Grenze International Journal of Engineering and Technology, June Issue

Video Sentiment Analysis: A Fusion Approach using CNN and LSTM Techniques

Anagha Bhamare[1], Rushali A.Deshmukh[2], Aditya Jadhav[3] and Vaishnavi Amati[4]
[1-4]Department of Computer Engineering, JSPM'S Rajarshi Shahu College of Engineering, Pune, India
Email: anaghabhamare17@gmail.com, rushalideshmukh78@gmail.com, vaishnaviamati2507@gmail.com, adijadhav1920@gmail.com, vaishnaviamati2507@gmail.com

***Abstract*—For the past few decades, textual sentiment analysis leveraging text mining techniques has been responsible for the majority of dedication done on sentiment analysis. However, there is still much to learn about the topic of audio sentiment analysis. In the proposed study, we use speaker differentiated voice transcripts to perform sentiment analysis to identify the feelings of the various speakers. To identify effective algorithms for this task, we examined various methods for sentiment analysis and speaker discrimination. We started out by reviewing the approaches that various academics have proposed for sentiment analysis across all data modalities. Then, we made an effort to put some of those tactics into action. We have focused on audio sentiment analysis in this work. New methods are being employed to analyze audio data as a result of the ongoing research into audio sentiment analysis. Here, we try to utilize machine learning to divide audio into several emotions. In this work, audio sentiment was explored in addition to multimode analysis. The categorization of emotions utilizing a range of data formats, such as text, audio, and video, is a component of multimodal sentiment analysis. Our method also yielded reasonable outcomes when merging the individual modality with attention networks for audio sentiment analysis. The aim was to build a system that can identify six various emotions in an audio recording, including anger, pleasure, disgust, sadness, fear, and surprise, when data is fed into it**

***Index Terms*— Convolutional Long Short-Term Memory (CLSTM), Convolutional Neural Network (CNN), OpenCV, Opinion, Visual Sentiment Analysis (VSA).**

I. INTRODUCTION

Studying people's feelings and behaviors in reaction to events, conversation topics, or life in general is known as sentiment analysis. Sentiment analysis is employed in many applications, and in this case, we use it to understand human thought processes based on their interactions with one another. In many situations, it may be really helpful to understand how people feel For example, computers are able to identify and respond to non-lexical human communication, including emotions. In such a scenario, the computer may adjust the settings in accordance with the wants and preferences of the user after identifying the human's emotions. Due to its strong correlation with an individual's thoughts and emotions, one of the most touchiest academic topics is psychological well-being. Every day, more people use communal media sites like Instagram, Facebook, Flickr, and others, with photos and videos becoming more and more significant. These days, our facial expressions might reveal our feelings. By seeing one other's facial expressions, we may infer each other's moods. Sentiment analysis is essential to simplifying and enhancing the effectiveness of this recognition. We will evaluate

Grenze ID: 01.GIJET.10.2.622_1

"sentiment," which is short for "emotions," using the Sentiment evaluation System. Our goal is video-based sentiment prediction, while most earlier investigation has concentrated on sentiment analysis using text. While sentiment analysis may be addressed in a variety of ways by scholars in the domain of pattern extraction and NLP, the social networking environment poses a number of particular challenges. In addition to the enormous amounts of data that are shared, most spoken conversations in virtual communities are brief and casual. Furthermore, even on the most well-known social media platforms, people are increasingly representing themselves via photos and videos in addition to spoken words. The information contained in these video pictures is related to impact and emotion signals sent by the picture, in addition to semantic elements like objects or activities in the acquired image. Therefore, such information is crucial for figuring out the emotional impact in addition to the semantic. Consequently, images and videos are among the most popular means by which individuals convey their feelings and opinions on social media, which has grown in significance as a means of learning about people's attitudes and sentiments.

II. Literature Survey

Paper [1]: Machine learning techniques are used in this paper to determine the emotions. Since lexicon-based methods for text categorization don't require training on a dataset and perform well with enormous volumes of input data, they usually operate more rapidly. Although they said that machine learning models could fairly accurately detect the emotion of a text, VADER was the most effective lexicon-based technique.

Paper [2]: The study employed a dataset for textual sentiment analysis consisting of 8,600 tweets that were manually classified as belonging among the six emotions: surprise, anger, disgust, fear, joy, and sorrow. The study made use of the audio dataset from RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song).Despite the fact that the three modalities were first examined separately and subsequently integrated using weights, no clear accuracy has been found. On the other hand, this method approximated 70% accuracy based on test results.

Paper [3]: The study provided a generalized model for analyzing the speaker identities and content of an audio clip that features a dialogue between two people. Speaker identification and automated text conversion of the audio file are used to accomplish this. The method has several limitations, despite its effectiveness in interpreting the sentiments of the speakers in conversational discussion. There should only be one speaker at a time since the system can only manage discussions between two speakers at this time. It cannot understand simultaneous speech from two persons.

Paper [4]: Sentimental Analysis was used to gather and examine 200 samples in total. According to the research, deep learning pattern categorization and AI offer profound insights into the ASR system. Additionally, AI and Deep Learning might be employed during user-customer care executive calls to enable the computer system to monitor and detect potentially dishonest or improper behaviour from the customer care executive.

Paper [5]: They created a user-friendly platform that poses targeted queries on the user's requirements and provides him with an expedient reply. A programme like this may be tailored for a wide range of purposes, including medical and psychological as well as text, voice, and video communication.

Paper [6]: To identify the emotions, the study makes use of Deep Learning-Based Multi-Modal Emotion Recognition from Speech and Facial Expression. Deep learning is used in this study to support its capacity to identify emotions. The article contains a dataset of both spoken and facial emotions. The CNN and LSTM algorithms were employed by the system. Text and gesture modalities were not incorporated into multimodal models while they researched more effective feature extraction methods and multimodal fusion. Integration of facial expression and voice data has significantly improved the assessment approaches. They also discovered a notable improvement when comparing their method with the multimodal systems in use today.

Paper [7]: The work on topic classification and sentiment analysis in video transcriptions is presented in this publication. The MUSE-TOPIC SUBCHALLENGE delivered the information. The accuracy of the development set was 56.18%, whereas the test set's accuracy was 66.16%. Further research was required since the categories were hinted rather than definitely defined from continuous single.

Paper [8]: This study made use of Twitter, Flickr, and Instagram datasets. They unveiled ME2M, a user-friendly yet effective model for sentiment analysis in photos. The observed findings demonstrated the utility and applicability of the ME2M model. They didn't have any popularity predictions derived on sentiment analysis of images.

Paper [9]: Improvement of Sequence-to-Sequence Audio Converter by Adding Textual- Supervision's author used text-based phonetic data collecting. Machine learning techniques like the seq2seq VC model and HMM were applied. The sequence to sequence VC model is much improved by the suggested techniques; nonetheless,

insufficient training data still causes problems with model execution. Consequently, they faced the difficulty of achieving noticeably worse performance when limited instruction sets are accessible.
Paper [10]: The present study provided a summary of the latest growth in the domain of sentiment evaluation using many modes. The most well-known datasets and feature extraction methods in the domains have been categorized and scrutinized. The potency and effectiveness of the on two widely used multimodal sentiment datasets, 35 models have been assessed examination, the CMU-MOSEI and CMU-MOSI.

III. METHODOLOGY

The main concept of the study is analyzing emotions in both audio and video. Multiple inputs are taken into account for better and more trustworthy outcomes. To categorize the mood of the session, the supplied audio would be transformed into text and subsequently processed for sentiment analysis. In addition, face emotion detection using CNN will be utilized to identify facial expressions. We can create an assessment of the individual's mental health that may be utilized for further diagnosis by combining the data from the two inputs. A CNN is a Deep Learning system that has the ability to distinguish between distinct elements of an image, establish the point at which an image has undergone processing and assign values to different aspects of the image. Establish the point at which an image has undergone processing and assign values to different aspects of the image. It is only a free-source library that offers functions like face identification, object tracking, and landmark discovery, among others. Problems involving regression and classification may be solved using the machine learning technique known as "Support Vector Machine". However, it is frequently applied to classification issues. These datasets may be reduced in dimension using PCA, which increases accuracy while reducing data redundancy. CNN translates written speech into a machine-readable format that follows a logical structure to extract the sentiment from every word, phrase, and eventually video.

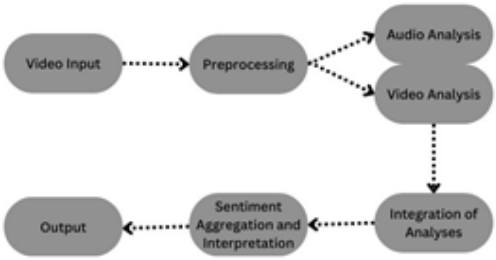


Figure.1 Data flow diagram of Sentiment Analysis

1. Data Collection: Videos are collected from various sources, including inserted videos, live streaming, or camera feeds, to ensure a diverse dataset for training and testing.
2. Video Preprocessing: The collected videos are divided into frames to facilitate both visual and audio sentiment analysis on a frame-by-frame basis.
3. Visual Sentiment Analysis (CNN): Convolutional Neural Networks (CNNs) are employed for visual sentiment analysis. Each frame undergoes feature extraction through layers of convolutional and pooling operations to capture visual patterns related to sentiment.
4. Audio Sentiment Analysis (LSTM): Long Short-Term Memory (LSTM) networks are applied to the audio data to capture temporal dependencies and nuances in the audio stream, providing a nuanced understanding of the sentiment expressed through sound.
5. Aggregate Sentiment Analysis: The final step involves aggregating the sentiment predictions from both the visual and audio analyses. This fusion approach enhances accuracy and gives a more comprehensive recognition of the sentiment expressed in the video.
6. Accuracy Enhancement Techniques: To further refine the model, techniques such as transfer learning, data augmentation, and hyper-parameter tuning are applied. These techniques aim to improve the accuracy and robustness of the sentiment analysis model.
7. Implementation Details: The model is implemented using popular deep learning frameworks such as TensorFlow, Keras models and the codebase is made available for reproducibility and further research.
8. Experimental Results: The paper includes a detailed presentation of experimental results, including accuracy metrics, confusion matrices, and comparisons with existing sentiment analysis approaches to illustrate the efficacy of the proposed CNN-LSTM fusion model.

A. FER-2013 Identification of Facial Expressions

Kaggle offers the FER-2013 Facial Emotion Detection collection, which was first presented at the International Conference on Machine Learning (ICML).







Figure.2 Furious Figure.3 Revulsion Figure.4 Furious Figure.5 Happy Figure.6 Neutral Figure.7 Sad Figure.8 Surprise

Every face in this dataset has been classified according to several emotion categories as shown in TABLE I, and each image's grayscale is 48 pixels by 48 pixels. There are 35,887 photos in the FER- 2013 collection that have seven different expression types classified by seven different classification characteristics as in [21] (TABLE I).

B. Minimilist Classification of Expressions on the Face

A micro-expression within the interpersonal psychology is a discipline of communication that is easy to observe and identify on the face. Face expressions have a crucial function in interpersonal communication because they transmit information about emotions as well as our goals and ambitions. Naturally, having the required conversation is made simpler by comprehension and the ability to read expressions on the face. Face identification, feature extraction, and classification of facial expressions are the three phases involved in classifying human facial expressions. This study's authors employed a technique that allowed for the broad-scale classification of facial expressions and included seven universal human emotions.
Human faces are recognized as follows:

1. Eyebrows drawn downward (exudes anger)
2. Raised and conjoined brows (Shows Fear)
3. Pulling up of the upper lip (Shows Disgust)
4. Neutral expression in the eyes (Shows neutral)
5. Raises cheeks (exudes happiness)
6. Raised lip corners (depicts sadness)
7. Open mouth (indicating astonishment)

IV. CONVLSTM

ConvLSTM the process of identifying the emotions in the speech segments is sequential and exhibits a high link between the subsequent speech segments. The suggested SER model processes each segment individually and detects correlations during sequential prediction to identify the emotions present in an unprocessed speech signal. As a result, it loses the spatial cues by modelling the sequential cues as a one- dimensional vector.In this architecture, the convolution LSTM (ConvLSTM) is used in place of a conventional CNN network to preserve both the spatial and temporal signals, such as the spatiotemporal information in the successive speech segments. In the ConvLSTM, the state-to-state as well as input-to-state transitions are carried out using the convolution process. To better mine the correlation between the speech segments, the ConvLSTM records the spatiotemporal cues during the convolution procedure.

$$ic = (Wix * xt + Wih * ht{-}1 + Wic \copyright Ct{-}1 + bi \quad (1)$$

$$fc = (Wfx * xt + Wfh * ht{-}1 + Wfc \copyright Ct{-}1 + bf \quad (2)$$

$$oc = (Wox * xt + Woh * ht{-}1 + Woc \copyright Ct + bo \quad (3)$$

$$gt = \tanh (Wgx * xt + Wgh * ht{-}1 + bg) \quad (4)$$

$$ct = ft \copyright ct{-}1 + ic \copyright gt \quad (5)$$

$$ht = ot \copyright \tanh(ct) \quad (6)$$

The sigmoid function is represented by "σ" the convolution operation is denoted by *, the element-wise operation is denoted by ©, and the hyperbolic tangent function is represented by tanh in the equations above. We used different symbols to represent different functionality during the convolution process. Likewise, when performing the t th step of the ConvLSTM, denoted by the subscript "t" in the equations, the various gates—namely, the input gate, storage gate, output gate, and input modulation gate—are represented by it, ft, ot, and gt, respectively. Moreover, xt illustrates the input data, while ct and ht, respectively, reflect the cell state and hidden

state. During the convolution process in the ConvLSTM, the first dimension uses the temporal information from the input tensor, but the subsequent two and three dimensions use the spatial data. The core concept of both the ConvLSTM and the succeeding layer is that the result of the preceding layer is used as the input for the subsequent layer and the conventional LSTM networks. Their approach is where they diverge most.
Since a typical RNN just has one hidden state that is transferred across time, learning long-term dependencies may be challenging for the network. In order to solve this issue, LSTMs introduce memory cells, which are long-term information storage units. Because LSTM networks can learn long-term relationships from sequential data, they are a good fit for applications like time series forecasting, speech recognition, and language translation. For the study of images and videos, LSTMs may additionally be employed in addition to several neural network architectures, such Convolutional Neural Networks (CNNs) .One kind of RNN, or recurrent neural network, that may preserve long-term dependencies in sequential data is the LSTM, or long short-term memory. Text, audio, and time series are examples of sequential data that LSTMs can process and analyze. They overcome the vanishing gradient issue that besets conventional RNNs by controlling the information flow using gates and a memory cell, which enables them to maintain or discard information as needed. LSTMs are extensively employed in many different applications, encompassing natural language processing, speech recognition, and time series forecasting.

V. Result Analysis

In this study the system was evaluated at different phases of the ability to recognize small facial expressions in design. The outcomes demonstrated how quickly and effectively the face expression detection system could employ the CNN architectural paradigm. The evidence in Table II indicates that using a separate convolution layer is the most effective way to carry out data training. After the system is put into place, analysis of the outcomes is crucial.

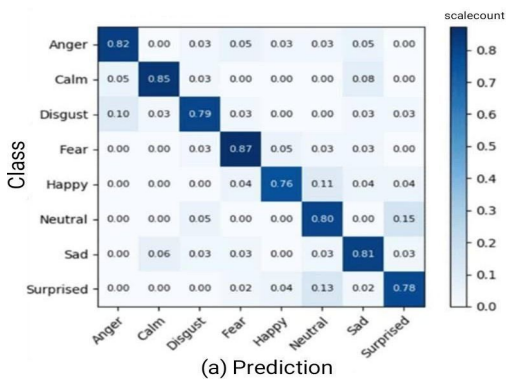
TABLE I: RESULT OF FACIAL EXPRESSION TESTING

	precision	recall	f1-score	support
angry	0.96	0.97	0.97	1484
disgust	0.97	0.95	0.96	1558
fear	0.96	0.97	0.96	1505
happy	0.96	0.95	0.96	1619
neutral	0.97	0.98	0.97	1558
sad	0.96	0.97	0.96	1478
surprise	0.98	0.97	0.97	528
accuracy			0.96	9730
macro avg	0.96	0.96	0.96	9730
weighted avg	0.96	0.96	0.96	9730

A. Prediction Test of Facial Expression

The system successfully recognizes the expression after ten iterations of the experiment for each of the seven expressions. The consequences of miss-expressing fear and anger occurred once and twice, respectively, whereas miss-expressing contempt occurred thrice. The report's Table III presents the results. Which expressions are easier to predict and which are harder to predict is shown in this table.

TABLE II: CONFUSION MATRIX



(a) Prediction

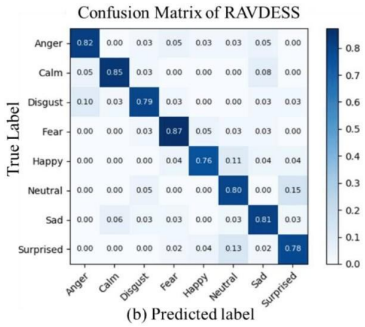
B. Result Analysis of Audio Data

24 professional actors—12 men and 12 women—were used in the RAVDESS dataset to record the pre-scripted language while evoking various emotions. Anger, fear, sorrow, and happiness are among the eight distinct the suggested dataset includes categories for speech emotions. Apart from neutral, which contains fewer audio files than the other classes in this dataset, all of the classes have comparable numbers of audio files. Each emotion was captured at a sample rate of 48 kHz, with an average audio clip lasting 3.5 seconds. On the other hand, to estimate the prediction concert, we trained and tested our suggested model using this dataset. Accuracy of our model on test data is 96.34 % using RAVDEES dataset.

TABLE III: OVERALL PERFORMANCE OF RAVDESS DATASET

	precision	recall	f1-score	support
angry	0.96	0.97	0.97	1484
disgust	0.97	0.95	0.96	1558
fear	0.96	0.97	0.96	1505
happy	0.96	0.95	0.96	1619
neutral	0.97	0.98	0.97	1558
sad	0.96	0.97	0.96	1478
surprise	0.98	0.97	0.97	528
accuracy			0.96	9730
macro avg	0.96	0.96	0.96	9730
weighted avg	0.96	0.96	0.96	9730

TABLE IV: CONFUSION MATRIX OF RAVDESS



VI. CONCLUSION

The intent of the research was to develop a transportable system, versatile, affordable, and, most all, flexible. It's a reliable technique to make sure social product reviews are accurate. This is where our suggested Sentimental Analysis method comes into play—machine learning. Achieving high-accuracy emotive video detection was our primary objective. We may also analyze video reviews with the use of this analyzing tool. Nowadays, many social media sites, such as Facebook, Twitter, and YouTube, need audio and video surveillance. With the use of technology, we are able to examine product usage and identify opinions on a certain item. The exponential growth of social media has made multimedia data an essential means of transferring human ideas and viewpoints. Social network analysis is becoming more and more popular as a potential field of study. Based on a cursory examination, we examined the most popular approaches for social media sentiment analysis using text.This paper aimed to give a comprehensive analysis of the Visual Sentiment Analysis topic, associated difficulties, and region approaches. There have been investigations into meaningful meaning using actual corporate software that might benefit from video and photo emotion analysis. The ConvLSTM network for speech signals was created and presented in this study. It is readily applicable to various speech processing-related issues, such energy forecasting and automatic speaker recognition. In order to reach a consensus, the suggested framework may be quickly used in different speech classification tasks. Semantics and sentiment analysis are two fields in which the fields of machine learning and deep learning have seen significant application. Professionals in a variety of fields benefit from computers' ability to interpret emotions and reactions, not only coders. Counsellors and even the users themselves may recognize and monitor their emotions on a regular basis with the help of this recognition. We have effectively carried out an extensive investigation on the many machine learning approaches that have been used for real-time sentiment analysis over the last few decades through this work.

REFERENCES

[1] Dr. Munish Mehta , Kanhav Gupta , Shubhangi Tiwari , Anamika A Review on Sentiment Analysis of Text, Image and Audio Data (IEEE 2021)

[2] Dr. Ashwini Rao , Akriti Ahuja , Shyam Kansara , Vrunda Patel Sentiment Analysis on User- generated Video, Audio and Text (IEEE 2021)

[3] Maghilnan S, Rajesh Kumar M Sentiment analysis on speaker specific speech data (IEEE 2017)

[4] Rohit Raj Sehgal , Shubham Agarwal, Gaurav Raj Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems (IEEE 2018)

[5] Rishit Jain, Revant Singh Rai, Sajal Jain, Ruchir Ahluwalia, Jyoti Gupta Real time sentiment analysis of natural language using multimedia input (Springer 2023)

[6] L. Cai, Jiangong Dong, Min Wei , Multi-Modal Emotion Recognition From Speech and Facial Expression Based on Deep Learning (IEEE 2020)
[7] L. Stappen and A. Baird, E. Cambria, B. W. Schuller , Sentiment Analysis and Topic Recognition in Video Transcriptions (IEEE 2021)
[8] H. Zhang, J. Wu, H. Shi, Z. Jiang, D. Ji, T. Yuan, G. Li, Multidimensional Extra Evidence Mining for Image Sentiment Analysis (IEEE 2020)
[9] J. X. Zhang , Z. H. Ling , Y. Jiang , L. J. Liu , C. Liang , L. R. Dai, Improving Sequence-To-Sequence Voice Conversion by Adding Text- Supervision (ResearchGate 2019)
[10] Sarah A.Abdu, Ahmed H. Yousef, Ashraf Salem, Multimodal Video Sentiment Analysis Using Deep Learning Approaches, a Survey.
[11] J. Choi, H. Gill, S. Ou, Y. Song, J. Lee, Design of voice to text conversion and management program based on Google Cloud Speech API (CSCI, 2018)
[12] A. Ahuja, S. Kansara, V. Patel , Sentiment Analysis on User-generated Video, Audio and Text (ICCCIS 2021)
[13] U. Doshi, V. Barot, S. Gavhane , Emotion Detection and Sentiment analysis of static images (IEEE 2020)
[14] D. Kushwaha, D. De, V. Mohindru ,Anuj Gupta , Sentiment Analysis and Mood detection on an android platform Using ML Integrated with IOT (Researchgate,2020).
[15] N. Mittal, D. Sharma, M. L. Joshi, Image Sentiment Analysis using Deep Learning (IEEE/WIC/ACM International Conference on Web Intelligence (WI) 2018)
[16] R. K. Madhupi, C. Kothapalli, V. Yarra, S.Harika, Automatic human emotion recognition system using facial expressions with CNN (IEEE,2020)
[17] H. Bhuiyan, J. Ara, R. Bardhan, Md. R. Islam, Retrieving YouTube video by sentiment analysis on user comment (IEEE ICSIPA,2017)
[18] P. Das, A. Ghosh, R Majumdar, Determining attention mechanism for “Visual Sentiment Analysis” of an Image using SVM Classifier in Deep learning based architecture (ICRITO,2020)
[19] T. Liu, J. Wan, X. Dai, F. Liu, Q. You, J. Luo , Sentiment Recognition for Short Annotated GIFs Using Visual-Textual Fusion (IEEE 2019)
[20] Lukas Christ, Shahin Amiriparian, Alice Baird, The MuSe 2023 Multimodal Sentiment Analysis Challenge: Mimicked Emotions, Cross-Cultural Humour and Personalisation (Researchgate 2023).
[21] Rushali A. Deshmukh, Vaishnavi Amati, Anagha Bhamare, Aditya Jadhav, Visual Sentiment Analysis: An Analysis of Emotions in Video and Audio (ICICNIS 2023).