IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125123

MULTIMODAL DEPRESSION DETECTION FROM FACIAL LANDMARK FEATURES USING LSTM MODEL

D.SYLVIA SHARON [1], J ANGEL OZNI [2], S. SOMALAKSHMI [3]

Department of Information Technology, DMI College of Engineering, Chennai, Tamil Nadu, India[1]

Department of Information Technology, DMI College of Engineering, Chennai, Tamil Nadu, India[2]

Department of Information Technology, DMI College of Engineering, Chennai, Tamil Nadu, India[3]

Abstract: This paper proposes massive and growing burden imposed on modern society by depression has motivated investigations into early detection through automated, scalable, and non-invasive methods, including those based on speech. However, speech-based methods that capture articulatory information effectively across different recording devices and in naturalistic environments are still needed. This article presents a novel multi-level attention-based network for multi-modal depression prediction that fuses features from audio, video, and text modalities while learning the intra and inter modality relevance. Multi-level attention reinforces overall learning by selecting the most influential features within each modality for decision-making. We perform exhaustive experimentation to create different regression models for audio, video, and text modalities. Evaluations of both landmark duration features and landmark n-gram features on the DAIC-WOZ and SH2 datasets show that they are highly effective, either alone or fused, relative to existing approaches.

Keywords: Depression classification, landmark *n*-grams, speech articulation, smartphone speech, naturalistic environments

I. INTRODUCTION

Depression a major mental disorder reported to affect 10-15% of the world's population, places severe health, security, productivity, and economic burdens on modern society. Early detection and treatment of depression can help relieve this economic burden while increasing the productivity and quality of life of depressed individuals. However, treatment of depression is expensive and often delayed due to the scarcity of trained psychological clinicians and often the late diagnosis of mental disorder symptoms. Furthermore, the cost of early detection by either spot or large-scale screening is prohibitive due to the aforementioned reasons. Therefore, alternative technology-based screening methods have been sought in the form of inexpensive, automatic systems, to facilitate large-scale early detection and connect with timely intervention. The lack of effective depression screening candidate technologies has attracted research attention for more than a decade. To date, there have been several studies on the automatic detection of depression ranging from voice, facial video, EEG signals, head pose, eye gaze, etc. Among these modalities, video, text, and speech, which have demonstrated promising effectiveness and efficiency as an indicator of depression, remain notably non-invasive and easily accessible. In this project, we present a novel framework that invokes attention mechanisms at several layers to identify and extract important features from different modalities to predict the level of depression. The network uses several low-level and mid-level features from audio, text, and video modalities. However, most studies to date on speech-based depression detection have primarily focused on laboratory-collected data, recorded from a single channel in a clean environment. The increasing adoption of smartphones coupled with the emergence of voice assistants provide unprecedented opportunities for new automated medical screening methods through sampling the human voice the ability to accumulate a sufficiently large quantity of data to statistically model variations in speech patterns for depressed and non-depressed individuals across populations and audio recording devices type; and the ability to administer individually tailored questionnaires, analyze voice samples and provide clinical screening feedback across large populations. However, conventional features developed from clean lab-based datasets may not generalize as well in real-world applications due to the dramatic differences in speech recording such as noise conditions, handset hardware, and design protocols. This shortcoming motivates the design of a new category of effective features for detecting depression in both environments. Although the inherent relationship between verbal content and mental illness level is more prominent, the visual features also play a pivotal role to reinstate the deep association of depression to facial emotions.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.125123

It has been observed that patients suffering from depression often have distorted facial expressions e.g., eyebrow twitching, dull smile, frowning faces, aggressive looks, restricted lip movements, reduced eye blinks, etc. With the quantum of proliferating video data and the availability of high-end built-in cameras in wearables and surveillance sectors, analyzing facial emotions and sentiments is the growing trend amongst the vision community.

II. METHODOLOGY

1.Data collection:
Video recordings of individuals are collected using a camera or other recording device. The recordings may be made in a controlled environment or in a naturalistic setting.

2.Facial landmark feature extraction:
The facial landmark features of the individuals are extracted using computer vision techniques. This involves identifying specific points on the face, such as the position of the eyebrows, eyes, nose, mouth, and chin. The position of these points is tracked over time to capture changes in facial expressions.

3.Head movement extraction:
The head movements of the individuals are extracted using computer vision techniques. This involves tracking the movement of the head, such as nodding or shaking, as well as the orientation of the head.

4.Data pre-processing:
The extracted facial landmark features and head movements are preprocessed to prepare them for input into the LSTM model. This may involve normalization, scaling, or other data transformations.

5.LSTM model training:
An LSTM model is trained using the preprocessed facial landmark features and head movements as input. The model is trained using a labeled dataset of individuals with and without depression.

6.Model evaluation:
The performance of the LSTM model is evaluated using metrics such as accuracy, precision, recall, F1 score, and AUC of the ROC curve.

Depression:
Depression is a complex mental health condition that can have a significant impact on a person's quality of life. Deep learning, a subset of machine learning, has shown promise in predicting and diagnosing depression based on various factors.

Deep learning
Depression is a mental health condition that can have a significant impact on a person's life, and deep learning has shown potential for identifying and predicting depression based on various factors. One approach to using deep learning for depression is to analyze patterns in neuron imaging data. Studies have used deep learning algorithms to analyze functional magnetic resonance imaging (fMRI) data to predict depression. By analyzing changes in brain activity, these algorithms can identify patterns that are indicative of depression. Another approach is to use deep learning to analyze speech patterns and language use. People with depression often exhibit changes in the way they speak, such as slower speech or a reduction in the use of positive words.

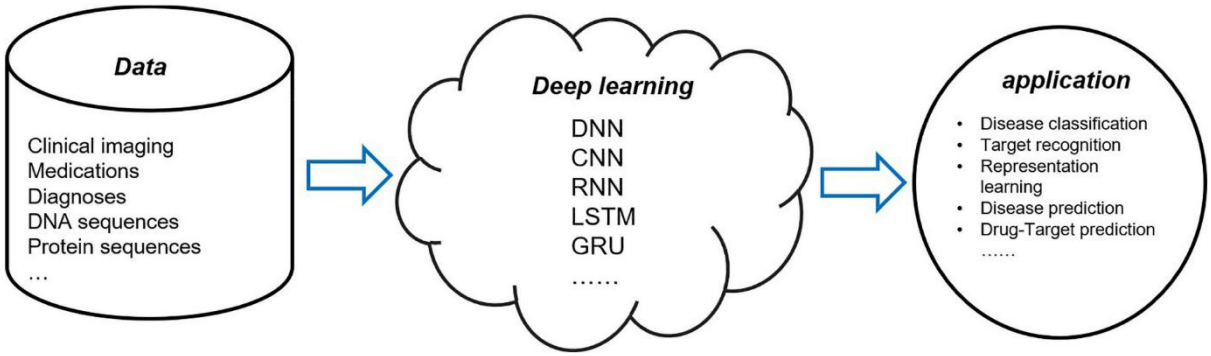
Deep learning algorithms can be trained to identify these patterns in speech and predict the likelihood of depression. Furthermore, deep learning can be used to analyze social media data to detect signs of depression. People often express their emotions and thoughts on social media, and these can be analyzed using natural language processing techniques and sentiment analysis. By analyzing social media data, deep learning algorithms can identify patterns that are indicative of depression, such as negative emotions and social isolation. In addition to these approaches, deep learning can also be used to develop personalized treatment plans for people with depression. By analyzing a person's medical history, lifestyle, and other factors, deep learning algorithms can help healthcare professionals develop individualized treatment plans that are tailored to the specific needs of each person.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125123

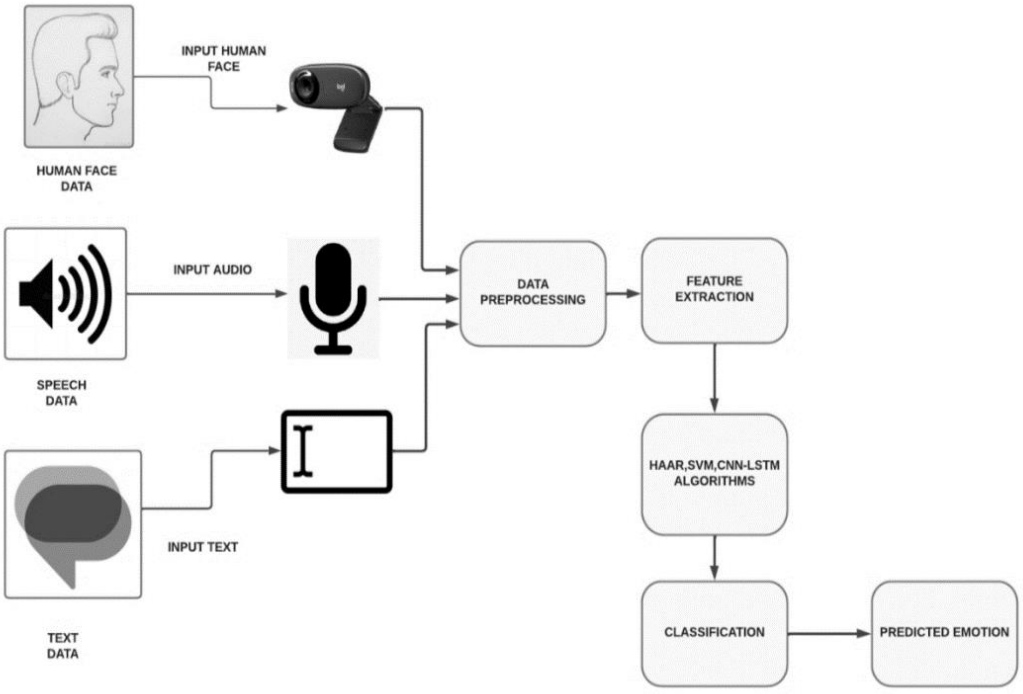


Long And Short Term Model (LSTM)

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that is well-suited for modeling sequential data. LSTM has been applied in depression detection by analyzing speech patterns and language use. The use of LSTM for depression detection involves analyzing speech features such as pitch, duration, and intensity. LSTM can be trained on speech data to identify patterns that are indicative of depression, such as long pauses between words, slower speech rate, and lower pitch. Additionally, LSTM can be trained to identify specific words and phrases that are associated with depression, such as negative emotions and feelings of hopelessness.

Furthermore, LSTM can be used to analyze social media data to detect signs of depression. By analyzing social media posts and messages, LSTM can identify patterns that are indicative of depression, such as negative emotions and social isolation. Overall, LSTM has shown promise in depression detection by analyzing speech and language use, as well as social media data. However, it is important to note that the effectiveness of LSTM in depression detection depends on the quality and quantity of the data used for training. Additionally, more research is needed to fully understand the potential of LSTM and other deep-learning techniques for depression detection and treatment.

III. MODELING AND ANALYSIS



Initially, data will collect in various forms from the person such as human face data, speech data, and text data as input. Secondly, the collected input data are pre-processed and then the feature extraction is done. Using haar, svm, and CNN-lstm algorithms the data are classified. Based on the classification the person's emotions are predicted.

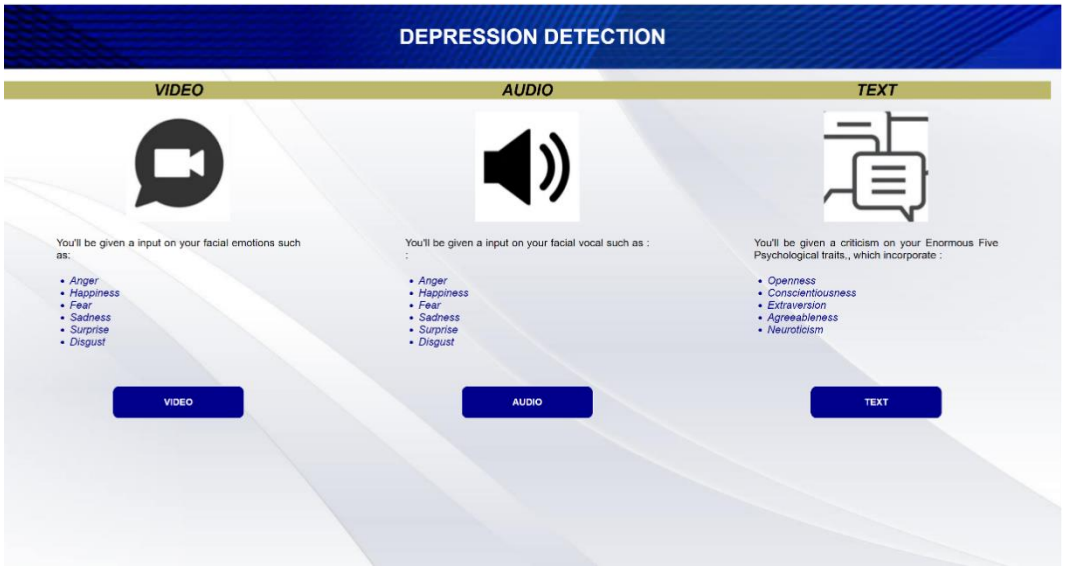
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

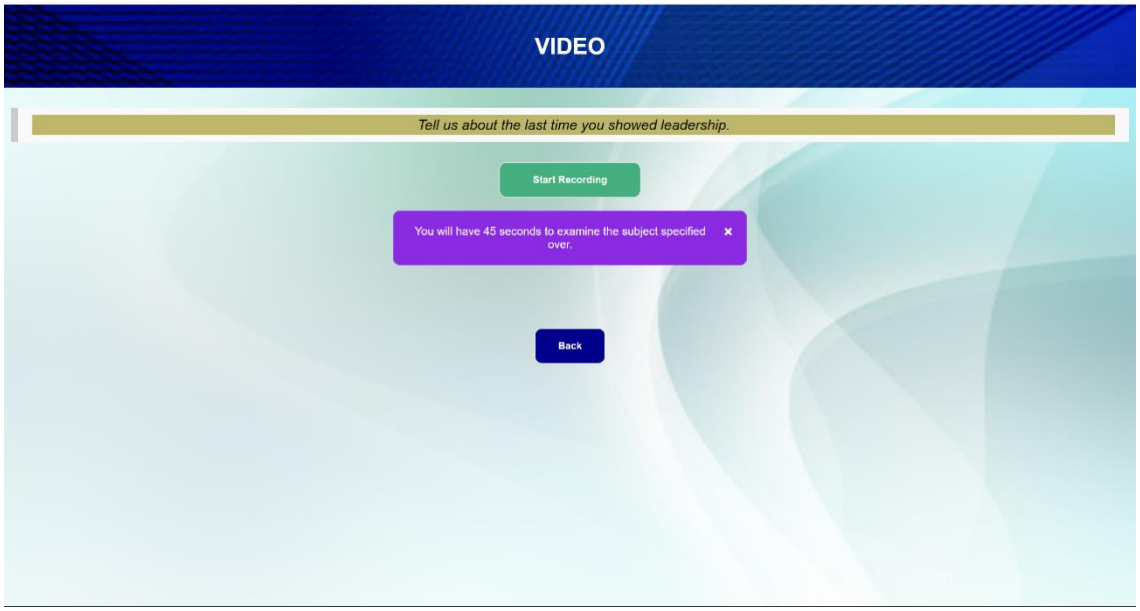
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125123

IV. RESULT AND DISCUSSION



The above figure shows the front view of the depression detection application. there are three types of classification ways to find depression using video, audio, and text. The technology used in this is deep learning and back end our Python. we use java script to interact with the python.



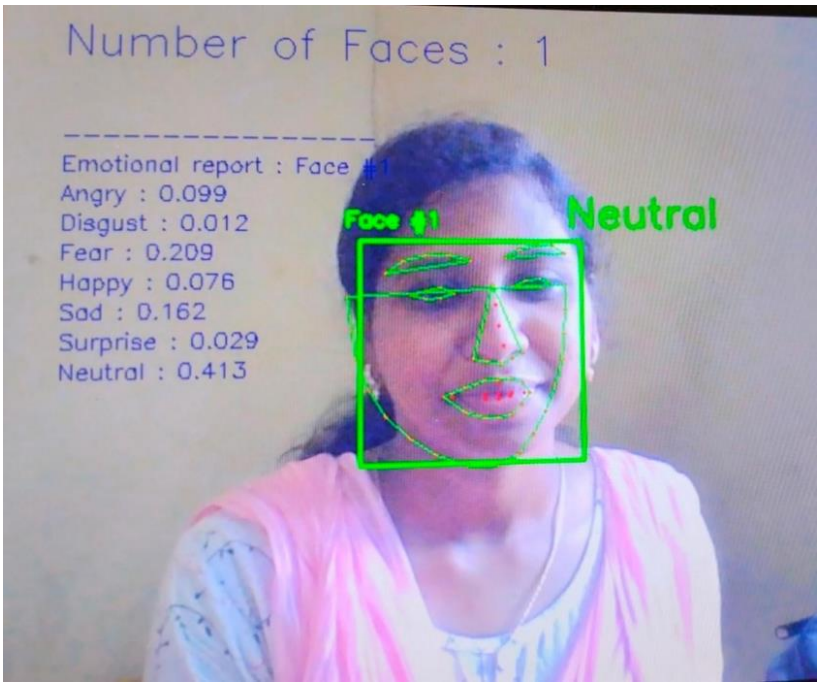
This is a video page. The facial landmark a is detected and takes the input us the face and predicted the depression level.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

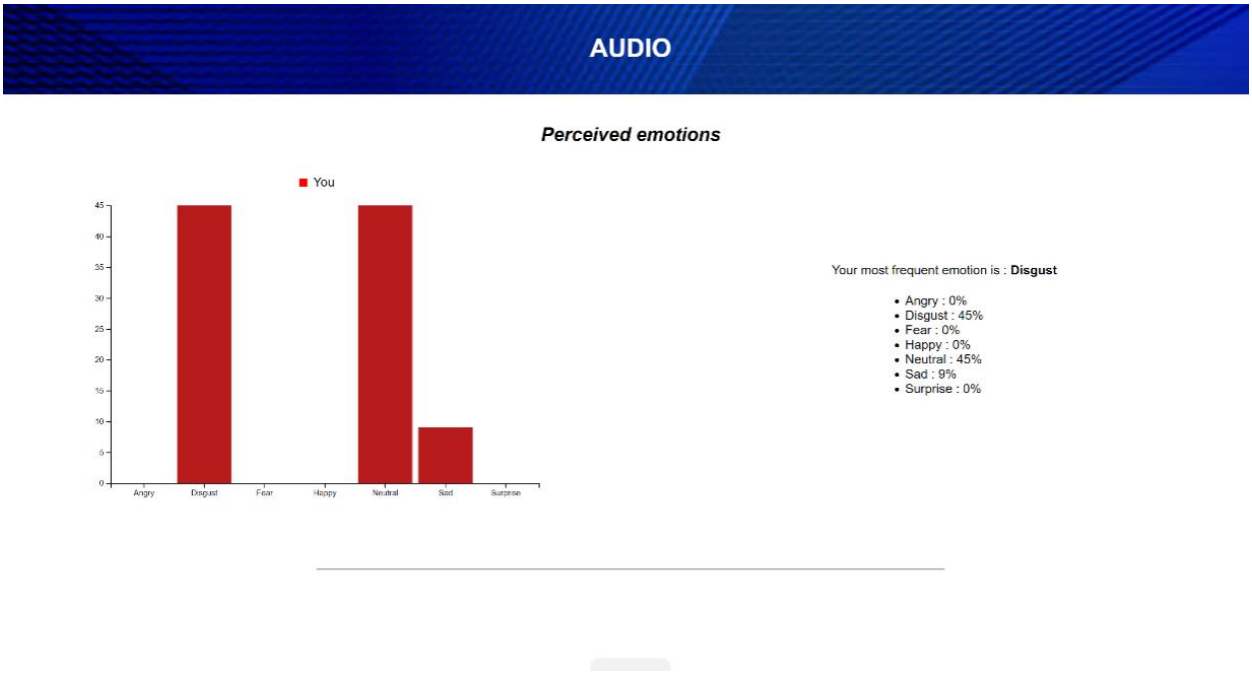
International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125123



In the above figure, the output of the depression detection using video is displayed. Users allow to know about the percentage of their happiness, disgust, fear, sadness, surprise, and neutrality.



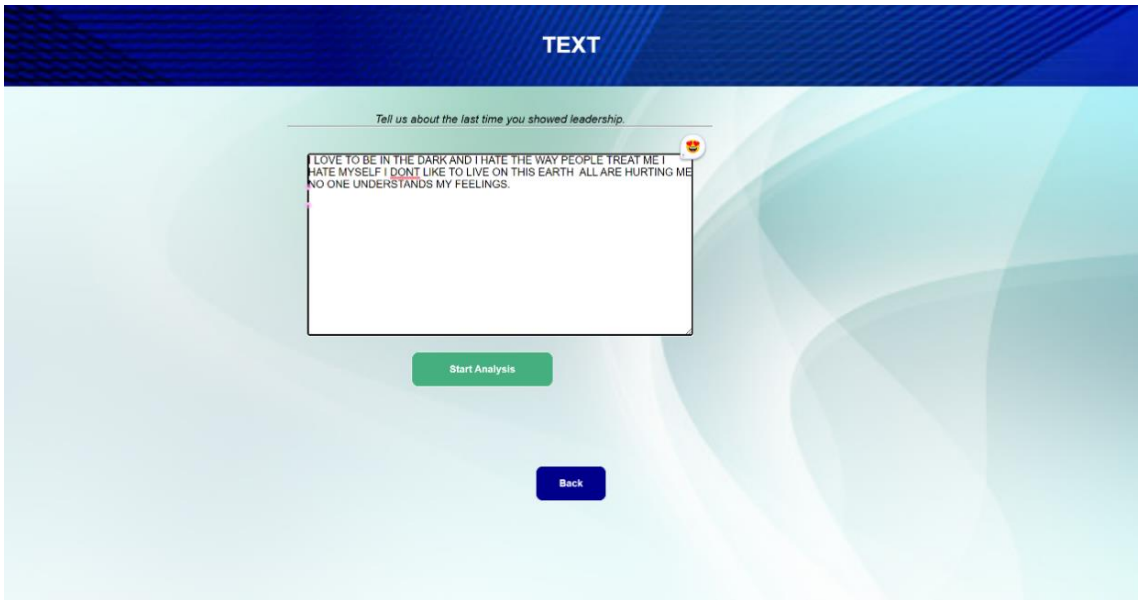
The above figure show that the output of the depression detection using audio is displayed. Users allow to know about the percentage of their angry, disgusted, fearful, happy, neutral, sad, and surprised.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

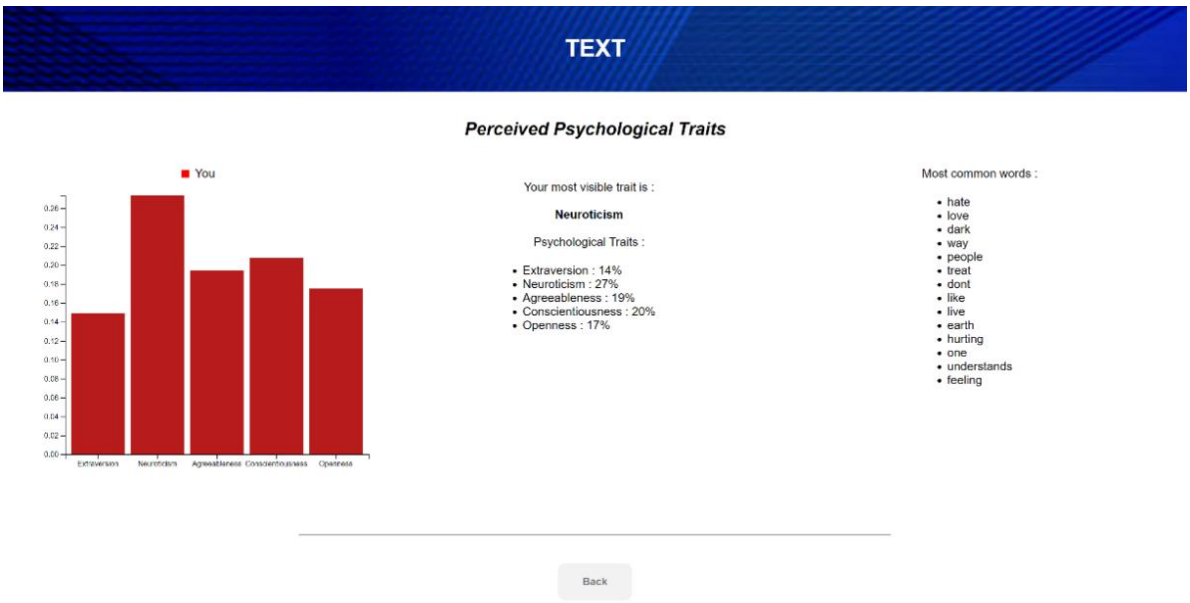
International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125123



The user can enter their text on this page and through this the depression of the person can be identified by clicking the start analysis the text that is entered by the user is verified and the output will be displayed.



This is the output of the text. Here we can see the percentage of the text and also the graphical representation of the text there are some classifications through which the depression of the person can be detected extraversion, neuroticism, agreeableness, conscientiousness, and openness.

V. CONCLUSION

A multi-level attention-based early fusion network that fuses audio, video, and text modalities to predict the severity of depression. For this task, we observed that the attention network gave the highest weights to the text modality and almost equal weight age to audio and video modalities. The use of multi-level attention led us to obtain significantly better results in all individual and fusion models compared to both the baseline and state-of-art. Using attention to each feature and each modality had a twofold advantage overall. Firstly, this gives us a deep and better understanding of the importance of each feature within a modality toward depression prediction. Secondly, attention simplified the network's overall computational complexity and reduced the training and test time.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.125123

REFERENCES

[1]. M.-I. Georgescu, R. T. Ionescu, and M. Popescu,(2019) "Local learning with deep and handcrafted features for facial expression recognition," IEEE Access, vol. 7, pp. 64827–64836.
[2]. K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, (2019)"Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp. 5866–5870.
[3]. D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou,(2019)"Deep affect prediction inthe-wild: Aff-wild database and challenge, deep architectures, and beyond," Int. J. Comput. Vis., vol. 127, pp. 1–23.
[4]. D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, (2020)"Analysing affective behavior in the first ABAW 2020 competition".
[5]. Z. Du, S. Wu, D. Huang, W. Li, and Y. Wang,(2019) "Spatio-temporal encoderdecoder fully convolutional network for video-based dimensional emotion recognition," IEEE Trans. Affect. Comput., early access.
[6]. P. Barros, N. Churamani, and A. Sciutti, (2020)"The FaceChannel: A light-weight deep neural network for facial expression recognition," in Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), Buenos Aires, AR, USA.
[7]. M. Koujan, L. Alharbawee, G. Giannakakis, N. Pugeault, and A. Roussos,(2020)"Real-time facial expression recognition 'in the wild' by disentangling 3D expression from identity," in Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), Buenos Aires, AR, USA.
[8]. G. Zhao, H. Yang, and M. Yu,(2020) "Expression recognition method based on a lightweight convolutional neural network," IEEE Access, vol. 8, pp. 38528–385.