Multimodal Sentiment Analysis Model using Machine Learning

Archit Aggarwal[1], Akash Kumar[2], Ankesh Patel[3]

[1]Student at VIT University, Vellore Pursuing Bachelor's in Computer Science and Engineering
[2]Student at VIT University, Vellore Pursuing Bachelor's in Computer Science and Engineering
[3]Student at VIT University, Vellore Pursuing Bachelor's in Electronics and Communication Engineering

---***---

Abstract - *The paper presents a model to perform sentiment analysis by three mediums – video, audio, and text to predict their sentiments in different categories. Comparative analysis of various machine learning models used to perform the study is also presented in the paper. The model can help in preparing personnel for interviews. This paper presents a model for users to assess themselves based on our sentimental analysis models.*

Key Words: **audio, interviews, machine learning, model, predict, sentiments, text**

1.INTRODUCTION

The model consists of three separate modules to perform sentiment analysis in three mediums - audio, video, text. The data is captured and processed using NLP and OpenCV to detect the user's performance concerning other users and also assess the sentiments exhibited by the user. The architecture consists of three separate modules to perform sentiment analysis in three mediums - audio, video, text. The data is captured and processed using NLP and OpenCV to detect the user's performance concerning other users and also assess the sentiments exhibited by the user.

The machine learning algorithm used will be for video and audio, and textual analysis. The Algorithms used will be chosen in such a way as to avoid overfitting and give accurate results to the users.

The paper begins with an introduction to the datasets used to perform the analysis and the machine learning algorithms used. After that, we explained the methodology and the modeling design. Outcomes are then compared for different modes at the end.

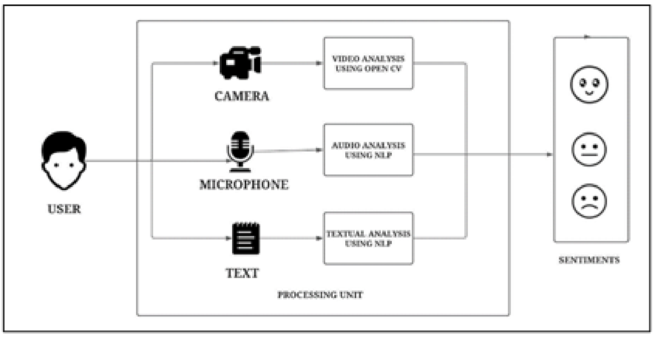

Fig -1: Architecture Diagram

2. LITERATURE SURVEY

In 2020, A. Ishaq et al. [1] provides a method to perform aspect-based sentiment analysis by integrating three steps: Semantic feature mining, transformation using Word2vec, and CNN for opinion mining. The technique proposed exhibited high accuracy and precision rate for unstructured text data.

In 2018, Pontes et al. [2] proposed a system combining recurrent and convolution networks to perform semantic textual similarity on text data. The proposed method works on two levels: Local and global. CNN analyzes the relevance of a word while LSTM analyzes the entire sentence.

In 2021, Syed Aley Fatima et al. [3] explored the use of Xception and CNN to identify facial emotions and was able to achieve better results if compared with the previous works. FER 2013 dataset was used. A real time vision system was made which gives you the life emotions (by evaluating facial emotions) of the user. For example, whether he/she is happy, sad, angry or neutral.

In 2019, Yu Yong et al. [4] discusses the use of RNN for analysis of text, audio and video. However, low performance is observed due to the long gap between input data. LSTM with introduction of gate functions handle such cases well. LSTM is now used for deep learning applications. This study also presents the future research scope in various fields.

In 2021, Manna et al. [5] discusses different algorithms used for sentiment analysis for emotions. This is done through one month voice recording and shows the broad aspects of the currently used algorithm. They also focus on understanding the emotions and then implementing models

In 2018, Thomas M et al. [6] discusses how Recurrent Neural Network Models are one of the models of neural networks that do not rely on length about a window while a natural language processing situation is considered. Input is fed in batches, and the accuracy of the model is evaluated each time in batches. The methods of RNN-LSTM model sentimental Analysis and Deep Learning using RNN can also be used for the sentimental Analysis of other language domains and to deal with cross-linguistic problems.

In 2019, Min Hu et al. [7] spends significant time on the issue of acknowledgment of facial feeling articulations in video successions. It proposes an included structure of two networks: a neighbourhood organization and a worldwide organization, principally founded on nearby upgraded movement history picture (LEMHI) and CNN-LSTM fell organizations separately. In the nearby organization, outlines from the unnoticed video are accumulated solidly into a solitary edge through a particular strategy, LEMHI. Contrasted and cutting-edge strategies, the proposed system set up a prevalent presentation.

3. DATASET

The paper presents methods to perform sentiment analysis by three mediums - text, audio and visual.

For the textual analysis, we have used the "Stream of Consciousness" dataset. The dataset is a collection of essays portraying different sentiments and personality traits. The dataset is widely used for personality recognition by machine learning.

RAVDESS dataset has been used to predict audio sentiments in our model. The dataset contains recordings of 24 professional actors. The recording includes expressions like calm, happy, sad, etc., at two different pitch levels.

The video analysis model is based on the FER2013 dataset. The dataset consists of approximately 30,000 images exhibiting different emotions. It consists of 48x48 images, which are grayscale and centered, which reduces the pre-processing steps.

Fig -2: Sample images from the FER2013 dataset

4. ALGORITHMS USED

Support Vector Machine or SVM model is a machine learning model that uses a supervised method of learning. The SVM's purpose is to find the optimal decision boundary to classify data points into different, correct categories. The decision boundary that exhibits the most accuracy is known as the hyperplane.

Xception is a CNN model that is 71 layers deep. We can also load a pre-trained version of the network trained on more than a million images from the ImageNet database. The pre-trained network is strong enough to classify images into thousands of object categories, like football, dogs, cat, cars, etc. Therefore, this model can identify a lot of objects in the images. Hence, this model can identify a lot of objects in the images and therefore can be used to capture the facial expressions on the user's face while he/she is speaking.

The convolutional neural network (CNN) is a feedforward neural network, and it is also the most mature field of deep learning algorithm application. It has a strong characterization learning ability. The network structure consists of three layers: a convolutional, pooling, and fully connected layer. The convolutional neural network convolves the input word vector sequence, generates a feature map, and then uses max pooling on the feature map to get the feature corresponding to the kernel.

LSTM is a time-cycle neural network, an advanced version of RNN, which is now more widely used in industry. Before introducing LSTM, the recurrent neural network should be taught because LSTM is developed from RNN. In recent years, the recurrent neural network has been widely used in natural language processing, such as language analysis and text classification.

5. METHODOLOGY

5.1 Textual Analysis

The text-based personality recognition steps are done as follows:

1. Retrieving data from Text
2. Performing custom Natural Language Processing steps:
   - Performing tokenization of the entire text
   - Converting all tokens to lowercase
   - Removal of predefined English stopwords
   - Assignment of POS tags on the tokens remaining after stopwords removal
   - Perform lemmatization with POS to achieve higher accuracy
   - Token sequence of the data is padded to make input vectors
3. Word2Vector embedding which uses 300-dimension is used for mapping
4. Predict the sentiments using our model which is pre trained

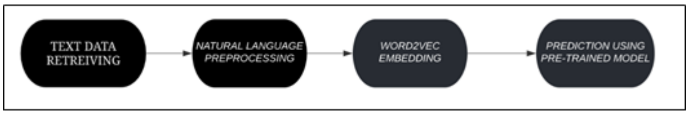
TEXT DATA RETREIVING
NATURAL LANGUAGE PREPROCESSING
WORD2VEC EMBEDDING
PREDICTION USING PRE-TRAINED MODEL

Fig -3: Textual Analysis Pipeline

5.2 Video Analysis

The pipeline used for video sentiment analysis was made up in the following ways:

1. Webcam is to be launched to capture the video.
2. Use Histogram (Orient gradients) to identify the human faces in the video.
3. Perform zoom on the face to capture sentiments (emotions).
4. Correct the dimensions to 48*48 pixels for processing.
5. Predict the sentiments using our model which is pre-trained.

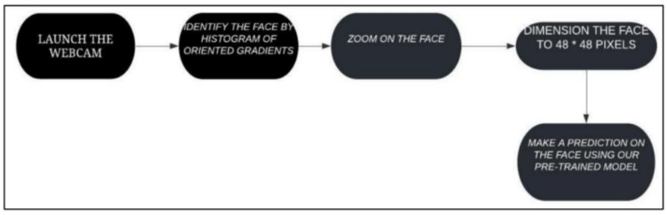
LAUNCH THE WEBCAM
DENTIFY THE FACE BY HISTOGRAM OF ORIENTED GRADIENTS
ZOOM ON THE FACE
DIMENSION THE FACE TO 48 * 48 PIXELS
MAKE A PREDICTION ON THE FACE USING OUR PRE-TRAINED MODEL

Fig -4: Video Analysis Pipeline

5.3 Audio Analysis

The pipeline for audio analysis is accomplished by the following structure and steps:

1. Recording the voice of the subject
2. Implementing audio signal discretization on the voice recorded
3. Performing extraction of long-Mel-spectrogram
4. Splitting the obtained spectrogram into various timeframes/rolling windows
5. Predict the sentiments using our model, which is pre-trained.

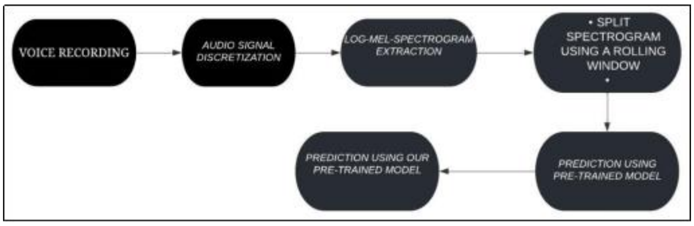
VOICE RECORDING
AUDIO SIGNAL DISCRETIZATION
LOG-MEL-SPECTROGRAM EXTRACTION
• SPLIT SPECTROGRAM USING A ROLLING WINDOW
PREDICTION USING PRE-TRAINED MODEL
PREDICTION USING OUR PRE-TRAINED MODEL

Fig -5: Audio Analysis Pipeline

6. MODELING DESIGN

6.1 Textual Model

In this paper, we have selected a neural network architecture consisting of both CNN and recurrent neural networks. The convolution layer performs feature extraction, which helps to identify repetitions and patterns in the text. The LSTM cell is used to grasp the sequential nature of the language. LSTMs are preferred because they can store selective information for a longer duration than regular neural networks.

Firstly, the final version consists of 3 blocks, such as the subsequent four layers performing different functions - 1D convolution layer, max pooling, spatial dropout, batch normalization. The number of convolution filters is doubled after each block. Finally, an utterly related layer of multiple nodes is added for classification.

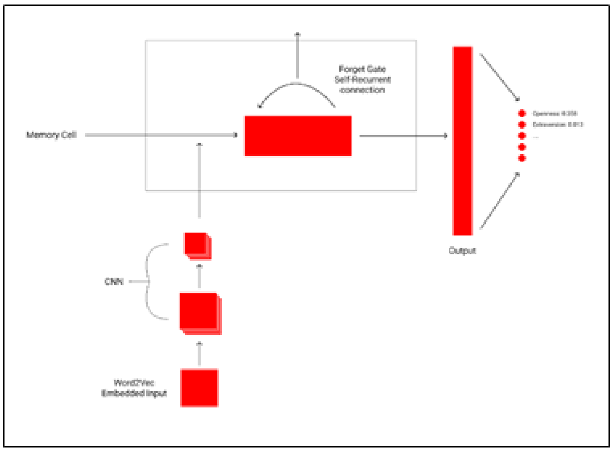

Fig -6: Textual Analysis Model

6.2 Video Model

The model we are using is XCeption model. The reason behind this was that it was more effective than the other models.

We tuned the model with the following:

1. Augmentation of Data
2. Early Stopping to prevent Overfitting
3. Reducing learning rate on the plateau
4. L2- Regularization
5. Balancing of Class weight

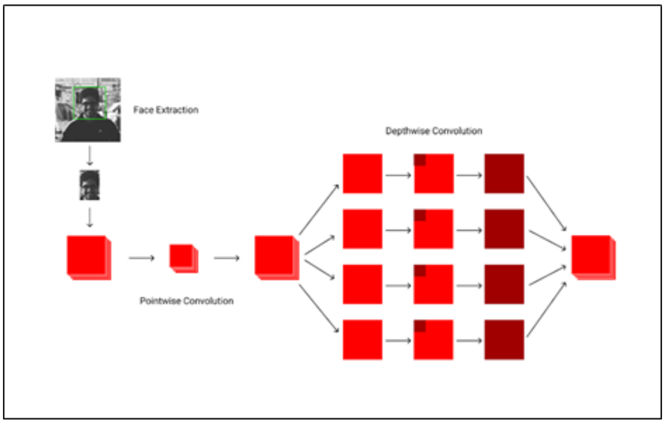

Fig -7: Video Analysis Model

6.3 Audio Model

The model uses a variation of the typical CNN, the Time Distributed CNN. It uses a rolling window (of constant time-step and length) along with the long-mel-spectrogram. The home windows with the help of several Local Feature Learning Blocks (LFLBs) can be used as the access of the network. The output of these CNNs is sent for further processing in a recurrent neural network made up of a duo of LSTM cells. In the end, a completely connected layer along with a SoftMax activation function is used to detect the sentiment in the voice.

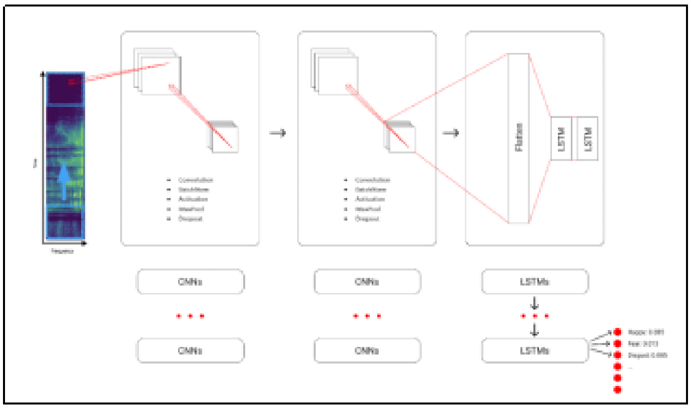

Fig -8: Audio Analysis Model

7. RESULT AND CONCLUSIONS

The above models exhibited correct predictions for all three methods. The results are summarized below.

Table.1. Audio Accuracy comparison between SVM and CNN

Model	Accuracy
SVM	68%
CNN	77%

From Table 1, it can be inferred that both the models provided high accuracy however, the time distributed CNN should be preferred because of the accuracy of 77%.

Table.2. Video Accuracy

Model	Accuracy
Xception	64.5%

Xception Model delivered reasonably correct results on the video tests. The model was trained on the FER2013 dataset and was able to predict sentiments from live video footage.

Table.3. Text Accuracy

Model	Accuracy
LSTM	68%

The LSTM Model along with recurrent neural network and 1D CNN was able to efficiently detect patterns, word repetitions in the text. The model was able to correctly predict the sentiments of the text in most cases through the use of punctuations and the word patterns in the text.

8. FUTURE SCOPE

The model can be improved and used as an interview-taking tool by recruiters to hire new employees. The model can help employers to hire employees with the right mindset. Nowadays, employers look for skills and need employees who have the right attitude towards the work, and our product can help them find one. The project can be altered for specific job-type questions, such as adding technical or creative questions according to the job. Furthermore, the system can perform psychometric analysis of various participants and track their progress in multiple fields such as athletics and education. It can use it to assess employee's workplace sentiments. The system is in its developmental stage but can be developed into an application readily available to be used on portable devices such as smartphones or tablets.

REFERENCES

[1] Ishaq, A., Asghar, S., & Gillani, S. A. (2020). Aspect-Based Sentiment Analysis Using a Hybridized Approach Based on CNN and GA. IEEE Access, 8, 135499–135512. https://doi.org/10.1109/access.2020.3011802

[2] Pontes, E. L., S. Huet, A. Linhares and J. Torres-Moreno. "Predicting the Semantic Textual Similarity with Siamese CNN and LSTM." CORIA-TALN-RJC (2018).

[3] Fatima, S. A., Kumar, A., & Raoof, S. S. (2021). Real Time Emotion Detection of Humans Using Mini-Xception Algorithm. IOP Conference Series: Materials Science and Engineering, 1042(1), 012027. https://doi.org/10.1088/1757-899x/1042/1/012027

[4] Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Computation, 31(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

[5] Manna, Debasmita & Baidya, Shaon & Bhattacharyya, S. (2021). Sentiment Analysis of Audio Diary. 10.1007/978-981-15-5546-6_9.

[6] Thomas, M., & C.A, L. (2018). Sentimental analysis using recurrent neural network. International Journal of Engineering & Technology, 7(2.27), 88. https://doi.org/10.14419/ijet.v7i2.27.12635

[7] Hu, M., Wang, H., Wang, X., Yang, J., & Wang, R. (2019). Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks. Journal of Visual Communication and Image Representation, 59, 176–185. https://doi.org/10.1016/j.jvcir.2018.12.039

BIOGRAPHIES

Archit Aggarwal is a Final Year Student at VIT University pursuing bachelor's in Computer Science and Engineering. Specializes in Machine Learning, Cloud Computing and Web Development.

Akash Kumar is a Final Year Student at VIT University pursuing bachelor's in Computer Science and Engineering. Specializes in Machine Learning and Cyber Security.

Ankesh Patel is a Final Year Student at VIT University pursuing bachelor's in Electronics and Communication Engineering. Specializes in Machine Learning and IoT.