International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Audio and Video Based Emotion Detection

R.Mohamed Yousuf , Dr.T.C.Subbulakshmi

Student(IT) – Professor(IT)

Francis Xavier Engineering College, Tirunelveli, India

ABSTRACT

The 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, facial expression and text. In response to the pressing need for accurate depression detection, researchers in affective computing have turned to behavioral cues such as facial expressions and prosodic features from speech to predict mental disorders like depression and Post-traumatic Stress Disorder. This project introduces a novel framework that harnesses attention mechanisms across multiple layers to identify and extract crucial features from various modalities, facilitating the prediction of depression levels. Leveraging low-level and mid-level features from text, audio, and video data, the proposed network employs attention mechanisms at different levels to quantify the significance of each feature and modality, resulting in enhanced prediction performance. Through a series of meticulous experiments involving individual features from diverse modalities as well as their combinations, the study showcases the efficacy of the approach. Additionally, the project addresses the challenge of identifying effective depression-related features in adverse recording conditions and across different smartphone devices. Introducing two innovative sets of features rooted in speech landmarks, the study achieves promising outcomes on disparate datasets. Employing a blend of machine learning and deep learning algorithms, this research contributes to advancing accurate depression detection by fusing behavioral cues from multiple sources.

I. INTRODUCTION

Depression, a major mental disorder reported to dafflict 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 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 a number of studies on automatic detection of depression ranging from voice, facial video, EEG signals, head pose, eye gaze, etc. Among these modalities, video, text, speech, which has demonstrated promising effectiveness and efficiency as an indicator of depression , remains notably non- invasive and easily accessible. In this project, we present a novel framework that invokes attention mechanism at several layers to identify and extract important features from different modalities to predict 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 individual 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 short coming motivates the design of a new category of effective features for detecting depression under 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. It has been observed that patients suffering from depression often have distorted facial expressions for 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 availability of high end built-in cameras in wearables and surveillance sectors, analyzing the facial emotions and sentiments is the growing trend amongst the vision community.

II. LITERATURE SURVEY

2.1 Shaoxiong Ji, Xue Li, Zi Huang, Erik Cambria (2021) Suicidal Ideation And Mental Disorder Detection With Attentive Relation Networks.

Mental health is a critical issue in modern society, and mental disorders could sometimes turn to suicidal ideation without effective treatment. Early detection of mental disorders and suicidal ideation from social content provides a potential way for effective social intervention. However, classifying suicidal ideation and other mental disorders is challenging as they share similar patterns in language usage and sentimental polarity. This paper enhances text representation with lexicon-based sentiment scores and latent topics and proposes using relation networks to detect suicidal ideation and mental disorders with related risk indicators. The relation module is further equipped with the attention mechanism to prioritize more critical relational features. Through experiments on three real-world datasets, our model outperforms most of its counterparts.

2.2 Seyed Habib Hosseini-Saravani et al., (2020) Depression Detection In Social Media Using A Psychoanalytical Technique For Feature Extraction And A Cognitive Based Classifier

Depression detection in social media is a multidisciplinary area where psychological and psychoanalytical findings can help machine learning and natural language processing techniques to detect symptoms of depression in the users of social media. In this research, using an inventory that has made systematic observations and records of the characteristic attitudes and symptoms of depressed

patients, we develop a bipolar feature vector that contains features from both depressed and non-depressed classes. The inventory we use for feature extraction is composed of 21 categories of symptoms and attitudes, which are primarily clinically derived in the course of the psychoanalytic psychotherapy of depressed patients, and systematic observations and records of their characteristic attitudes and symptoms. Also, getting insight from a cognitive idea, we develop a classifier based on multinomial Naïve Bayes training algorithm with some modification. The model we develop in this research is successful in classifying the users of social media into depressed and non-depressed groups, achieving the F1 score 82.75 %

2.3 Erik Cambria et al.,(2022) SenticNet 7:A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis

In recent years, AI research has demonstrated enormous potential for the benefit of humanity and society. While often better than its human counterparts in classification and pattern recognition tasks, however, AI still struggles with complex tasks

that require commonsense reasoning such as natural language understanding. In this context, the key limitations of current AI models are: dependency, reproducibility, trustworthiness, interpretability, and explainability. In this work, we propose a commonsense-based neurosymbolic framework that aims to overcome these issues in the context of sentiment analysis. In particular, we employ unsupervised and reproducible subsymbolic techniques such as auto-regressive language models and kernel methods to build trustworthy symbolic representations that convert natural language to a sort of protolanguage and, hence, extract polarity from text in a completely interpretable and explainable manner

2.4 Janardan Misra (2020),Auto NLP: NLP Feature Recommendations For Text Analytics Applications

While designing machine learning based text analytics applications, often, NLP data scientists manually determine which NLP features to use based upon their knowledge and experience with related problems. This results in increased efforts during feature engineering process and renders automated reuse of features across semantically related applications inherently difficult. In this paper, we argue for standardization in feature specification by outlining structure of a language for specifying NLP features and present an approach for their reuse across applications to increase likelihood of identifying optimal features

2.5 Shaoxiong Ji et al.,(2020) Suicidal Ideation Detection: A Review Of Machine Learning Methods And Applications

Suicide is a critical issue in modern society. Early detection and prevention of suicide attempts should be addressed to save people's life. Current suicidal ideation detection (SID) methods include clinical methods based on the interaction between social workers or experts and the targeted individuals and machine learning techniques with feature engineering or deep learning for automatic detection based on online social contents. This article is the first survey that comprehensively introduces and discusses the methods from these categories. Domain-specific applications of SID are reviewed according to their data sources, i.e., questionnaires, electronic health records, suicide notes, and online user content. Several specific tasks and data sets are introduced and summarized to facilitate further research. Finally, we summarize the limitations of current work and provide an outlook of further research direction

III. METHODOLOGY

3.1 EXISTING SYSTEM

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. Current speech processing methods typically segment speech into short 10-20 millisecond frames before extracting low-level descriptors. Recently, landmark-based features were shown to provide discriminative information for speech-based depression classification, particularly using relatively simple counts of consecutive landmark Diagrams.

3.1.1 DISADVANTAGE

- There are a few drawbacks to the well-structured frame-based approach. First, all frames are treated equally, which undermines the fact that some frames contain less information than others.
- Second, frame-based features such as spectral features are vulnerable to channel variability, especially for smartphone speech, which is commonly, collected various handset types.

3.2 PROPOSED SYSTEM

The huge need for depression detection and the challenges involved motivated the affective computing research community to use behavioral cues to learn to predict depression, Post-traumatic stress disorder, and related mental disorders. Behavioral cues such as facial expression, prosodic features from speech have proven to be excellent features for depression prediction in this project, we present a novel framework that invokes attention mechanism at several layers to identify and extract important features from different modalities to predict level of depression. The network uses several low-level and mid-level features from text, audio and video modalities. We show that attention at different levels gives us the ratio of importance of each feature and modality, leading to better results. We perform several experiments on each feature from different modality and combine several modalities. More precisely, we propose duration based features, which are statistics calculated from two types of bigrams. To cope with the challenges of finding 12 effective depression-related features, especially for degraded recording conditions and diverse smartphones, herein we proposed two novel, effective sets of features based on speech landmarks, which delivered promising results on two dramatically different datasets.

3.2.1 ADVANTAGES

- By leveraging multiple modalities, our system can capture a richer set of cues and information related to emotions and depression. Each modality contributes different facets of an individual's emotional state, leading to a more comprehensive understanding.
- Text, speech, and facial expressions can all provide contextual information that aids in understanding emotions and depression. The integration of these cues can help discern the underlying
- Different individuals express their emotions differently. Using multiple modalities enables the system to better adapt to individual variations and deliver more personalized predictions.

3.3 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

3.3.1 ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

3.3.2 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

3.3.3 SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.**H/W SYSTEM CONFIGURATION:**

- Processor – I3, i5,i7
- RAM - 8 Gb
- Hard Disk - 500 GB

4.1 S/W SYSTEM CONFIGURATION:

- Operating System - Windows 7/8/10
- Front End - Html,Css
- Scripts – python language
- Tool – Python idle

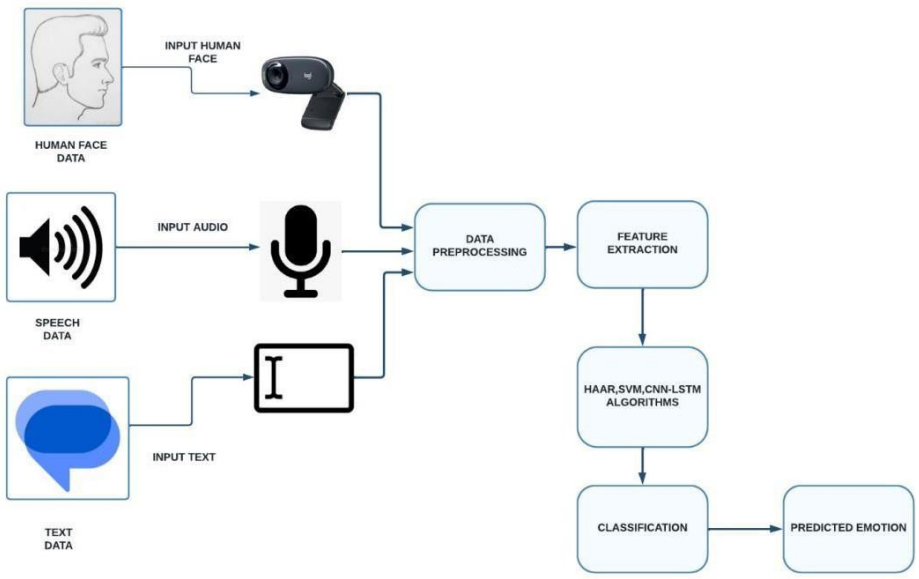


Fig 4.1 ARCHITECT_DIAGRAM

4.2 SOFTWARE DESCRIPTION

4.2.1 FRONT END:

The front end of the software has been developed using HTML and CSS, incorporating a user interface (UI) design. HTML (HyperText Markup Language) is utilized for structuring the content of web pages, while CSS (Cascading Style Sheets) is employed for styling and formatting the visual presentation of these pages. Together, they provide the foundation for creating the user interface that interacts with the end user.

4.2.2 WEB APPLICATION:

The Depression Detection web application is a powerful tool designed to assist users in identifying symptoms and patterns associated with depression. Developed using Flask, a lightweight Python web framework, this application offers a user-friendly interface that facilitates easy interaction and navigation.

At its core, the application leverages various algorithms and data analysis techniques to assess user inputs and provide personalized feedback. Users are prompted to answer questions or input information related to their mood, behavior, and experiences. Through careful analysis of this data, the application generates insights regarding the likelihood and severity of depression.

4.2.3 PROPOSED SYSTEM ALGORITHMS:

4.3.1 HAAR CASCADE ALGORITHM

The algorithm can be explained in four stages:

- Calculating Haar Features.
- Creating Integral Images.
- Using Adaboost.
- Implementing Cascading Classifiers.

HAAR CASCADE ALGORITHM

```
import numpy as np import cv2

f_cascade = cv2.CascadeClassifier("face.xml") e_cascade = cv2.CascadeClassifier("eye.xml")

image = cv2.imread("actor.jpg")

gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)

faces = f_cascade.detectMultiScale(gray, 1.3, 5) for (x,y,w,h) in faces:

img = cv2.rectangle(img,(x,y),(x+w,y+h),(255,0,0),2) roi_gray = gray[y:y+h, x:x+w]
```

roi_color = img[y:y+h, x:x+w]

eyes = e_cascade.detectMultiScale(roi_gray) for (ex,ey,ew,eh) in eyes:

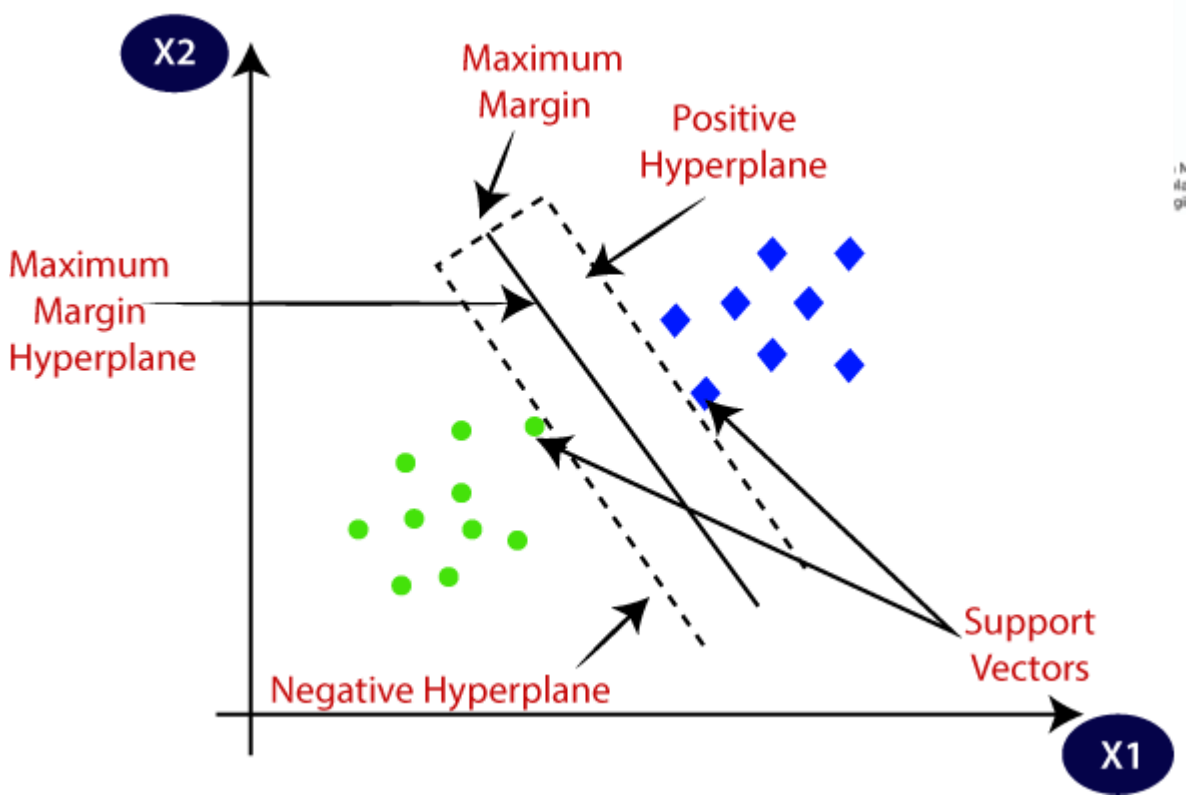
cv2.rectangle(roi_color,(ex,ey),(ex+ew,ey+eh),(0,255,0),2) cv2.imshow('img',image)

cv2.waitKey(0) cv2.destroyAllWindows()

4.3.2 SUPPORT VECTOR MACHINE(SVM)

SVM or support vector machine is the classifier that maximizes the margin. The goal of a classifier in our example below is to find a line or (n-1) dimension hyper-plane that separates the two classes present in the n-dimensional space.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



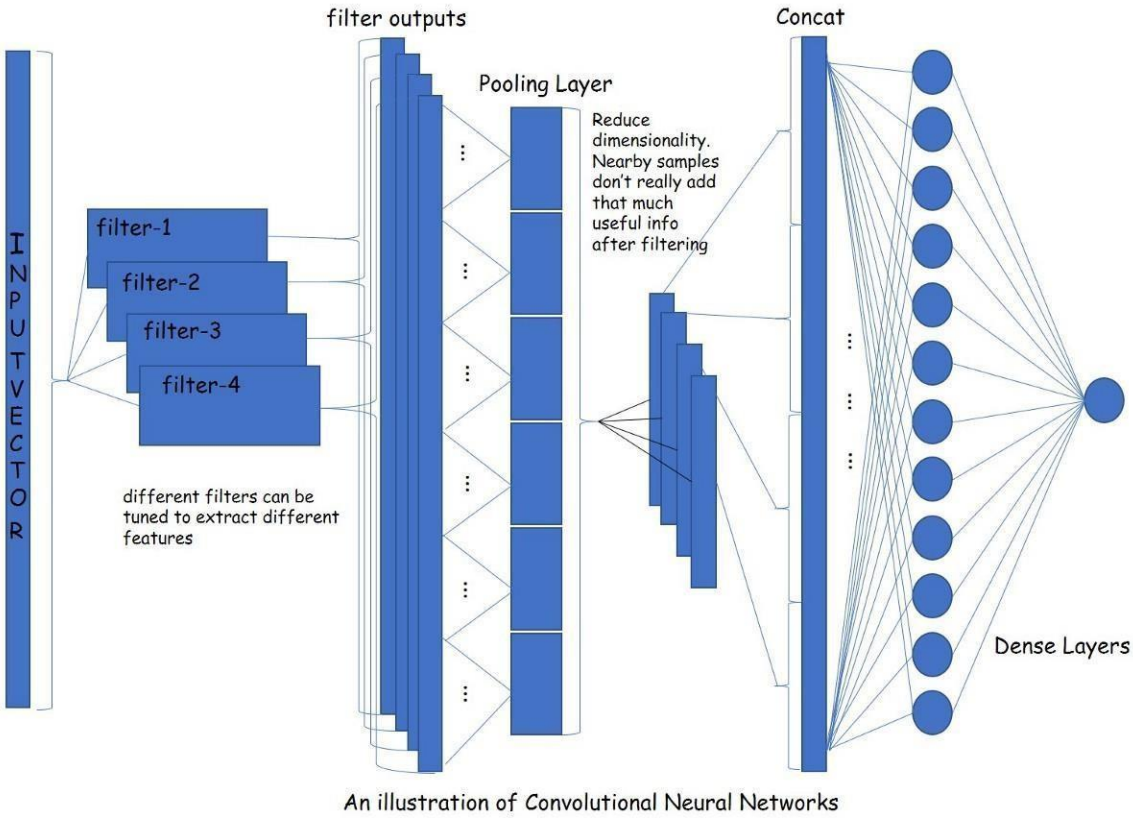
TYPES OF SVM

SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non- linear SVM classifier.

- **5.3.3 CONVOLUTIONAL NEURAL NETWORK (CNN) ALGORITHM**



An illustration of Convolutional Neural Networks

- I will be using Sequential method as I am creating a sequential model. Sequential model means that all the layers of the model will be arranged in sequence. Here I have imported Data Generator from keras preprocessing. The objective of Image Data Generator is to import data with labels easily into the model. It is a very useful class as it has many function to rescale, rotate, zoom, flip etc. The most useful thing about this class is that it doesn't affect the data stored on the disk. This class alters the data on the go while passing it to the model.

- function CONV(T_X, Ty, C_i) T_X □ 0, Ty □0,C_i □ 0 while T_y < Y do

- while T_x < X do

- while C_i< I do

- if T_x =0 and T_y = 0 then

c□0

READTILE(I Buf., Tx, Ty, Ci)

end if

|| Convop(Ci, [Buf.)

|| PREFETCHTILE(IBufc., Tx, Ty, Ci + 1

C □c end while

end while end while

end function.

CNN THEOREM:

There is a simple formula to do so: Dimension of image = (n, n) Dimension of filter = (f,f) Dimension of output will be ((n-f+1) , (n-f+1)) You should have a good understanding of how a convolutional layer works at this point. Let us move to the next part of the CNN architecture.

$$W_{out} = \frac{W - F + 2P}{S} + 1$$

4.2.4 LONG SHORT-TERM MEMORY ALGORITHM

LSTM (Long Short-Term Memory) is a type of Recurrent Neural Network (RNN) architecture specifically designed to handle the issue of vanishing gradients in traditional RNNs. It is commonly used in tasks involving sequence data such as speech recognition, text generation, language translation, etc. LSTMs introduce memory cells and gates that regulate the flow of information into and out of the cells, enabling the network to selectively remember and forget information over a longer period of time.

Pseudo code representation of an LSTM algorithm:

for each time step t:

input = previous hidden state and current input

forget gate = sigmoid(Wf * [input, h_t-1] + bf) input gate = sigmoid(Wi * [input, h_t-1] + bi) candidate memory = tanh(Wc * [input, h_t-1] + bc)

memory = forget gate * memory_t-1 + input gate * candidate memory output gate = sigmoid(Wo * [input, h_t-1] + bo)

hidden state = output gate * tanh(memory)

In this pseudo code, 'Wf', 'Wi', 'Wc', and 'Wo' are the weight matrices for the forget, input, candidate memory, and output gates, respectively. bf, bi, bc, and bo are the bias terms. input is the input at time step t and h_t-1 is the previous hidden state. The final output of the LSTM at each time step is the hidden state, h_t.

ARCHITECTURE OF LSTM-RNN

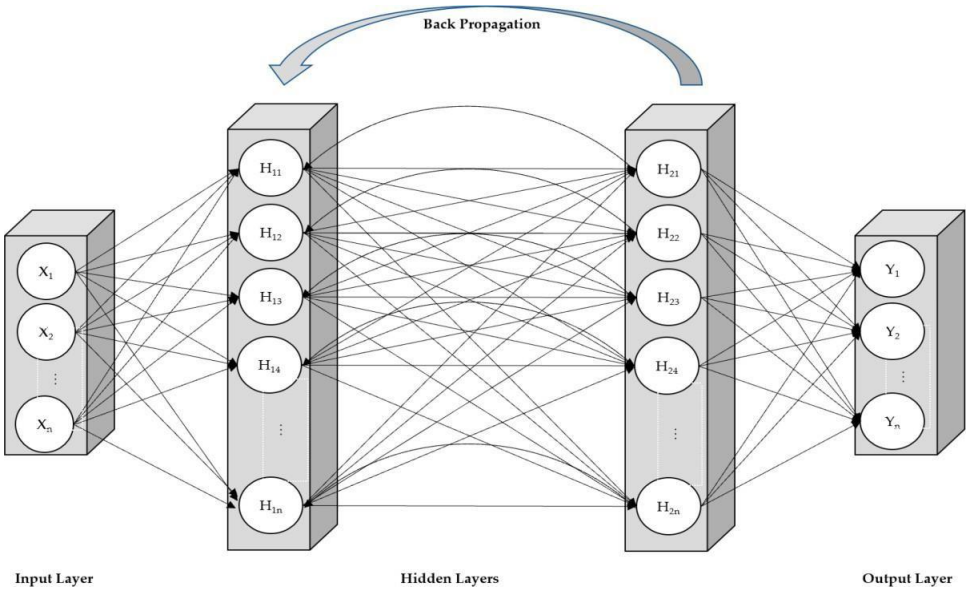


Fig 4.4 LSTM-RNN ARCHITECTURE

LSTMs deal with both Long Term Memory (LTM) and Short Term Memory (STM) and for making the calculations simple and effective it uses the concept of gates.

Forget Gate: LTM goes to forget gate and it forgets information that is not useful.

Learn Gate: Event (current input) and STM are combined together so that necessary information that we have recently learned from STM can be applied to the current input.

Remember Gate: LTM information that we haven't forget and STM and Event are combined together in Remember gate which works as updated LTM.

Use Gate: This gate also uses LTM, STM, and Event to predict the output of the current event wh**VII. FUTURE SCOPE**

The sign language recognition system has significant potential for expansion and improvement. As technology advances, several future developments can enhance its functionality:

Support for Additional Sign Languages – The system can be trained to recognize more sign languages, making it inclusive for a wider range of users.

Enhanced Accuracy and Efficiency – Further advancements in deep learning models could improve recognition accuracy and reduce processing time.

Mobile and Smart Device Integration – Making the system accessible through mobile applications and smart devices would provide users with greater flexibility and convenience.

Advanced Data Analytics – The Power BI dashboard could be enhanced with more detailed analytics, helping developers understand user patterns and improve system performance.

Customization and Interactive Feedback – Allowing users to personalize their experience and provide real-time feedback could make the system more user-friendly and adaptable to individual needs.

VI. FUTURE WORK

A multi-level attention based early fusion network which fuses audio, video and text modalities to predict severity of depression. For this task we observed that the attention network gave highest weights to the text modality and almost equal weightage 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 over each feature and each modality had a twofold advantage overall. Firstly, this gives us deep and better understanding of importance of each feature

within a modality towards depression prediction. Secondly, attention simplified the network's overall computational complexity and reduced the training and test time.

For future works, we will use the system and find issues in order to improve either system architecture, algorithms for a component or library used. Also, we encourage the use of novel methods in one or more of our proposed system components. We also encourage researchers to retrain the proposed CNN model with other datasets to improve its accuracy rate.Some further work is required to be done to optimize the model to achieve a greater percentage of efficiency. The dataset also needs to be expanded with a lot of samples of individuals, it also needs to be perfectly balanced as well as extra features like BDI reports could be added, and needs to be tried out on complex CNN architecture.

V REFERENCE

[1]. E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok,(2020) "SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis," in Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.

[2]. Gardini, E., Cavalli, A., and Decherchi, S. (2021). An ab initio local principal path algorithm. In International Joint Conference on Neural Networks, Accepted paper .

[3]. S. H. Hosseini-Saravani, S. Besharati, H. Calvo, and A. Gelbukh,(2020) "Depression detection in social media using a psychoanalytical technique for feature extraction and a cognitive based classifier," in Proc. Mex. Int.

[4]. S. Ji, X. Li, Z. Huang, and E. Cambria,(2021) "Suicidal ideation and mental disorder detection with attentive relation networks,", pp. 1–11.

[5]. Lei Cao, Huijun Zhang, and Ling Feng(2021). Building and using personal knowledge graph to improve suicidal ideation detection on social media. IEEE Transactions on Multimedia.

[6]. S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang,(2020) "Suicidal ideation detection: A review of machine learning methods and applications," IEEE Trans. Computat. Social Syst., vol. 8, no. 1, pp. 214–226.

[7]. Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang(2021). Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems, 8:214–226.

[8]. Yang, L., Li, J., Cunningham, P., Zhang, Y., Smyth, B., and Dong, R. (2021). Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. In Proceedings of ACL/IJCNLP, pages 306–316.

[9]. Yan, H., Dai, J., Ji, T., Qiu, X., and Zhang, Z. (2021). A unified generative framework for aspect-based sentiment analysis. In Proceedings of ACL/IJCNLP, pages 2416–2429.

\