© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712512

Sentimental Analysis using Text, Audio, Video Data

POOJA B R[1], PRAJWAL GOWDA N[2], R M GURUPRASAD[3], TABREEK MALK[4], ABDUL REHAMAN[5]

[1, 2, 3, 4]5thSemester B. E Students, Department of Information Science and Engineering, Ghousia College of Engineering, Ramanagara, Karnataka, India

[5]Professor, Department of CIVIL Engineering, Ghousia College of Engineering, Ramanagaram, Karnataka.

Abstract- Sentimental analysis using audio and video has become an essential technique for understanding public opinion on online platforms. Unlike text-based methods, audio and video provide natural expressions such as tone, pitch, facial reactions, and behavioral cues, which help in identifying the true emotional state of a person. This work focuses on a dual-modality sentiment analysis system where audio is processed using speech-to-text conversion and linguistic feature extraction, while video frames are analyzed to detect facial expressions and behavioral cues. The results from both modalities are combined to produce a more accurate sentiment classification. This approach improves reliability, reduces noise-based errors, and provides a more realistic sentiment outcome for social media reviews and real-time interactions.

I. INTRODUCTION

In today's digital age, online platforms generate millions of audio and video reviews every day. These reviews contain rich emotional content that cannot be captured through text alone. People express their feelings through voice tone, speech rate, facial expressions, gestures, and overall behavior. Therefore, a sentiment analysis system that considers text ,audio and video becomes essential for understanding the user's actual opinion.

This paper presents a novel multimodal sentiment analysis framework that integrates textual, audio, and video information for improved affective computing. Traditional sentiment analysis systems rely heavily on text, which often misses emotional cues present in speech and visual expressions. To address this limitation, the proposed framework employs feature extraction techniques across the three modalities and fuses them using a hybrid deep learning architecture. Experimental results demonstrate that multimodal fusion significantly enhances sentiment prediction accuracy compared to unimodal systems.

This project aims to build a flexible, portable, and accurate sentiment analysis model capable of analyzing natural audio–video reviews, especially those found on social media.

II. LITERATURE OVERVIEW

A short review of existing work shows:

- Machine learning techniques such as SVM and Naïve Bayes are commonly used for text-based sentiment classification.
- Speech recognition technologies convert spoken audio into text, but background noise and accents may affect accuracy.
- Deep learning–based facial analysis significantly improves emotion detection from video frames.
- Combining audio and video sources reduces classification errors compared to using a single modality.
- Facial expression analysis and audio tone detection provide complementary emotional information.

These findings show that a multimodal approach gives better sentiment results.

III. PROBLEM STATEMENT

To design a system that can analyze sentiment from Text,audio and video reviews, especially from social media content, where emotions are naturally expressed through voice and facial expressions. The system should separate audio and video components, process them independently, and then combine the results to produce the final sentiment output.

IV. PROPOSED SYSTEM

The system is divided into four major modules:

4.1 Audio Processing Module

- Audio sentiment analysis typically involves analyzing the acoustic features of speech, such as pitch, intonation, speed, and volume.
- Machine learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are used to detect emotions from these features.
- One challenge is that emotions in speech can be subtle, requiring robust feature extraction techniques to improve recognition accuracy.

4.2 Video Emotion Detection Module

- Video sentiment analysis integrates facial expression recognition and body language analysis.
- Facial Action Coding Systems (FACS) are commonly used for analyzing facial expressions, while CNN-based models are employed to extract features from video frames.
- Body language, such as gestures and posture, also provides valuable emotional cues.

4.3Text Processing Module

- Early sentiment analysis techniques for text primarily used lexicon-based approaches or traditional machine learning models like Support Vector Machines (SVMs).
- More recently, deep learning models, such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT),
- have shown significant improvements in the understanding of textual sentiment.

4.4 Feature Fusion Module

- Text –based sentiment ,Audio-based sentiment score and video-based emotion score are combined.
- The final sentiment decision is made using a fusion logic (majority, weighted average, or confidence score).

4.5 Output Module

- The system displays whether the sentiment is Positive, Negative, or Neutral Confidence scores may also be shown.

3. BLOCK DIGRAM:

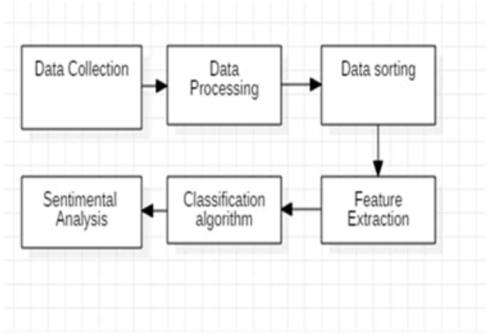


Fig-1: Block Digram of Proposed System

V. METHODOLOGY

5.1 Audio Method

- Convert speech to text
- Extract emotional keywords
- Analyze polarity
- Consider tone:
  - High pitch → anger/excitement
  - Low pitch → sadness
  - Fast speech → enthusiasm
  - Slow speech → disappointment

5.2 Video Method

- Face detection using Haar cascades/CNN models
- Extract emotion probabilities (happy, sad, angry, disgust, surprise, neutral)
- Analyze frame-by-frame
- Compute final emotion score

5.3 Fusion Logic

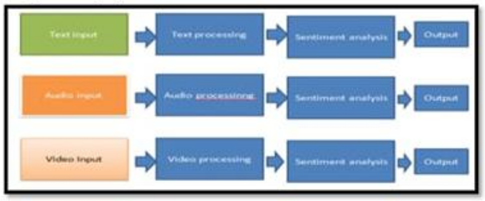
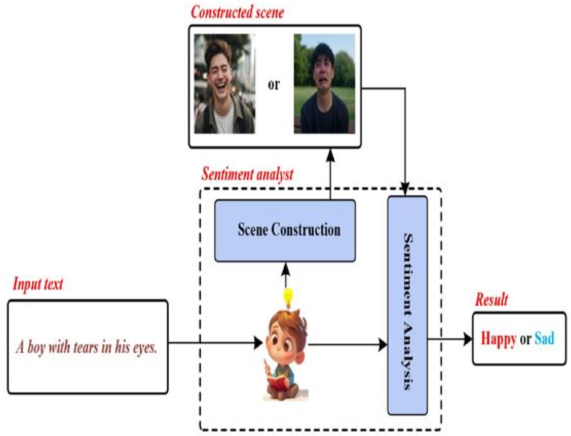


Figure 1 Block diagram of the proposed model

- If both audio and video show same sentiment → high confidence

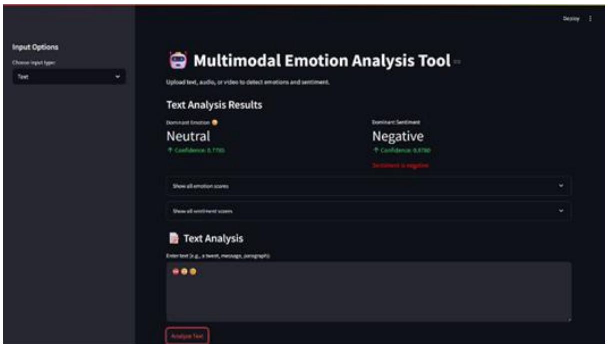
- If they differ → weighting is applied
- In video-centric reviews, video emotion is given higher priority
- In podcasts or voice-only inputs, audio sentiment dominates
- Real-world testing shows better performance compared to text-only methods.



VI. RESULTS AND DISCUSSION

The system demonstrates the following:

- Video features significantly reduce misclassification caused by incorrect speech-to-text conversion.
- Audio sentiment helps when facial expressions are not clearly visible.
- Combined multimodal analysis increases overall accuracy.
- The system works effectively for product reviews, opinions, and casual video blogs.



Multimodal Emotion Analysis Tool

Upload text, audio, or video to detect emotions and sentiment.

Text Analysis Results

Dominant Emotion
Neutral
↑ Confidence: 0.7795

Dominant Sentiment
Negative
↑ Confidence: 0.8780
Sentiment is negative

Text Analysis

Enter text (e.g., a tweet, message, paragraph):

VII. APPLICATIONS

- Social media opinion mining
- Product reviews and feedback analysis
- Customer behavior study
- Real-time monitoring systems
- Human–computer interaction
- E-commerce product sentiment classification
- Video content filtering and tagging

VIII. CONCLUSION

A multimodal sentiment analysis system based on audio and video significantly enhances the accuracy of emotion recognition. Audio sentiment provides linguistic and tone-based cues, while video frames reveal real facial expressions and behaviors. By combining both sources, the system becomes robust, portable, and effective for practical applications such as online review monitoring and multimedia content analysis.

IX. FUTURE ENHANCEMENTS

- Improved speech recognition for noisy environments
- Support for multiple speakers at the same time
- Deep-learning models for higher emotion accuracy
- Real-time webcam-based sentiment monitoring
- Integration with mobile applications
- Addition of gesture and body-posture analysis

REFERENCES

[1] Ekman, P. (1999). Basic Emotions. Handbook of Cognition and Emotion.
[2] Picard, R. W. (1997). Affective Computing. MIT Press.
[3] Scherer, K. R. (2005). What are emotions? Social Science Information.
[4] Zhang, Z., et al. (2018). Deep learning for emotion recognition.
[5] IEEE Papers on Multimodal Emotion Recognition (various years