Depression Detection Through Multi-Modal Data - Final Report

Pranjal Kaura, Priyanshi Jain, Manan Gupta
Indraprastha Institute of Information Technology - Delhi
{pranjal17079, priyanshi17358, manan17372}@iiitd.ac.in [GROUP 16]

Abstract

Conventionally depression detection was done through extensive clinical interviews, wherein the subject's responses are studied by the psychologist to determine his/her mental state. In our model, we try to imbibe this approach by fusing the 3 modalities i.e. word context, audio, and video and predict an output regarding the mental health of the patient. The output is divided into different levels to take into consideration the level of depression of the subject. We've built a deep learning model that fuses these 3 modalities, assigning them appropriate weights, and thus gives an output. This approach of fusion takes care of the following problems:

- *Presence of noise in one of the modalities.*
- *Control the level of contribution of a particular modality.*

1. INTRODUCTION

An accurate, autonomous, accessible approach to detect depression is the need of the hour. As society moves to more and more stressful environments, a higher percentage of the population is developing depressing tendencies. Only if we're able to detect it, can we work to cure it. The motivation to create such a model is our driving factor. To test our model, clinical interviews of the subjects need to be done for generation of the 3 modalities (as input to our model). It has been noted, through extensive research in this field, that a depressed subject displays various intricate signs, that can be better caught by studying all the 3 modalities together. Due to a change in mental behavior, various physiological and physiological changes can be detected. Research shows that a depressed subject often stammers while talking and thus uneven pauses can be caught in their speech. More occurrences of incorrect pronunciation is another attribute that the subject showcases. Using video modality, other factors such as abnormal eye contact, less frequent mouth movement, changed posture, etc can be caught. Using lexical analysis, the context of the words spoken by the subject can be analyzed, which also provides essential information regarding his/her mental health. Thus integrating all these channels, a more generic model can be built, which takes all these factors into consideration. Thus better predictions can be made due to the presence of more viable factors. Certain challenges that can be expected out of this model are:

- As our model is basically a DL model, large amount of dataset is required, in all 3 modalities.
- Alignment of these 3 modalities, according to their timestamp is another challenge. It's of utmost importance that our model receives these modalities in sync, to understand the correlation between them.
- Since video processing is involved, thus large amounts of computation power will be needed to train our model.

2. LITERATURE REVIEW

D. Huang uses a regression method based on PLS wherein a late fusion detection method is built for model prediction[1]. D.Devault has built a multimodal HCRF model which works on question-answer pairs. It analyses them for model prediction[2]. Gong et. al. use the same approach. Building on it, he combines the question-answer based model with his multi-modal approach, taking into consideration all the 3 modalities for model prediction[3]. Similar work is also done by Sun et al. They built a single model random forest-based classifier which works on the question-answer based approach. This classifier is used for model prediction[4]. Ma et al. propose an audio-based method for depression classification using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for a higher-level audio representation. Ma et al. works only on the audio based modality. He inputs the audio based data into a CNN and then further uses a LSTM network for model prediction[5].

In the work done by Shivakumar et al. [6], the temporal nature of audio/visual modalities is considered using a window-based representation instead of frame-level analysis. Utilizing complementary information from the text and audio features, J. Glass et al. proposed a model in which different LSTM branches for each modality are integrated via a feed-forward network [7]. However, while this work tries to predict depression based on late or early fusion methods [1, 3] or the sequential nature of their inputs [6, 7], learning the time-dependent relationships between language, visual and audio features in detecting depression is still unexplored. The major problems that these approaches face are the different predictive power of modalities and types of noise in the representation. In previous works, gating mechanism has shown to be effective in determining the predictive power of each modality.

Another approach[10] for the same problem explores paralinguistic, head pose and eye gaze behaviours. During the research phase, the authors found out that there are many physical attributes/changes that can be detected through appropriate sensors, when a subject is depressed. In this model, the authors try to detect features like dropped speech rate, lower articulation rate(speech features), Lesser eye contact, wavering eyes(eye features) and bent head(head features). To detect speech related features, the extraction from the dataset was done using a 2 fold approach i.e. Manual Labelling and Automated labeling (using manually segmented data). A total of 63 statistical features were extracted through manual labelling, and 19 speaking rate features were extracted using automated labelling(using PRAAT).For eye features, it was done by training a specialized CV application that is able to detect different attributes of the eye such as the eye lids, pupil, and it's extremities. Using this, it is able to make mathematical calculations that lead to features such blink time, blink frequency, gaze direction (left-right, up-down) etc. Previously, it was that found slower and less frequent head movements, increased eye contact avoidance and less social engagement with the clinical examiner, likely to also show in other social interactions. To extract head pose and movement behaviour, the face had to be detected and tracked before a 3 degrees of freedom (DOF) head pose could be calculated (yaw, roll and pitch). A subject-specific face active appearance model was trained and built, where 30 images per subject were selected for manual annotation, then used for the face model. These 3-DOF pose features, as well as their velocity and acceleration, were extracted to give a total of nine low-level features per frame. All of these eye and head duration features were detected when the feature in question is higher or lower than the average of the feature in question plus or minus the standard deviation of that feature for each subject's interview. For the method in this paper, the base of the model is an SVM classifier. It is used to classify the features into binary classes i.e. Yes(Depressed) or No(Not depressed). The extracted features are further sifted using feature extraction/Dimensionality reduction techniques like Statistical Analysis using t-test algorithm and Principal component analysis. Every feature was also normalised to bring down to one scale. For fusion, early, late and hybrid fusions are explored in this paper. For early fusion, feature fusion is explored that is basically concatenating extracted features from the raw data. In late fusion, results from each modality are combined after training them separately. This was done on labels (decision fusion) and scores (score fusion) from the classifier. In this paper, a comparatively new fusion technique is also explored which is hybrid fusion. In hybrid fusion, feature fusion of all modalities is performed first to create a new modality, which is then treated as an additional individual modality. The scores/decisions of this new modality are then fused with the scores/decisions of the individual modalities in either one or two levels. The dataset taken in this paper was relatively small due to which the results weren't conclusive.

The most recent approach [8] for this problem explores a model-based optimal fusion, that is, instead of using early fusion or late fusion technique, it focuses more on how much each modality should have an impact on the final result. Early fusion is basically concatenating the feature vectors of each modality after extraction into a single vector and feeding them to the model to learn the results. In the late fusion technique, we train individual models for each modality and then combine their results to get a final output by giving them some weights. What both of these approaches ignore is that learned representation of one modality can be undermined by the other modalities.

3. DATASET

3.1. DAIC-WOZ DATASET

The DAIC-WOZ dataset [9] was collected by the University Of Southern California. It is a part of a larger DAIC (Distress Analysis Interview Corpus) that contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and PTSD.

3.2. Modalities

The dataset contains audio and video recordings and extensive questionnaire responses. Additionally, the DAIC-WOZ dataset includes the Wizard-Of-Oz interviews, conducted by an animated virtual assistant called Ellie, who is controlled by a human interviewer in another room. The data has been transcribed and annotated for a variety of verbal and non-verbal features. Each participant's session includes a transcription of interaction, participant audio files, and facial features extracted from the recorded video.

3.2.1 Video Modality

The dataset contained facial features from the videos of the participant. The facial features consisted of 68 2D points on the face, 24 AU features that measure facial activity, 68 3D points on the face, 16 features to represent the subject's gaze, and 10 features to represent the subject's pose. This made for a total of 388 video features.

3.2.2 Audio Modality

The audio features are for every 10ms, thus the features are sampled at 100Hz. The features include 12 Mel-frequency cepstral coefficients (MFCCs), these are F0, VUV, NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, Rd, Rdconf, MCEP0-24,HMPDM0-24, HMPDD0-12. Along with the MFCCs we also have features for pitch tracking, peak slope, maximal dispersion quotients, glottal source parameters. Additionally, the VUV (voiced/unvoiced) feature flags whether the current sample is voice or unvoiced. In the case where the sample is unvoiced (VUV = 0), F0, NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, and Rd are set to 0.

3.2.3 Text Modality

The textual modality contains the transcript for the whole conversation of the patient with the RA in csv format. Individual sentences have been timestamped and further classified on the basis of their speaker. Expressions like laughter, frown etc have been added in angular brackets as and when they occur (for e.g. ¡Laughter¿). Differentiation between long/short pauses has not been made. Only word (not phenome) level segmentation has been recorded.

3.3. Dataset size

The dataset contains 189 sessions of interactions, ranging anywhere from 7 to 33 minutes. The dataset contains interviews with 59 depressed and 130 non-depressed subjects.

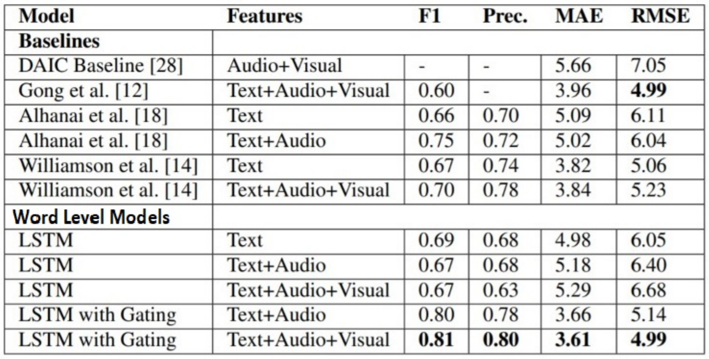
4. PROPOSED SOLUTION

In our system, we plan to first extract features and then apply some gating mechanism and hybrid fusion technique on the features extracted. For feature extraction: We have audio, visual, and textual modalities as our features that are integrated using time-stamps to learn the time-dependent interactions between them. The forced alignment will be done on a sentence level granularity. This is because we want the model to learn the context between words. This is the preprocessing part.

Now, we have aligned the textual, audio, and visual features at the sentence level. One important thing to note is that different modalities can have different impacts on the final result and there is some noise involved too while representing the features of different modalities. Now, on the extracted features, some gating mechanism will be applied to learn and control how much different modalities will be contributing to the final output. In our network, we'll use weight vectors with each modality to control and learn how much information will be transformed and carried to the next layers.

For each time step, the feature vectors from each modality will be concatenated and then passed to the word-level LSTM which comprises of the gating mechanism, Before the concatenation of the feature vectors, the audio and visual vectors will be also passed through gating mechanism to extract the important information.

The other approach that we can follow is to use a hybrid fusion technique, to reap the benefits of both early and late fusion. Hybrid fusion can be performed on one level or two levels. In hybrid fusion, feature fusion of all modalities is performed first to create a new modality, which is then treated as an additional individual modality. The scores/decisions of this new modality are then fused with the scores/decisions of the individual modalities in either one or two levels.



Model	Features	F1	Prec.	MAE	RMSE
Baselines					
DAIC Baseline [28]	Audio+Visual	-	-	5.66	7.05
Gong et al. [12]	Text+Audio+Visual	0.60	-	3.96	**4.99**
Alhanai et al. [18]	Text	0.66	0.70	5.09	6.11
Alhanai et al. [18]	Text+Audio	0.75	0.72	5.02	6.04
Williamson et al. [14]	Text	0.67	0.74	3.82	5.06
Williamson et al. [14]	Text+Audio+Visual	0.70	0.78	3.84	5.23
Word Level Models					
LSTM	Text	0.69	0.68	4.98	6.05
LSTM	Text+Audio	0.67	0.68	5.18	6.40
LSTM	Text+Audio+Visual	0.67	0.63	5.29	6.68
LSTM with Gating	Text+Audio	0.80	0.78	3.66	5.14
LSTM with Gating	Text+Audio+Visual	**0.81**	**0.80**	**3.61**	**4.99**

Figure 1. Baseline Results

5. BASELINE RESULTS

Our baseline paper[11] explores the DAIC-WOZ dataset. It uses a gating mechanism for learning the features extracted and a time-dependent recurrent framework. The results of the paper are given in Figure 1. It can be seen that LSTM with gating used on all the three modalities perform the best.

6. WORK DONE

The given DAIC dataset is skewed with a 7:3 ratio, of non-depressed class to depressed. To overcome the biases, the dataset was upsampled. The following models were applied to the dataset:

Model	Modality	Precision	Recall	F1 - Score
Late Fusion Using SVM	Text + Audio + Video	0.442	0.387	0.413
CNN	Text	0.569	0.618	0.587
	Audio	0.087	0.3	0.135
	Video	0.087	0.3	0.135
LSTM	Text	0.657	0.68	**0.667**
	Audio	0.574	0.455	0.464
	Video	0.49	0.679	0.567
LSTM w/o Gating (Sentence-Level)	Text + Audio	0.712	0.359	0.476
	Text + Video	0.604	0.467	0.476
	Text + Audio + Video	0.621	0.488	0.5
LSTM w Gating (Sentence-Level)	Text + Audio	0.638	0.652	**0.647**
	Text + Video	0.651	0.537	0.558
	Text + Audio + Video	0.638	0.652	**0.647**
LSTM w Gating (Word-Level)	Text + Audio + Video	0.489	0.651	0.553
BiLSTM (Word-Level)	Text + Audio + Video	0.79	0.321	0.187

Table 1.

- **SVM and Random Forest:** Firstly, SVM (with an RBF kernel) and Random forest were applied to the three modalities separately and then another SVM model was trained on the decision labels from the individual modalities to perform late fusion. For this purpose, the features of the audio and video modality were averaged over all the timestamps to give a total of 74 and 388 features, respectively. For the text modality, the word2Vec model obtained from google-news-300 was applied to transform each word into a vector of size 300. Further, the 3D vector obtained (sentences x words x 300 features) was first averaged over each word and then flattened.

- **CNN:** A CNN model having 6 layers was built with the first 4 layers having conv2D layers for the text modality and conv1D layers for audio and video modality and Max Pooling layers. Further flattening and fully connected layers were added with the ReLU activation function. Sigmoid activation was used in the last layer. For audio and video modality, features of the first 40,000 and timestamps, respectively, were taken. These values were chosen according to the available computation capability. For the text modality, after applying the Word2vec model, thresholds were set for the maximum number of words and sentences.

- **LSTM model with/without gating (Sentence-level):** Firstly, the data was converted to sentence level by averaging audio and video modality features for the given time-stamps of sentences i.e. sentence level force alignment was done.

  Next, we used highway layers for gating the audio and video features. Each highway layer comprises two non-linear transforms: a Carry and a Transform gate which define the degree to which the output is created by transforming the input and determining how much information should move forward. After feeding the audio and video features for a sentence to highway layers individually, they're concatenated with the corresponding text feature. The concatenated vector is then passed through a LSTM to get the final output.

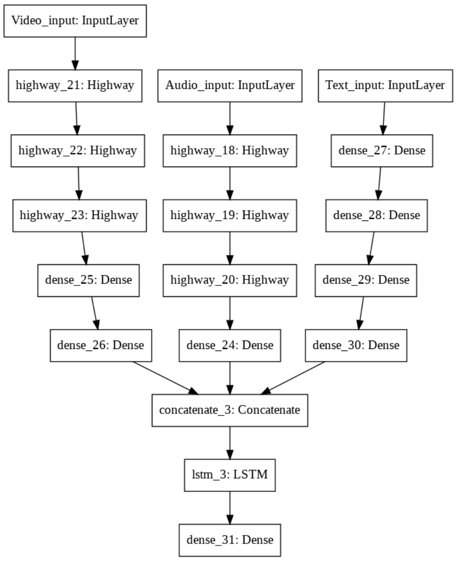


Figure 2. Model Architecture

Model Architecture: Figure 2. shows the basic model architecture. The audio and video features are first passed through 3 feedforward highway layers. Then, dense layers are used to reduce the dimensionality of both video and text features. After concatenation, LSTM with 128 hidden nodes is used. And finally, a dense layer is applied with sigmoid activation to get

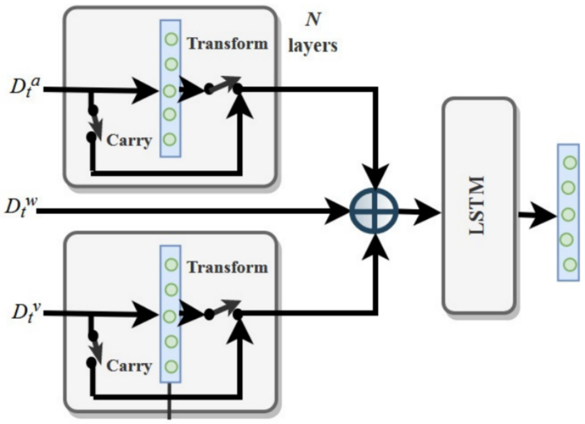


Figure 3. Sentence-level multi-modal fusion with gating

the output. A learning rate of 0.0001 is used. For the number of epochs, a EarlyStopping callback is used from Keras API.

- **(Bi)LSTM Model with Gating (Word-Level):** Forced alignment of the data was done on a word level basis. The alignment was done using a unitary approach wherein the sentence level time stamps were converted to word-level based on the number of words(present in the sentence) and the character length present in the sentence.

  Next, we used highway layers for gating the audio and video features. Each highway layer comprises two non-linear transforms: a Carry and a Transform gate which define the degree to which the output is created by transforming the input and determining how much information should move forward. After feeding the audio and video features for a sentence to highway layers individually, they're concatenated with the corresponding text feature. The concatenated vector is then passed through a (Bi)LSTM to get the final output.

  Model Architecture: The audio and video features are first passed through 3 feedforward highway layers. Then, dense layers are used to reduce the dimensionality of both video and text features. After concatenation, (Bi)LSTM with 128 hidden nodes is used. And finally, a dense layer is applied with sigmoid activation to get the output. A learning rate of 0.0001 is used. For the number of epochs, EarlyStopping callback is used from Keras API.

7. RESULTS

The results have been published by taking a weighted mean of the 2 classes, i.e. class 0(Not depressed) and class 1(Depressed). The data provided is in the ratio 7:3.

- **SVM Model:** The model did not perform well, as can be seen from the results in table Table 1. This could be due to the fact that averaging operations were performed across the 3 modalities. This could have led to the loss of a lot of information, leading to the model under-performing.

- **CNN Model:** This model performed better than the SVM one on the text modality because herein averaging across word vectors was not done. The audio and video modalities were still not giving satisfactory results. This could be due to the fact that the data points were too few and the features representing these modalities were too sparse.

- **LSTM with/without Gating at Sentence-Level:**

  - The results clearly indicate that our model works best for Text modality. The low values of the F1 score in the video and audio modality show that these features do not represent the depression class well. This could be the reason that when audio,video and text modality are combined, the results just fall short of that of the model which only uses Text modality.

  - Also, all models that use gating perform better than those models that do not. This could be because using gating, only the most important features are amplified, while the others are nullified. Thus a sort of feature extraction takes place at this level, which helps our LSTM model to learn from only the most favorable features.

- **LSTM with gating at word-level:** The results for word-level LSTM are not as good as expected. The reason could be that on a word level, the model does not get the context of the conversation as much as it does on a sentence level.

- **BiLSTM model:** This model shows clear partiality towards the 'depressed class'. The model is not able to learn much from the data.

8. CONCLUSION

A model was presented to detect if a person is depressed or not based on indicators from audio, video and lexical modalities. A sentence-level model with highway layers as gating mechanism was used for the task. According to our models, sentence-level seems to work best amongst other models. A mixture of early and late fusion was used to get better interpretation from each modality. For future scope, the features could be extracted on a better level. Some audio features like response time, number of pauses, silence rate can also be examined to get a better understanding about the symptoms. Interaction of bodily action sequences from motion capture data can be studied with the verbal behaviour to have a more extensive study.

9. INDIVIDUAL CONTRIBUTION

- Manan: Video, Worked together on combined modalities (33
- Priyanshi: Audio, Worked together on combined modalities (33
- Pranjal: Text, Worked together on combined modalities (33

10. References

[1] H.Meng, D.Huang, H.Wang, H.Yang, M. Ai Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression,"in Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. ACM, 2013, pp. 21–30.

[2] Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P.Morency, and J. Cassell, "Multimodal prediction of psychologicaldisorders: Learning verbal and nonverbal commonalities in adjacency pairs," in Semdial 2013 DialDam: Proceedings of the 17thWorkshop on the Semantics and Pragmatics of Dialogue, 2013,pp. 160–169.

[3] Y. Gong and C. Poellabauer, "Topic modeling based multi-modaldepression detection," inProceedings of the 7th Annual Workshopon Audio/Visual Emotion Challenge. ACM, 2017, pp. 69–76.

[4] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "Arandom forest regression method with selected-text feature for de-pression assessment," inProceedings of the 7th Annual Workshopon Audio/Visual Emotion Challenge. ACM, 2017, pp. 61–68.

[5] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet:An efficient deep model for audio based depression classifica-tion," inProceedings of the 6th International Workshop on Au-dio/Visual Emotion Challenge. ACM, 2016, pp. 35–42.

[6] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula,and P. Georgiou, "Multimodal and multiresolution depression de-tection from speech and facial landmark features," inProceedingsof the 6th International Workshop on Audio/Visual Emotion Chal-lenge. ACM, 2016, pp. 43–50.

[7] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," inProc. Inter-speech, 2018, pp. 1716–1720.

[8] http://www.eecs.qmul.ac.uk/ mpurver/papers/rohanian-et-al19interspeech.pdf

[9] University Of Southern California. "DAIC-WOZ Dataset." Dcapswoz.ict.usc.edu, dcapswoz.ict.usc.edu/.

[10] Alghowinem, Sharifa Goecke, Roland Wagner, Michael Epps, Julien Hyett, Matthew Parker, Gordon Breakspear, Michael. (2016). Multimodal Depression Detection:Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors. IEEE Transactions on Affective Computing. PP. 1-1. 10.1109/TAFFC.2016.2634527.

[11] Rohanian, Morteza Hough, Julian Purver, Matthew. (2019). Detecting Depression with Word-Level Multimodal Fusion. 1443-1447. 10.21437/Interspeech.2019-2283.