International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Multimodal Sentiment Analysis using Multiple Neural Networks and Natural Language Processing Models

M. Sathya Devi*[1], Dr. B. Indira[2]

Submitted:10/03/2024 **Revised**: 25/04/2024 **Accepted**: 02/05/2024

Abstract: Multimodal Sentiment Analysis (MSA) identifies emotional expressions over time using visual and audio information. MSA, an emerging field, detects and analyzes emotions across text, speech, and images, with practical applications in social media analysis, customer feedback analysis, healthcare monitoring, and investigations. In the realm of e-commerce, product reviews significantly influence consumer purchasing behaviours, with user-generated content offering diverse perspectives on products and services. Sentiment analysis systematically deciphers the emotional undertones within text, providing businesses with actionable insights. Videos and audio recordings introduce richer dimensions to these reviews, capturing nuanced expressions and vocal cues. Integrating both modalities into sentiment analysis offers a deeper comprehension of reviewers' emotional states, enhancing the accuracy of product perception and satisfaction assessments.

Our innovative approach merges the Haar cascade algorithm for facial detection and sentiment analysis in videos with the BERT model for audio sentiment analysis, synergizing visual and auditory cues. This integration improves accuracy and robustness, facilitating precise assessments of product sentiment. By fusing outputs from these modalities, we can identify the most salient emotional cues expressed by reviewers, providing businesses with a comprehensive understanding of consumer sentiment and aiding informed decision-making processes.

This paper presents a novel approach to sentiment analysis using facial expression and audio data from the CMU-MOSI dataset. The process involves extracting a focused video of the subject's face, converting the corresponding audio into text, and analyzing sentiment. The proposed methods achieved promising accuracy and hold immense potential when combined with other modalities.

Keywords: *audio signals, BERT model, CMU-MOSI dataset, facial expression, Haar Cascade Algorithm, MobileNet model, mouth state detection, sentiment analysis.*

1. Introduction

As humans, we express emotions in various ways, such as language, facial expressions, and tone of voice. With the growing use of digital communication platforms, these emotions are increasingly captured in different modalities, creating opportunities for researchers to understand and analyse them. This has led to the emergence of multimodal sentiment analysis, a field that aims to detect and analyse emotions across multiple modalities, such as text, speech, images, and videos. Sentiment analysis has gained significant attention in recent years, offering a more comprehensive understanding of human sentiment than traditional approaches that rely solely on text data. By incorporating different sources of data, multimodal sentiment analysis provides a more nuanced view of sentiment, capturing emotions that may be missed in text data alone. The practical applications of sentiment analysis are numerous and diverse, ranging from marketing and customer feedback analysis to healthcare monitoring and personalization or in the field of investigation.

For example, movie studios and production companies can use sentiment analysis to gauge the public's opinion of a new movie release and adjust their marketing strategies accordingly. If sentiment analysis reveals that the overall sentiment towards a movie is negative, the studio may decide to reposition their marketing message or even delay the release of the movie. Movie review websites can also benefit from sentiment analysis by automatically classifying reviews based on their sentiment, helping users quickly find reviews that match their preferences and generating summary statistics about the sentiment of reviews for a particular movie.

This paper presents an innovative approach to sentiment analysis in product reviews by integrating video and audio modalities, thereby extracting nuanced emotional insights, leveraging the Haar cascade algorithm for video analysis and the BERT model for audio processing. Our method offers a holistic understanding of reviewer sentiment. By scrutinizing facial expressions, gestures, speech patterns, and vocal intonations, we capture the intricate emotional context of product reviews, providing deeper insights into consumer perceptions and satisfaction. Additionally, our adaptable solution caters to diverse platform capabilities and user preferences, making it suitable for both offline and online product review systems. Through empirical evaluation and

[1] *Vasavi College of Eng, Telangana - INDIA*
[2] *Chaitanya Bharathi Institute of Technology, Telangana – INDIA*
** Corresponding Author Email: sathyamaranganti@staff.vce.ac.in*

comparative analysis, we demonstrate the effectiveness and robustness of our approach across various product categories and review platforms. This research significantly contributes to the advancement of sentiment analysis techniques and multimodal fusion, highlighting the potential for enhanced sentiment understanding in product review systems. Furthermore, our findings offer practical implications for businesses seeking to utilize consumer feedback effectively and make informed decisions. By integrating the video-based Haar cascade algorithm and the audio-based BERT model outputs, our method aims to identify prominent emotional cues expressed by reviewers, thereby providing a comprehensive assessment of product sentiment. Our proposed approach addresses the dynamic needs of both offline and online product review systems. It can efficiently analyze pre-recorded video and audio reviews offline, offering valuable insights into product sentiment without real-time processing requirements. Conversely, for online platforms hosting live video or audio reviews, our method seamlessly adapts to real-time processing, enabling immediate feedback and analysis of consumer sentiment as reviews are generated. This versatility makes our approach applicable across various review platforms, enhancing the depth and accuracy of sentiment analysis in product reviews.

The entire process of sentiment analysis is done based on fusion model, in which the idea of considering light neural networks for each of the sub process, the efficiency of the process can be improved. As part of this, we used Haar Cascade model to identify the face from the video frame and extract it to check whether mouth is open or not. This part of implementation is called 'mouth classifier'. This is used to extract the audio part of the mp4 video and the combined audio is stored in a separate file for further processing. The audio is converted to text using Google API. From text, the sentiment can be identified using the BERT model. Rather than using a single model to perform all the tasks, the light weight neural networks can work efficiently. This is represented using the architecture diagram in figure-1

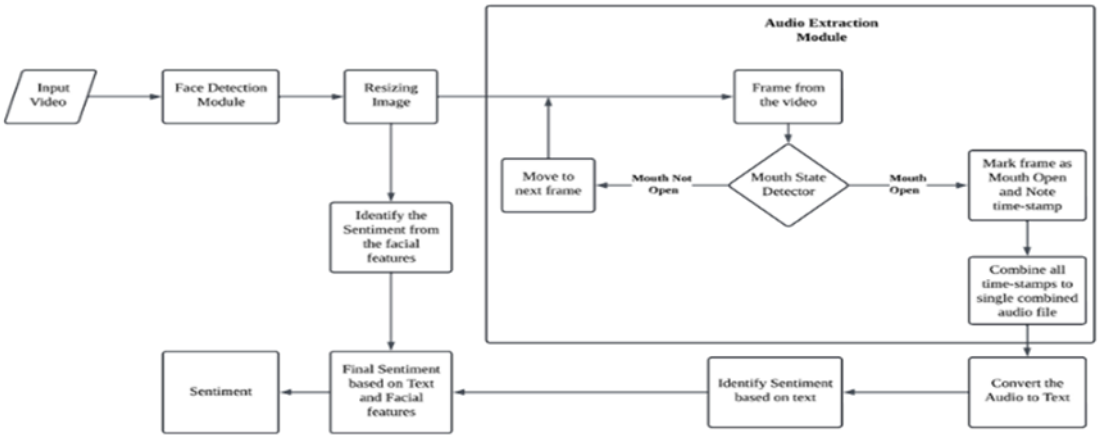

Fig-1 MMSA Architecture

2. Literature Review

Qiupu Chen et al [1] proposes a novel weighted cross-modal attention mechanism with a sentiment prediction auxiliary task for multimodal sentiment analysis. The proposed method aims to effectively fuse information from different modalities, including text, image, and audio, to improve the sentiment analysis accuracy. In this approach, the authors mentioned about three main components: a cross-modal attention mechanism, a sentiment prediction auxiliary task, and a weighting scheme.

Sarah A. Abdu et al [2] provides a comprehensive survey of deep learning approaches for multimodal video sentiment analysis. The authors begin by introducing the importance of multimodal sentiment analysis in the context of video data, where emotions can be conveyed through different modalities, such as facial expressions, speech, and scene context. The paper then presents a detailed review of the literature on deep learning approaches for multimodal video sentiment analysis, including feature extraction techniques, fusion methods, and classification algorithms.

Atitaya Yakaewl et al [3] proposes two main components: a feature extraction module and a classification module. The feature extraction module uses a combination of convolutional and recurrent neural networks to extract features from different modalities, including facial expressions, speech, and scene context. The classification module uses a lightweight deep neural network architecture, called the MobileNet, to classify the extracted features into different sentiment categories.

Kaldi [4] is an open-source speech recognition toolkit used in research and industry. It provides a comprehensive set of tools to build a state-of-the-art speech recognition

system, that includs feature extraction, acoustic modeling, language modeling, and decoding. Kaldi uses a neural network-based acoustic model for speech recognition, which is trained on a large dataset of labeled speech data.

Google Cloud Speech-to-Text API [5] is a deep learning approach that uses a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to transcribe speech into text. The model is pre-trained on large datasets of speech data, which allows it to recognize a wide range of accents and speaking styles. It can handle multiple languages and can also recognize specialized vocabulary and context-specific jargon.

The Google Cloud Speech-to-Text API provides a simple interface for developers to integrate speech recognition into their applications. The API can transcribe both streaming and non-streaming audio and can provide real-time transcription results. It also includes features such as automatic punctuation which can help improve the accuracy and usefulness of the transcribed text.

PocketSphinx [6] is an open-source speech recognition engine developed by Carnegie Mellon University. It is a lightweight and efficient speech recognition system that is designed to work on mobile devices and embedded systems with limited resources. PocketSphinx is based on Hidden Markov Models (HMMs) and uses statistical language models to recognize speech. It can recognize continuous speech and can be trained to recognize new words and phrases.

Gina Khayatun Nufus, Mustafid Mustafid, and Rahma [7] proposes a methodology utilizing Word2vec and Long Short-Term Memory (LSTM) models for aspect-level sentiment classification in video-on-demand application reviews. The integration of Word2vec embeddings with LSTM architecture enhances sequential data modelling, facilitating accurate sentiment classification.

Gaurav Meena, Krishna Kumar Mohbey, Sunil Kumar [8] focuses on image-based sentiment analysis. Through transfer learning and pre-trained deep CNN models, the study aims to detect and classify emotions in images, achieving superior results compared to traditional machine learning approaches.

Nur Hasifah A Razak, Muhammad Firdaus Mustapha

[9] presents a systematic methodology for sentiment analysis of Amazon product reviews across different categories. Utilizing RapidMiner, the study involves data acquisition, preprocessing, and model validation to analyze sentiment trends and their impact on product attitudes.

Dorca Manuel-Ilie, Pitic Antoniu Gabriel, Crețulescu Radu George [10] proposes a comprehensive approach to sentiment analysis that integrates deep learning concepts. By annotating topics and emotions, customizing neural network architectures, and systematic model evaluation, the methodology aims to optimize sentiment analysis performance across diverse datasets.

Amira Samy Talaat [11] this study combines BERT with BILSTM and BIGRU algorithms. Through preprocessing, model comparison, and evaluation on multiple datasets, the research demonstrates the superiority of hybrid models over classical machine learning methods.

3. Design and Implementation

The implementation has been done in two ways. In the first case, we have extracted the facial features from a already available video, and in the other case, we tried to capture the images directly from the camera. In both the cases the Haar Cascade algorithm is used to identify the face[3]. In the first case, the audio is extracted from the video and processed for further sentiment identification, whereas in the second case the audio is taken from the system's microphone.

3.1 Case: 1

In the first case, the implementation is 4-stage pipeline to perform sentiment analysis.

The first stage of the pipeline is face detection, where the presence of a human face is identified in each frame of the video. This is achieved using object detection algorithms (like Haar cascade).

In the second stage, the mouth state detector is used to identify audio fragments where the mouth is open or closed. This involves analyzing the movement of the lips in the video frames to determine the state of the mouth.

The third stage of the pipeline involves the conversion of the previously obtained audio fragments into text using natural language processing tools. This enables the extraction of text from the audio, which can then be used for further analysis.

The final stage is performing the sentiment analysis on the obtained text to give the final sentiment, one of *Negative*, *Neutral*, *Happy*.

This multi-stage approach ensures simple development, easy understanding and modularity in the architecture which ensures future expansion and customization.

3.1.1. Face Recognition

The dataset used is the famous datasets in sentiment analysis, the Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity Dataset or simply know as CMU-MOSI. The CMU MOSI dataset is a multimodal dataset designed for research in multimodal sentiment analysis and emotion recognition. It was

created by researchers at Carnegie Mellon University and includes audio, visual, and textual data from a diverse set of sources, including movies, TV shows, and online videos. The opinion video clips is annotated with sentiment ranging from -3 to 3.

We chose Haar Cascade classifier due to its superior performance in detecting subjects in close proximity. By integrating Haar Cascade classifier with OpenCV2, we can efficiently identify the frames where the subject is present and zoom into their face to obtain a 200x200x3 square video for further analysis. In parallel with the video processing, we extract the audio from each video sample. This audio data is then used in subsequent steps of our pipeline.

3.1.2 Audio Analysis: Mouth State Detection and Conversion

In the second stage of our pipeline, we aim to detect whether the mouth is open or closed in each frame of the video. This enables us to obtain timestamps for each frame where the mouth is open or closed. By doing this, we can extract continuous audio fragments where the mouth is open and combine them to obtain a single audio segment consisting solely of audio samples where the mouth is open. To achieve this, we make use of OpenCV and dlib.

The first step in detecting whether the mouth is open or closed is to take the processed 200x200x3 video from the previous stage and project 68 facial landmarks onto the current frame using the shape_predictor_68_face_landmarks.dat. The points corresponding to 49-55 are for the upper lip area, whereas the points from 56 to 61 correspond to the lower lip area. We obtain the coordinates where these points are located and take the mean of the upper and lower lip coordinates. This gives us the mean location of the lips, to get the average tip position of the upper lip as well as the lower lip. Finally, we calculate the distance between these coordinates and if it exceeds a threshold, we declare the mouth to be open; otherwise, it is considered closed. At the same time, we keep track of the timestamps of the frames where the mouth is open.

Next, we take the wav file corresponding to the video obtained earlier and slice the audio segments which correspond to the continuous timestamps where the mouth was open. These audio segments are then combined into a single combined audio segment. Naturally, the combined audio segment would be quicker than the original, we slow it down by a factor of 0.5 to make it more understandable.

Overall, this approach allows us to obtain a single audio segment consisting only of audio samples where the mouth is open, which can be further analyzed in the next stage of our pipeline.

Fig 2 Frame with closed label mouth is closed according to mouth state detector

Fig 3 Frame with open indicating label indicating mouth is open according to mouth state detector

3.1.3 Speech to Text Conversion

The Google Cloud Platform's Speech to Text API is an advanced technology that uses a deep learning approach to transcribe spoken words into text with exceptional accuracy. It provides several advanced features such as automatic punctuation and multi-language support, making it a preferred choice for speech recognition tasks. By integrating this API into our pipeline, we obtain highly accurate and reliable transcription results, which we use for sentiment analysis in the final stage.

3.1.4 Sentiment Analysis

In the final stage of the pipeline, BERT model is used to identify the sentiment of the text obtained from the previous stage. There are only two classes used here for the sentiment specification ie., 'Posiitve' and 'Negative'.

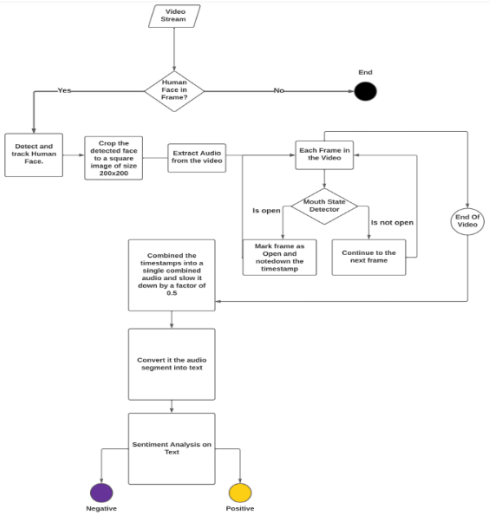

Fig 4 Case 1 Flow of Sentiment Analysis

3.2 Case 2

In the second case, the pipeline has 3 stages to perform the sentiment analysis.

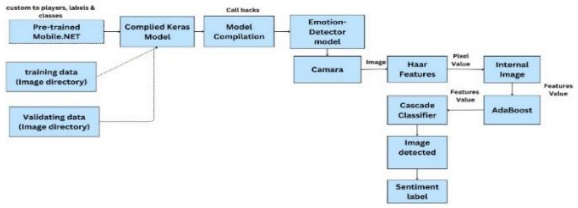

Fig 5 Case 2 Flow of Sentiment Analysis

3.2.1 Face Recognition and Sentiment Representation

The first stage is reading the input from the camera and identifying face using Haar Cascade. By considering the ROI ie Region of Interest, and using a predefined CNN model 'Emotion_Detection' to identify the facial sentiment.

Here the pre-trained model MobileNet model is used with weights from ImageNet as the 'Emotion Detection' model. MobileNet works as the base CNN in identifying the sentiment based on the facial features. The dataset used in this case was prepared by Pierre-Luc Carrier and Aaron Courville, as part of an ongoing research project. They have graciously provided the workshop organizers with a preliminary version of their dataset to use for this contest.

https://www.kaggle.com/competitions/challenges-in-representationlearning-facial-expression-recognition-challenge/data

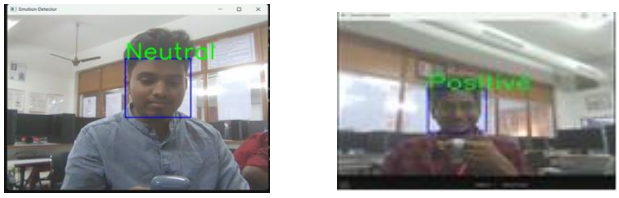

Fig 6 Video based sentiment showing neutral and positive sentiment

3.2.2. Audio Analysis and Sentiment Analysis

The second stage takes the input from the microphone. The system continuously listens to the user for any audio input from the user, through a microphone. Using Google API, the audio signal is converted to text. Using BERT model we can get sentiment of the converted text.

3.2.3. Combined Sentiment Analysis from Audio and Video

The third stage is the integrating stage where the two outputs from video and audio are taken and combined. When integrating the two inputs ie sentiment from audio and video, we have considered weightage for audio as 60% and video as 40%. The combined sentiment is obtained by considering these two inputs. The number of classes taken here are 'Extremely Negative', 'Negative', 'Neutral', 'Positive', 'Extremely Positive'.

4. Results and Discussion

4.1 Case 1

In the first case, the model was good in classifying the negative labels. The plan was to classify text into either one of positive or negative, it must be noted that experimentation was done on multi-class (Neutral, Positive, Negative) classification also. This gave us an accuracy of only around 54%. This is due to the lack of perfectly neutral sentences, causing a major class imbalance problem hence poor performance.

From the table, the precision shows that the classes Negative is not identified for the video input. It could be because the number of samples for the 'negative' class are not sufficient to and the dataset could be imbalanced, because of which the results are not very encouraging.

	precision	recall	f1-score	support
0	0.65	0.69	0.67	153
1	0.72	0.68	0.70	177
accuracy			0.69	330
macro avg	0.69	0.69	0.69	330
weighted avg	0.69	0.69	0.69	330

Fig 7 Classification Report

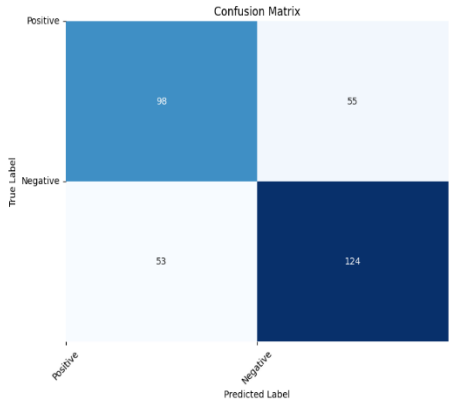

Fig 8 Confusion Matrix

4.2 Case 2

Table 1. Results when considered separately and combined

Sentiment	Precision		Recall		Accuracy (Individual)		Accuracy Combined)
	VH	AB	VH	AB	VH	AB	Video + Audio
EN	0.58	0.91	0.56	0.97	0.96	0.97	0.96

N	0.50	0.95	0.62	0.90	0.96	0.97	0.99
Ne	0.60	0.88	0.66	0.85	0.97	0.99	0.98
P	0.87	0.98	0.81	0.96	0.97	0.99	0.98
EP	0.53	0.91	0.80	0.97	0.96	0.97	0.96

EN – Extremely Negative

N - Negative

Ne – Neutral

P – Positive

EP – Extremely Positive

VH - Video using Haar Cascade Classifier

AB – Audio using BERT Model

The dataset used for image sentiment identification (MobileNet) model consists of 28,709 samples which were divided into seven classes Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral. This dataset was used and the number of classes represented here are five, Extremely Negative, Negative, Neutral, Positive, Extremely Positive. There could be a variation of precision and recall as it is a multiclass classification and few samples are identified differently. From the table, the precision shows that the Negative classes are not identified for video input. It could be because the number of samples for negative class are not sufficient.

5. Conclusion

In both the cases, using predefined models can help us to identify the sentiment by considering both audio and video inputs, using different ways of extracting the features. The inputs can be provided from a webcam for live video or can be taken from an mp4 file.

As we are using different models for sentiment analysis for audio and video separately, this can be a better way of identifying the overall sentiment. This approach can be used for different types of applications. By considering the characteristics of audio, the efficiency of the model can be improved.

Acknowledgements

The authors would like to thank Mr.D.Kiran and Mr.K.Sunil for their support during the process of implementation of this paper.

References

[1] Qiupu Chen, Guimin Huang and Yabing Wang, "The Weighted Cross-Modal Attention Mechanism With Sentiment Prediction Auxiliary Task for Multimodal Sentiment Analysis" Published in: IEEE/ACM Transactions on Audio, Speech, and Language Processing. Date of Publication: 20 July 2022. DOI: 10.1109/TASLP.2022.3192728. Publisher: IEEE.

[2] Sarah A. Abdu, Ahmed H. Yousef, Ashraf Salem, Multimodal Video Sentiment Analysis Using Deep Learning Approaches, a Survey, Information Fusion, Volume 76, 2021, Pages 204-226, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2021.06.003.

[3] Yakaew, Atitaya & Dailey, Matthew & Racharak, Teeradaj. (2021). Multimodal Sentiment Analysis on Video Streams using Lightweight Deep Neural Networks. 442-451.

[4] 10.5220/0010304404420451.

[5] Chi Sun, Xipeng Qiu, Yige Xu & Xuanjing Huang. (2019). How to Fine-Tune BERT for Text Classification? LNAI,volume 11856

[6] https://doi.org/10.1007/978-3-030-32381-3_16

[7] Google. (2023). Google Cloud Speech-to-Text API [Computer software]. Retrieved from https://cloud.google.com/speech-to-text.

[8] Kaldi website, URL: https://kaldi-asr.org/. Accessed on [14th April 2023].

[9] Gina Khayatun Nufus, Mustafid Mustafid, and Rahmat "Sentiment Analysis for Video on Demand Application User Satisfaction with Long Short Term Memory Model" E3S Web Conference Volume 317,2021. The 6th International Conference on Energy, Environment, Epidemiology, and Information System (ICENIS 2021)

[10] Gaurav Meena, Krishna Kumar Mohbey, Sunil Kumar "Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach"

[11] Nur Hasifah A Razak, Muhammad Firdaus Mustapha. "Amazon Product Sentiment Analysis using RapidMiner"

[12] Dorca Manuel-Ilie, Pitic Antoniu Gabriel, Crețulescu Radu George "Sentiment Analysis Using Bert Model" https://www.researchgate.net/publication/376670839_Sentiment_Analysis_Using_Bert_Model

[13] Amira Samy Talaat. "Sentiment analysis classification system using hybrid BERT models".

[14] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00781-w