (which happens for Gen_{concat}, MSE: 20.06), further corroborating the findings of Qureshi et al. [38].

VI. CONCLUSION

We hypothesize that gender information may play an important role in the estimation of depression. To verify the hypothesis, we designed 1 gender-unaware model (Gen_{less}), and 4 gender-aware models (Gen_{concat}, Gen_{pred}, Gen_{SP}, Gen_{ASP}), and trained them on the task of depression estimation. From the results, we find that indeed gender information improves the performance of depression estimation, although a simple concatenation of gender binary might not be enough at modality level. We also find that for 3 of the modalities (COVAREP, facial action units, text), simultaneously learning to predict gender makes depression estimation more accurate. Lastly, depression estimation using the other 4 modalities (Formants, facial landmarks, eye gaze and head pose) is different for males and females.

ACKNOWLEDGMENT

Sriparna Saha would like to acknowledge the support of Serb Women in Excellence Award 2018 for conducting this research.

REFERENCES

[1] Beck, A.T. and Alford, B.A., 2009. Depression: Causes and treatment. University of Pennsylvania Press.
[2] Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B., Berry, J.T. and Mokdad, A.H., 2009. The PHQ-8 as a measure of current depression in the general population. Journal of affective disorders, 114(1-3), pp.163-173.
[3] Kroenke, K., 2012. Enhancing the clinical utility of depression screening. CMAJ, 184(3), pp.281-282.
[4] Scherer, S., 2016. Multimodal behavior analytics for interactive technologies. KI-Künstliche Intelligenz, 30(1), pp.91-92.
[5] Morales, M., Scherer, S. and Levitan, R., 2018, June. A linguistically-informed fusion approach for multimodal depression detection. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (pp. 13-24).
[6] Joan, G. and Kaite, Y., 2015. Gender and depression. Current Opinion in Psychology, pp. 53-60.
[7] Cummins, N., Vlasenko, B., Sagha, H. and Schuller, B., 2017. Enhancing speech-based depression detection through gender dependent vowel-level formant features. Conference on artificial intelligence in medicine in Europe, pp. 209-214.
[8] Lopez-Otero, P. and Docio-Fernandez, L. 2021. Analysis of gender and identity issues in depression detection on de-identified speech. Computer Speech & Language, 65, pp. 101-118.
[9] Parker, G. and Brotchie, H., 2010. Gender differences in depression. International review of psychiatry, 22(5), pp. 429-436.

[10] Andersson, G. and Titov, N., 2014. Advantages and limitations of Internet-based interventions for common mental disorders. World Psychiatry, 13(1), pp.4-11.

[11] Dewan, P., 2015, May. Towards emotion-based collaborative software engineering. In 2015 IEEE/ACM 8th International Workshop on Cooperative and Human Aspects of Software Engineering (pp. 109-112). IEEE.

[12] Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A. and Morency, L.P., 2013, April. Automatic behavior descriptors for psychological disorder analysis. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) (pp. 1-8). IEEE.

[13] Chatterjee, M., Stratou, G., Scherer, S. and Morency, L.P., 2014, May. Context-based signal descriptors of heart-rate variability for anxiety assessment. In 2014 ieee international conference on acoustics, speech and signal processing (ICASSP) (pp. 3631-3635). IEEE.

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

[15] Morales, M.R. and Levitan, R., 2016, December. Speech vs. text: A comparative analysis of features for depression detection systems. In 2016 IEEE spoken language technology workshop (SLT) (pp. 136-143). IEEE.

[16] Wolohan, J.T., Hiraga, M., Mukherjee, A., Sayyed, Z.A. and Millard, M., 2018, August. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In Proceedings of the First International Workshop on Language Cognition and Computational Models (pp. 11-21).

[17] De Choudhury, M., Gamon, M., Counts, S. and Horvitz, E., 2013, June. Predicting depression via social media. In Proceedings of the International AAAI Conference on Web and Social Media (ICWSM) (pp. 128-137).

[18] Benton, A., Mitchell, M. and Hovy, D., 2017. Multi-task learning for mental health using social media text. arXiv preprint arXiv:1712.03538.

[19] Losada, D.E. and Crestani, F., 2016, September. A test collection for research on depression and language use. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 28-39). Springer, Cham.

[20] He, L., Jiang, D., Yang, L., Pei, E., Wu, P. and Sahli, H., 2015, October. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (pp. 73-80).

[21] Dibeklioğlu, H., Alnajar, F., Salah, A.A. and Gevers, T., 2015. Combining facial dynamics with appearance for age estimation. IEEE Transactions on Image Processing, 24(6), pp.1928-1943.

[22] Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M.C. and Sahli, H., 2016, October. Decision tree based depression classification from audio video and language information. In Proceedings of the 6th international workshop on audio/visual emotion challenge (pp. 89-96).

[23] Conklin, A.I., Yao, C.A. and Richardson, C.G., 2018. Chronic sleep deprivation and gender-specific risk of depression in adolescents: a prospective population-based study. BMC public health, 18(1), pp.1-7.

[24] Lopez-Otero, P. and Docio-Fernandez, L., 2021. Analysis of gender and identity issues in depression detection on de-identified speech. Computer Speech and Language, 65, p.101118.

[25] Gratch, J., Artstein, R., Lucas, G.M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S. and Traum, D.R., 2014, May. The distress analysis interview corpus of human and computer interviews. In LREC (pp. 3123-3128).

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

[27] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J. and Bartlett, M., 2011, March. The computer expression recognition toolbox (CERT). In Face and gesture 2011 (pp. 298-305). IEEE.

[28] Degottex, G., Kane, J., Drugman, T., Raitio, T. and Scherer, S., 2014, May. COVAREP—A collaborative voice analysis repository for speech technologies. In 2014 ieee international conference on acoustics, speech and signal processing (ICASSP) (pp. 960-964). IEEE.

[29] Cer, D., Yang, Y., Kong, S-Y., Hua, N., Limtiaco, N., St. John, R, Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y-H., Strope, B. and Kurzweil, R., 2018, May. Universal Sentence Encoder. In arXiv preprint arXiv:1803.11175.

[30] Luong, T., Pham, H. and Manning, CD., 2015, Sept. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1412-1421).

[31] Liu, P., Qiu, X. and Huang, X., 2017. Adversarial multi-task learning for text classification. arXiv preprint arXiv:1704.05742.

[32] Yadav, S., Ekbal, A., Saha, S., Bhattacharyya, P. and Sheth, A., 2018. Multi-task learning framework for mining crowd intelligence towards clinical treatment. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (pp. 271-277)

[33] Morales, M.R., 2018. Multimodal depression detection: An investigation of features and fusion techniques for automated systems. CUNY Academic Works (2018).

[34] Qureshi, S. A., Saha, S., Hasanuzzaman, M. and Dias, G., 2019. Multi-task Representation Learning for Multimodal Estimation of Depression Level. IEEE Intelligent Systems, vol. 34, no. 5 (pp. 45-52)

[35] Liu, P., Qiu, X. and Huang, X., 2017. Adversarial multi-task learning for text classification. 2017, July. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 1-10).

[36] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

[37] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[38] Qureshi, S.A., Hasanuzzaman, M., Saha, S. and Dias, G., 2019. The Verbal and Non Verbal Signals of Depression–Combining Acoustics, Text and Visuals for Estimating Depression Level. arXiv preprint arXiv:1904.07656.

[39] Qureshi, S.A., Dias, G., Hasanuzzaman, M. and Saha, S., 2020. Improving depression level estimation by concurrently learning emotion intensity. IEEE Computational Intelligence Magazine, 15(3), pp.47-59.