Depression Recognition Using Machine Learning

Alvin Abraham[1]
[1] Student, Dept. Of Computer Science & Engineering, Mangalam College of Engineering, India,

Albin Varkey[2]
[2] Student, Dept. Of Computer Science & Engineering, Mangalam College of Engineering, India,

Geo Sabu[3]
[3]Student, Dept. Of Computer Science &Engineering, Mangalam College of Engineering, India,

Dheeraj N[4]
[4]Assistant Professor, Dept. of AI & ML, Mangalam College of Engineering, India,

***Abstract*— Traditionally, depression was identified through in-depth clinical interviews, during which the psychologist would analyze the subject's responses to ascertain his or her mental condition. In our model, we attempt to emulate this strategy by combining the three modalities of word context, audio, and video to forecast an outcome pertaining to the patient's mental health. To account for the subject's state of depression, the output is separated into various levels. We have developed a deep learning model that combines all three modalities, gives them the proper weights, and produces an output. The following issues are addressed by this fusion strategy:**
• Control the amount of contribution from each modality;
• The presence of noise in one of the modalities.

1. INTRODUCTION

The current state of affairs calls for an effective, independent, and accessible method to identify depression. More and more people are becoming depressed as society creates conditions that are more and more stressful. We can only attempt to treat it if we can find it in the first place. Our motivating force is the desire to develop such a model. Clinical interviews with the individuals must be conducted in order to generate the three modalities that will be used as input in our model's testing. It has been discovered via significant research in this area that a depressed subject exhibits a variety of complex symptoms that can be detected more effectively by combining the three modalities. A shift in mental behaviour can cause a variety of physiological and physiological changes. According to research, people who are depressed tend to stutter when they speak, causing unnatural pauses to appear in their speech. Another feature that the topic emphasises is more instances of erroneous pronunciation. Other indicators, such as unusual eye contact, less frequent mouth movement, altered posture, etc., can be detected using the video modality. Lexical analysis can be used to examine the subject's speech in context, which also reveals crucial details about his or her mental state. A more general model that takes into account all of these elements can be developed by merging all of these channels. As a result of the availability of more reliable components, better forecasts can be made. This model will likely face the following difficulties:

• Because our model is essentially a DL model, a sizable dataset in each of the three modalities is needed.
• Another difficulty is aligning these 3 modalities in accordance with their timing. It is crucial for our model to receive these modalities simultaneously in order to comprehend how they are correlated.
• Since video processing is required, training our model will require a lot of computing resources.

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 questionanswer 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 audiobased 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 etal.(6), the temporal nature of audio/ visual modalities is considered using a window-grounded representation rather of frame- position analysis. exercising reciprocal information from the textbook and audio features,J. Glass etal. proposed a model in which different LSTM branches for each modality are integrated via a feed-forward network(7). still, while this work tries to prognosticate depression grounded on late or early emulsion styles(1, 3) or the successional nature of their inputs(6, 7), learning the time-dependent connections between language, visual and audio features in detecting depression is still unexplored. The major problems that these approaches face are the different prophetic power of modalities and types of noise in the representation. In former workshop, gating medium has shown to be effective in determining the prophetic power of each modality

Another approach(10) for the same problem explores paralinguistic, head disguise and eye aspect behaviours. During the exploration phase, the authors set up out that there are numerous physical attributes changes that can be detected

through applicable detectors, when a subject is depressed. In this model, the authors try to descry features like dropped speech rate, lower articulation rate(speech features), lower eye contact, wavering eyes(eye features) and fraudulent head(head features). A aggregate of 63 statistical features were uprooted through homemade labelling, and 19 speaking rate features were uprooted using automated labelling(using PRAAT). For eye features, it was done by training a technical CV operation that's suitable to descry different attributes of the eye similar as the eye lids, pupil, and it's extremities.

Using this, it's suitable to make fine computations that lead to features similar blink time, blink frequence, aspect direction(left- right, over- down) etc. preliminarily, it was that set up slower and lower frequent head movements, increased eye contact avoidance and lower social engagement with the clinical monitor, likely to also show in other social relations. To prize head disguise and movement geste , the face had to be detected and tracked before a 3 degrees of freedom(DOF) head disguise could be calculated(yaw, roll and pitch). A subject-specific face active appearance model was trained and erected, where 30 images per subject were named for homemade reflection, also used for the face model. These 3-DOF disguise features, as well as their haste and acceleration, were uprooted to give a aggregate of nine low- position features per frame. All of these eye and head duration features were detected when the point in question is advanced or lower than the normal of the point in question plus or disadvantage the standard divagation of that point for each subject's interview. For the system in this paper, the base of the model is an SVM classifier. It's used to classify the features into double classes i.e. Yes(Depressed) or No(Not depressed). The uprooted features are farther sifted using point birth/ Dimensionality reduction ways like Statistical Analysis using t- test algorithm and star element analysis.

For emulsion, beforehand, late and cold-blooded mixtures are explored in this paper. For early emulsion, point emulsion is explored that's principally concatenating uprooted features from the raw data. In late emulsion, results from each modality are combined after training them independently. This was done on markers(decision emulsion) and scores(score emulsion) from the classifier. In this paper, a comparatively new emulsion fashion is also explored which is cold-blooded emulsion. In cold-blooded emulsion, point emulsion of all modalities is performed first to produce a new modality, which is also treated as an fresh individual modality. The scores opinions of this new modality are also fused with the scores opinions of the individual modalities in either one or two situations. The dataset taken in this paper was fairly small due to which the results were n'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

A. 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.

B. Modalities

The dataset contains audio and video recordings and extensive questionnaire responses. Additionally, the DAICWOZ 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.

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.

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, MCEP024,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) 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.

C. 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 mongrel emulsion fashion, to reap the benefits of both early and late emulsion. mongrel emulsion can be performed on one position or two situations. In cold-blooded emulsion, point emulsion of all modalities is performed first to produce a new modality, which is also treated as an fresh individual modality. The scores opinions of this new modality are also fused with the scores opinions of the individual modalities in either one or two situations.

Table 1: Baseline Results

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**

5.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 models were applied to the dataset;

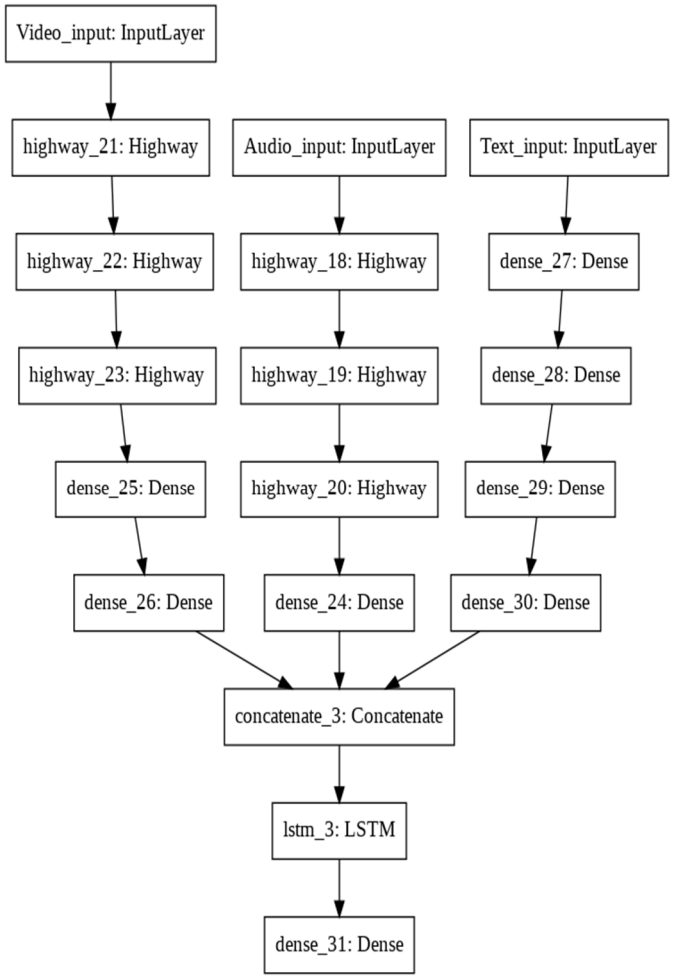

Figure 1. Model Architecture

Random Forest: Firstly, SVM (with an RBF kernel) and Random forest were applied to the three modalities separately and then another SVMmodel 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.

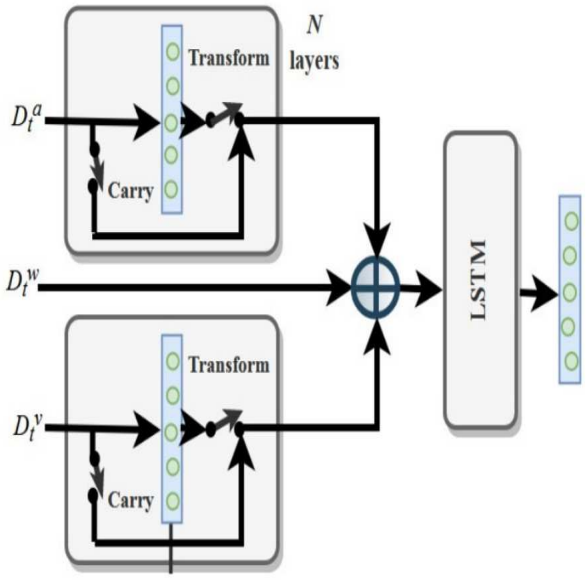

Figure 2. Sentence-level multi-modal fusion with gating

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.

5. 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.

6. 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.

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] 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.

[5] 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.