Detection of Depression using Multi-Modal data

Divija Lekkalapudi   
CSE Department  
PES UNIVERSITYBengaluru,India  
PES1UG20CS134

Dhathri Gurram  
CSE Department  
PES UNIVERSITYBengaluru,India  
PES1UG20CS141

Gaurav Mahajan  
CSE Department  
PES UNIVERSITYBengaluru,India  
PES1UG20CS150

*Abstract*— Traditionally, depression was detected through extensive clinical interviews in which the psychologist studied the subject's responses to determine his/her mental state. This approach is something we try to incorporate into our model. By combining two modalities of text and audio out of three, namely word context, audio, and video, and predicting the patient's mental health. The output is divided into levels based on the subject's level of depression. We created a deep learning model that combines the two aforementioned model modalities, assigns appropriate weights to them, and generates an output. This fusion strategy addresses the following issues:

• Noise in one of the modalities

• Limit the amount of contribution Keywords—component, formatting, style, styling, insert (key words)

# Introduction

# A reliable, self-contained, and easily accessible method for detecting depression is urgently required. As society becomes more stressful, an increasing percentage of the population develops depressive tendencies. We can only work to cure it if we can detect it. Our motivation for developing such a model is our driving force. Clinical interviews with subjects are required to generate the two modalities for testing our model (as input to our model). Extensive research in this field has revealed that a depressed subject exhibits a variety of intricate signs, which can be better detected by studying all three modalities concurrently. Various physiological and physiological changes can occur as a result of a change in mental behaviour and attitude. According to research, a depressed person frequently stammers while speaking, resulting in uneven pauses in their speech. Another feature of the subject is the increased occurrence of incorrect pronunciation. Other factors, such as abnormal eye contact, less frequent mouth movement, changed posture, and so on, can be detected using video technology. The context of the subject's words spoken can be analysed using lexical analysis, which also provides important information about his/her mental health. By integrating all of these channels, a more generic model can be constructed that takes all of these factors into account. As a result of the presence of more viable factors, better predictions can be made. This model is likely to present the following challenges:

# Because our model is essentially a DL model, a large amount of dataset in both modalities is required.

* Another challenge is aligning these modalities based on their timestamps. It is critical that our model receives these modalities in sync in order to understand the relationship between them.
* A significant amount of computation power will be required to train our model.

# LITERATURE REVIEW

Chen and Zhang [1] recommended One of the most important forms of verbal communication is written communication.

Text has its own unique characteristics for expressing a person's emotional or mental state. Sentiment analysis analyses, processes, and induces texts with emotional colours using NLP, text analysis, Ml, and other methods. In this social media age, text sentiment analysis can be used to analyse a specific person's mental and emotional state using data extracted from social media sites such as Twitter. SVM and CNN methods can be used to improve the data's accuracy and recall rate. CNN is used as the automatic feature learner in this model, and SVM is the final text classifier. The CNN-SVM model outperforms the CNN and NLPCC-SCDL best models in terms of accuracy, demonstrating that this model is more practical for dealing with text sentiment analysis and classification.

R.Danush Vikram *et al.* [2] recently discovered depressive signals in tweets from people suffering from severe anxiety and depression disorders. A large number of tweets from various users were extracted and clinically tested, and behavioural attributes such as emotion, language, linguistic styles, and so on were extracted from this data set. According to their findings, word frequency and topic modelling are useful features for developing prediction models. They concluded that combining sentiment analysis with the percentage of depressed tweets improved the precision of detecting depression when using an SVM classifier to classify the tweets. SVM predicts the likelihood of depression with a 70% accuracy.

They also used LE-LSTM, in which the data obtained from Twitter was cleaned and normalised before being segmented, stemmed, and lemmatized. The data was fed into a machine learning classifier, which could distinguish between depressed and non-depressed text tweets. They have created a framework that works well and produces accurate results. The Lexicon Enhanced LSTM model (LE-LSTM) keeps words in memory for a long time, allowing it to compare words for depression detection from textual tweets. This project's accuracy is 92.8%.

The goal of Muley et al’s [3 ]paper is to develop a framework that can identify the client's depression level by connecting visual inputs and the BDI-II inventory result. This paper is based on a nonverbal examination of depression and abstracts the visible signs identified within the inspected thoughts. Depending on the programme bundles such as OpenFace, CERT, and SEMAINE API are used for preprocessing. Depending on the specific investigation objectives, various choice strategies may be connected, including classification strategies such as SVM, Closest Neighbor, etc. to address categorical questions and regression strategies such as SVR, Straight relapse, etc. The exactness of the framework is computed in accordance with the confusion matrix. Preprocessing, feature extraction, depression detection: classification, correlation, and regression are all steps taken in this approach. The web application allows the user to input video and solve the BDI-II inventory. The converted videos are fed into a neural network model for face detection. The output is fed into a trained CNN model for feature extraction and classification, resulting in an emotion vector output. The model is evaluated based on precision, recall, and accuracy; however, the facial cicatrix may affect accuracy, which can be remedied by increased data availability. The output is divided into four depression categories: Minimal Level, Mild Level, Moderate Level, and Severe Level. The output is analysed using graphs based on time and demographic factors such as age and gender. Reports are generated to provide details such as the user's basic information, BDI score, depression level detected, and curative measures. The accuracy of the machine learning model achieved using CNN is 66.45%.

Yu Ching Huang et al. [4] sought to predict one's depression proclivity by analysing the image, text, and behaviour of his/her Instagram postings. They use an efficient data collection mechanism that does not require users to take any screening tests. It outperforms previous studies that used the CES-D (Center for Epidemiological Studies Depression) questionnaire. introduce a deep learning model that uses text, image, and behaviour as features to predict a user's depression proclivity. The central idea behind this approach is that the internal layers of CNN act as an extractor of mid-level image representations, which can then be pre-trained on a large dataset (i.e., ImageNet) and used for depressive and non-depressive image classification. Second, they apply the previously trained weights parameters to the target task, remove the pre-trained model's output layer and replace it with two fully connected classification layers that output image probabilities. The degree of depression is then generated. Before the training and testing processes, they use min-max normalisation to normalise all feature values into [0, 1]. The model is trained using backpropagation and the Adam optimizer. As our loss function, they use cross-entropy. Using both image and text features strengthens the model more than using either image and behaviour features or behaviour and text features alone. We discovered that while user behaviour features on Instagram contribute less significantly to distinguishing between depressive and non-depressive users, they still improve the overall model's performance.

Sonam Gupta et al. [6] chose and trained the tweet data, preprocessed the datasets, and removed the raw data from the dataset. Following value extraction, tweet data was trained and the training dataset was cross-validated. For depression detection in tweets, we used five machine learning classifiers: support vector machines, decision trees, logistic regression, K-nearest neighbour, and LSTM. In the depression detection approach, the results show that the LSTM classification model outperforms the other baseline models. Oversampling and undersampling of class imbalance approaches, such as SMOTE and RUS, are implemented and analysed to deal with the imbalanced dataset.

Tuka Alhanai et al. [7] proposed that medical professionals diagnose depression by interpreting individuals' responses to a variety of questions, probing lifestyle changes, and continuing thoughts. An effective automated agent, like a professional, must understand that responses to queries have varying prognostic value. Given the question-and-answer nature of depression screening tests, they were interested in modelling depression using sequences of responses rather than formally conditioning on the type of questions asked. A data-driven system like this has the advantage of requiring little prior knowledge of the structure of an interview or interaction. Furthermore, for a model to be truly data-driven, feature engineering should be kept to a minimum. They demonstrate an automated depression-detection algorithm in this study. To detect depression, they used data from 142 people who had undergone depression screening and modelled the interactions with audio and text features in a Long-Short Term Memory (LSTM) neural network model.

•A regularised logistic regression model, CONTEXT FREE MODEL, was built without conditioning on the type of questions asked. A WEIGHTED MODEL of a regularised logistic regression model with conditioning on the type of questions asked. SEQUENCE MODELLING is an LSTM model that uses response sequences without knowing the type of questions that prompted the response.

Morteza Rohanian et al. [8] use a time-dependent recurrent approach in a deep learning setup to propose a model that can perform modality fusion incrementally after each word in an utterance. To deal with noisy modalities, authors use fusion gates to control how much the audio or visual modality contributes to the final prediction. Current approaches employ either early feature-level fusion, in which features from various modalities are combined into a new feature set for classification, or late prediction-based fusion, in which separate classifiers are trained on each modality to predict the depression state, and the output of those classifiers is combined into a single prediction. There has been work on combining temporal and spatial data in other related tasks using multimodal fusion to predict a cognitive state.

Work on combining temporal information from two or more modalities in a recurrent approach in audio/visual emotion classification and image captioning tasks has been done in other related tasks using multimodal fusion to predict a cognitive state. This work demonstrated the ability to learn complex decision boundaries that other models using different fusion methods struggle with. One major issue with these models is dealing with the different predictive power of each modality, as well as the various levels and types of noise. In different multimodal tasks, adding gating mechanisms has been shown to be effective in dealing with the level of contribution of each modality to the final prediction. Pre-processing: For word timings, use Forced Alignment. We align words with their corresponding audio time at each time step. They employ feed-forward highway layers with gating units that learn to control information flow through the network by weighting visual and audio inputs at each time step. Each highway layer contains two non-linear transforms: a Carry (Cr) and a Transform (T r) gate, which define the extent to which the output is created by transforming and carrying the input (how much information should move forward or be changed in successive training epochs). The gating mechanism is first applied to the audio and visual feature input vectors D a t and D v t, which are then concatenated with the current word embedding D w t after passing through N highway layers (where the best value N is determined by optimising on held out data). The resulting Mean Absolute Error (MAE) loss is used as the signal for training our highway layers using the REINFORCE rule after training our LSTM with gating. They demonstrated a model that learns depression indicators from audio and visual modalities, as well as lexical information in transcript texts. As a gating mechanism, they used word-level multimodal fusion with feed-forward highway layers. Their main goal was to capture intermodal dynamics in a joint multimodal representation.

# DATASET

## DATASET DAIC-WOZ

The University of Southern California collected the DAIC-WOZ dataset [9]. It is a subset of the larger DAIC (Distress Analysis Interview Corpus) collection, which includes clinical interviews designed to aid in the diagnosis of psychological distress conditions such as anxiety, depression, and PTSD*.*

## Modality

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.

1. 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.

1. 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.

## 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.

# SOLUTION

We used the two modalities of text and audio to determine depression from patient interviews. We employed the Support Vector Machine (SVM), the Random Forest Classifier (RF), and the Convolutional Neural Network (CNN).

We built these models to work with multiple modalities and tracked the results.

The given DAIC dataset is skewed with a 7:3 non-depressed class to depressed class ratio. The dataset was upsampled to compensate for the biases. The dataset was subjected to the following models:

SVM and Random Forest: First, SVM (with an RBF kernel) and Random Forest were applied to the two modalities separately, and then another SVM model was trained on the individual modalities' decision labels to perform late fusion. For this purpose, the audio and video modality features were averaged across all timestamps, yielding a total of 74 and 388 features, respectively. The word2Vec model obtained from google-news-300 was used to transform each word into a vector of size 300 for the text modality. Furthermore, the obtained 3D vector (sentences x words x 300 features) was averaged over each word before being flattened.

CNN: A CNN model with six layers was created, with the first four layers consisting of conv2D layers for text modality, conv1D layers for audio and video modality, and Max Pooling layers. With the ReLU activation function, additional flattening and fully connected layers were added. In the final layer, sigmoid activation was used. The first 40,000 and timestamps were used for audio and video modality, respectively.These values were chosen based on the computation capability available. After applying the Word2vec model to the text modality, maximum number of words and sentences were set.

# RESULT

The findings were published by taking a weighted mean of the two classes, namely class 0 (Not depressed) and class 1. (Depressed). The data provided is in the 7:3 ratio.

SVM Model: The model did not perform well. This could be because averaging operations were performed across all three modalities. This could have resulted in the loss of a large amount of data, causing the model to underperform.

CNN Model: This model outperformed the SVM model on the text modality because averaging across word vectors was not performed. The audio and video modalities continued to produce unsatisfactory results. This could be because the data points were too few and the features representing these modalities were too few.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MODEL** | **MODALITY** | **PRECISION** | **RECALL** | **F1-SCORE** |
| **SVM** | **TEXT** | **0.42** | **0.32** | **0.353** |
| **SVM** | **AUDIO** | **0.61** | **0.33** | **0.43** |
| **SVM** | **TEXT-AUDIO** | **0.442** | **0.387** | **0.413** |
| **RF** | **TEXT** | **0.25** | **0.31** | **0.51** |
| **RF** | **AUDIO** | **0.28** | **0.36** | **0.31** |
| **CNN** | **TEXT** | **0.569** | **0.618** | **0.587** |
| **CNN** | **AUDIO** | **0.83** | **0.34** | **0.49** |

**Table 1.Results**

##### CONCLUSION

A model was presented to detect if a person is depressed or not based on indicators from audio, video and lexical modalities. 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 5 symptoms. Interaction of bodily action sequences from motion capture data can be studied with the verbal behaviour to have a more extensive study. The results that we obtained could be hybridised with the video modality which wasn’t considered because of technical limitations.

##### FUTURE

We could use the LSTM model to enhance the purpose of detection. We could modify the model with gating techniques to get a better result.

Since the dataset collected was for the year 2017, we anticipate the new and updated dataset post the Covid-19 era. This upcoming dataset would have the potency to give us brilliant insights on how the world has faired with regards to mental health.

REFERENCES

1. Y. Chen and Z.Zhang, Research on text sentiment analysis based on CNN and SVM
2. R.Danush Vikram ,K.S.Sudhishna ,M.Lingeshraj ,N.Gokulswaruban, AI to Detect Social Media Users’ Depression Polarity Score and Diagnose using Auto Curative Therapy
3. Akshada Mulay,Anagha Dhekne,Pranjali Deshpande,Shivani Kadam Automatic Depression Level Detection Through Visual Input.
4. Yu Ching Huang,Chieh-Feng,Chiang and Arbee . P. Chen Predicting Depression Tendency based on Image, Text and Behavior Data from Instagram
5. Shephali Santosh Nikam,Aishwarya Vijay Patil,Gauri Shashikant Patil, Sharvari Pramod Patil and B. D. Jitkar , AI Therapist Using Natural Language Processing
6. Sonam Gupta ,Lipika Goel Arjun,Singh Ajay Prasad,Mohammad Aman Ullah , Psychological Analysis for Depression Detection from Social Networking Sites
7. Tuka Alhanai, Mohammad Ghassemi, James Glass, Detecting Depression with Audio/Text Sequence Modelling of Interviews