The International Journal of Engineering and Science (IJES)
|| Volume || 14 || Issue || 4 || Pages || PP 71-77 || 2025 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Sentiment Analysis of Video Data Using AI Technology

Vikash Kumar Pradhan[1], Akram Pradhan[2], Gautam Kumar Sethi[3], Vikram Pradhan[4]
Bachelor of Technology (CSE) Kalinga University Raipur, India

Dr. Omprakash Dewangan
Assistant Professor Faculty of CS & IT Kalinga University Raipur, India

----------ABSTRACT----------
This research is proposed that, with the rises of short video platforms and the growing popularity of video content on social media, entertainment and customer feedback platforms, understanding human emotions has become very crucial. Users are shift from text-based interactions to short videos, vast amounts of multimodal data are being generated. Traditional sentiment analysis methods, which rely on a single type of data, struggle to accurately interpret emotions in these videos. This research presents a new approach to sentiment analysis by combining text, audio, and visual data using deep learning techniques. We use BERT to analyze the sentiment of text, LSTM to interpret emotions in audio, and CNN to recognize facial expressions. These individual models work together to classify the overall sentiment of a video as positive, negative, or neutral which is known as a Fusion model (Multimodel). By combining these different methods, our model offers a more accurate and reliable way to understand emotions in video content. This approach can be applied to areas like social media monitoring, customer feedback analysis, and human-computer interaction, providing a deeper and more complete analysis of emotions.

Keywords— *LSTM, BERT, CNN, Machine Learning,Artificial Intelligence, Video data, Sentiment Analysis.*

Date of Submission: 08-04-2025 Date of acceptance: 19-04-2025

I. INTRODUCTION

With the rapid growth of social media and mobile internet, short video content has become a key form of communication and user interaction. These videos are rich in multimodal information text, speech, and visuals that together reflect human emotions. Analyzing sentiment from such content is both valuable and challenging, as it requires understanding emotional cues across multiple data types.

Emotions influence how people think, learn, communicate, and make decisions. Positive emotions can improve attention, memory, and social connection, while negative emotions may limit focus and lead to cautious behavior. In online videos, emotional expressions are naturally conveyed not just through words, but also tone of voice and facial expressions, making single-modal analysis (like text-only) insufficient for accurate sentiment classification.

Traditional sentiment analysis methods often ignore audio and visual information, resulting in incomplete or biased interpretations. Multimodal analysis, on the other hand, combines text, audio, and visuals for a fuller understanding of emotional content. However, challenges such as imbalanced datasets and noisy labels still affect performance and reliability.

To address this, our research introduces a deep learning-based multimodal sentiment analysis framework. It uses BERT to analyze transcribed text, LSTM to interpret audio signals, and CNN to detect visual emotion cues from video frames. These features are fused to classify sentiment as positive, negative, or neutral. This approach improves accuracy and has practical applications in social media monitoring, customer feedback analysis, and emotion-aware systems.

II. LITERATURE REVIEW

Historically, sentiment analysis has been centered around text, relying on rule-based approaches, lexicons, and machine learning models like Naïve Bayes and Support Vector Machines. However, these methods often struggle to capture emotions effectively, especially in video content, where speech tone and facial expressions provide essential emotional context.

The advent of deep learning has significantly advanced sentiment analysis. Models such as CNNs have been used for emotion detection in images, LSTMs excel in processing sequential data like audio, and transformers

like BERT have improved text-based sentiment analysis. Recent research emphasizes the effectiveness of multimodal sentiment analysis, which combines various data sources such as speech, facial expressions, and text to achieve higher accuracy. Studies have demonstrated that CNNs are highly effective in recognizing facial expressions, LSTMs perform well in audio emotion classification, and BERT enhances the accuracy of text-based sentiment analysis.

Despite these advancements, there are still challenges in synchronizing multimodal data, reducing noise, and improving real-time performance. To tackle these issues, our research proposes a fusion model that integrates BERT for text, LSTM for audio, and CNN for visual analysis, enhancing sentiment classification in video content. This approach leverages the strengths of each modality, offering a more reliable and robust framework for sentiment analysis.

III. RESEARCH DESIGN AND METHODOLOGY

This study generally focuses on implementing a multimodal sentiment analysis framework that processes video data by extracting textual, audio and visual features to classify emotions accurately. Unlike traditional sentiment analysis, which relies on single-modal data. Our approach combines the strengths of BERT, LSTM, and CNN models to classify video content into positive, negative, or neutral sentiments. The process involves data acquisition, feature extraction, multimodal fusion and classification. The methodology consists of the following steps:

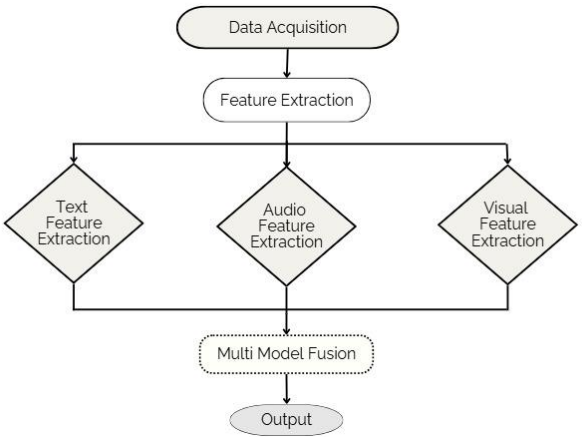

Fig. Structure of Proposed Sentiment Analysis of Video Data Methodology.

3.1. Data Acquisition

A video file serves as the primary input, from which multiple modalities are extracted. The video contains speech, facial expressions, and background sounds, all of which contribute to sentiment classification. To ensure consistency, the video is preprocessed, and relevant data is extracted:

- **Speech Extraction:** The audio stream is separated from the video using moviepy and converted into a uniform format for analysis.
- **Visual Frame Extraction:** Keyframes are captured at specific intervals to analyze facial expressions.
- **Text Extraction:** Speech-to-text conversion is performed using SpeechRecognition techniques.

3.2. Feature Extraction

To accurately classify sentiment from video content, three distinct feature extraction methods are applied, each targeting a specific modality: text, audio, and visual. These extracted features provide a comprehensive representation of the emotions conveyed in the video.

Text Feature Extraction

Speech from the video is first isolated and converted into text to analyze linguistic sentiment. The following steps are performed:

- **Audio Extraction:** The "moviepy" library is used to extract the audio stream from the video file.

$A(t) = \text{moviepy}(t)$

- **Speech-to-Text Conversion:** The extracted audio is processed using the "SpeechRecognition" library, with Google Speech Recognition API transcribing speech into text.

$T = \text{SpeechRecognition}(A(t))$

- **Text Tokenization**: The transcribed text is tokenized using a pre-trained "BERT tokenizer" (bert-base-uncased), which segments text into subword units for better handling of unseen words.

T_{tokens} = BERT_Tokenizer(T)

- **Semantic Feature Extraction:** The BERT-based "TextEncoder" converts the tokenized text into context-aware embeddings, capturing deep semantic relationships and sentiment cues.

F_{text} = BERT_TextEncoder(T_{tokens})

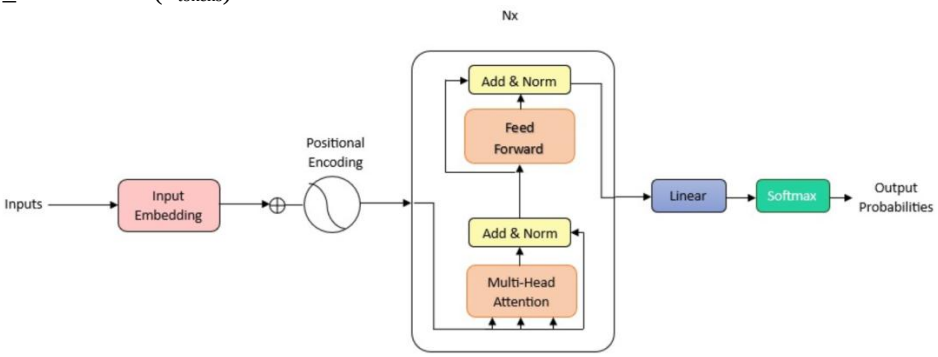

Fig. BERT TextEncoder Architecture.

Audio Feature Extraction

Audio signals contain rich emotional information that can be extracted through specialized feature engineering and deep learning methods. The following steps are used to analyze audio-based sentiment:

- **Feature Computation:** The extracted audio in text extraction phase is processed using the "Librosa" library to compute Mel-Frequency Cepstral Coefficients (MFCCs), which model speech characteristics such as pitch, tone, and energy levels.

$$MFCC = \sum_{m=1}^{M} X(m).cos\left(\frac{(m-0.5)\pi k}{M}\right)$$

- **Sequence Learning**:Since speech is a sequential data, an LSTM-based "AudioEncoder" processes the MFCC features to capture temporal dependencies in the audio signal.

$h_t = \sigma(W_h . MFCC + b_h)$

$F_{audio} = h_N$

Where:

- $W_h$and b_h are LSTM weight matrices and bias.
- $h_t$is the hidden state at time step t.
- h_N is the final hidden state, representing the extracted audio feature vectorF_{audio}.

- **Emotional Representation:** The LSTM model learns patterns in voice modulation and pitch variation, helping classify positive, negative, or neutral emotions from speech.

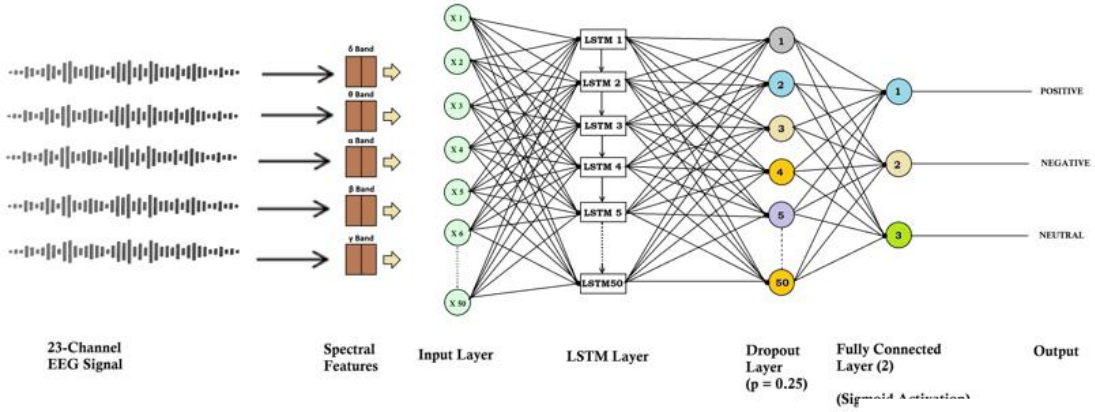

Fig. LSTM Architecture

Visual Feature Extraction

Facial expressions are one of the most direct and powerful indicators of human emotion. To extract visual sentiment cues, the following approach is used:

- **Frame Sampling:** A keyframe is extracted from the video at regular intervals using "OpenCV (cv2)" to capture facial expressions.

I_n= ExtractFrames($V(t)$, Δt)

- **Preprocessing:** The extracted frame is resized and normalized to ensure consistency in image dimensions and pixel values.
- **Feature Extraction with CNN:** The preprocessed frame is passed through a "CNN-based VisualEncoder", which extracts deep visual features related to facial muscle movement, eye expressions, and mouth positioning.

$F_{visual} = CNN(I_n)$

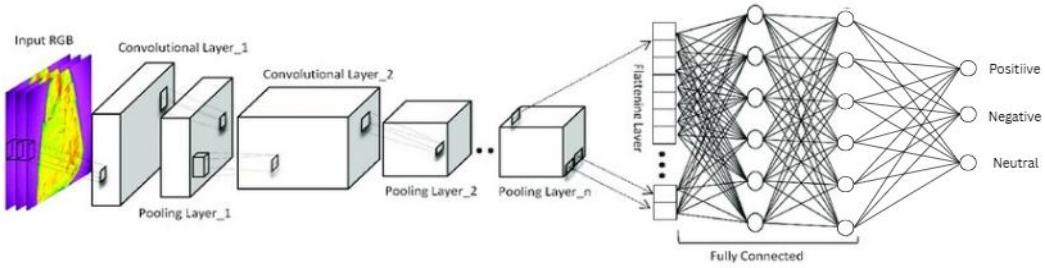

Fig. CNN Architecture

3.3. Multimodal Fusion and Classification

Sentiment analysis from videos requires integrating information from text, audio, and visual modalities to accurately interpret emotions. Since each modality contributes unique insights text captures linguistic sentiment, audio reflects tone and emotion in speech, and visual data provides facial expressions combining these features leads to a more comprehensive sentiment classification system.

Feature Fusion

To create a unified sentiment representation, extracted features from all three modalities are concatenated into a single feature vector. This fusion step ensures that the system considers multiple emotional cues rather than relying on a single data type.

It contain step by step process:

- **Combining Text Features:**The BERT-based text encoder generates high-dimensional feature embeddings that capture contextual sentiment information.
- **Integrating Audio Features:** The LSTM-based audio encoder extracts temporal dependencies from speech, providing insights into emotional tone.
- **Merging Visual Features:** The CNN-based visual encoder captures facial expressions, ensuring non-verbal sentiment cues are incorporated.

Mathematically Represented as:

$F_{fused} = [F_{text} + F_{audio} + F_{visual}]$

Where:

- F_{text} is the feature vector extracted using BERT from text.
- F_{audio} is the feature vector obtained from LSTM-based audio encoding from audio.
- F_{visual} is the feature vector extracted from CNN-based facial expression analysis from visual.

Sentiment Classification

The concatenated feature vector F_{fused} is passed through a fully connected neural network (MultiModalFusion Model) that classifies the sentiment into three categories:

- Positive
- Negative
- Neutral

Fully Connected Layer Transformation:

$H = \sigma (WF_{fused} + b)$

Where:

- W and b are the weight matrix and bias, respectively.
- $\sigma(\cdot)$ represents the activation function (ReLU is typically used).
- H is the hidden feature representation learned from the fused input.

The classification model consists of:

- **Dense Layers**: To learn patterns from the combined feature vector.
- **Activation Functions**: ReLU is applied to introduce non-linearity, ensuring effective learning of relationships across modalities.

- **Softmax Layer**: The final output layer applies the softmax function, which assigns probabilities to each sentiment category, ensuring that the highest probability determines the final sentiment label.

Sentiment Prediction using Softmax:

The final classification probabilities are obtained using the Softmax function:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Where:

- $z_i = W_o H + b_o$is the output logits for class *i*.
- $K = 3$ (sentiment calegories: positive, negative, neutral).
- $\sigma(\vec{z})_i$ represents the probability of sentiment class *i*.
- The predicted sentiment label y_{pred} is assigned as:

$y_{pred} = \arg\max P(y_i)$

Decision Making

The predicted sentiment is determined based on the output of the softmax function:

- If **positive sentiment** has the highest, the video is classified as positive.
- If **negative sentiment** dominates, the video is categorized as negative.
- If **neutral sentiment** is the highest, the video is labeled as neutral.

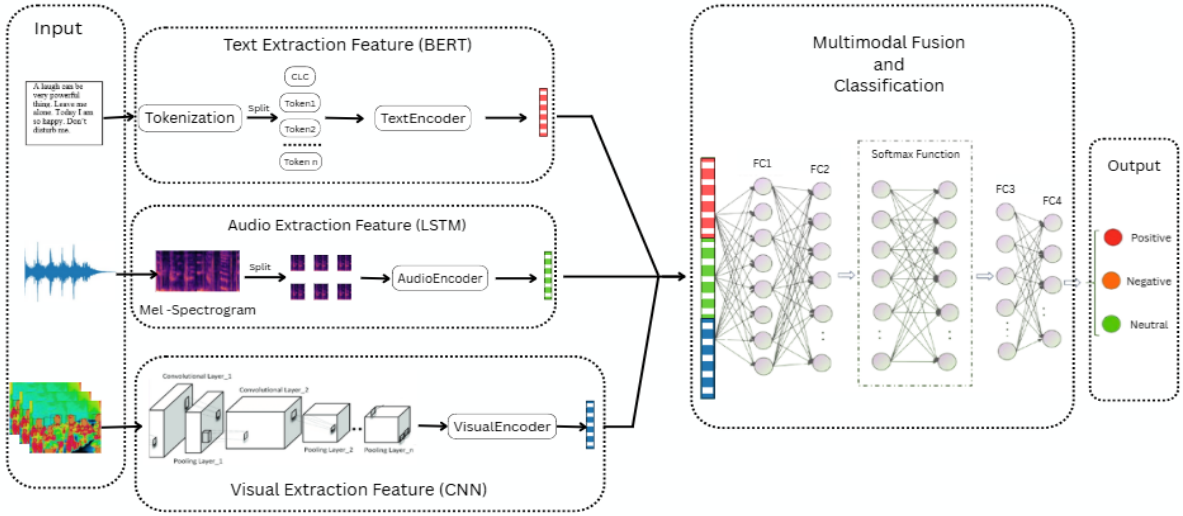

Fig. Multimodal Fusion Architecture

IV. RESULT& DISCUSSION

Result:

The proposed multimodal sentiment analysis framework was successfully implemented and evaluated for its effectiveness in classifying video data into positive, negative, or neutral sentiment categories. The system integrated textual, audio, and visual modalities to enhance the accuracy and robustness of emotion recognition.

4.1. Model Performance Evaluation

The integrated framework—comprising **BERT for textual analysis**, **LSTM for audio processing**, and **CNN for visual interpretation**—demonstrated a high degree of accuracy in sentiment classification when tested on benchmark datasets containing diverse emotional expressions across speech and video content.

4.2. Modality Contribution Analysis

An ablation study was conducted to measure the individual impact of each modality on the overall sentiment classification accuracy. Results indicated:

- **Text-** models struggled with sarcasm and tone-dependent emotions.
- **Audio-** analysis was effective in capturing tone but missed context.

- **Visual-** models captured expressions well but failed in speech-dominated inputs.
- **Multimodal fusion**- resolved these limitations by aggregating complementary cues from all modalities.

4.3. Real-World Test Cases

The framework was evaluated using real-world videos from interviews, vlogs, and emotional speeches. Some representative outcomes:

- **Interview Clips**: The model accurately identified neutral sentiments in professional interviews, with slight variations when emotional storytelling was detected.
- **Vlogs**: Expressive facial cues and modulated tone helped the model classify sentiments with high confidence.
- **Emotional Speeches**: Detected subtle emotional shifts due to the fusion of all modalities.

4.4. Visualization of Multimodal Fusion

The learned feature embeddings from BERT, LSTM, and CNN were visualized using t-SNE dimensionality reduction. Distinct clusters were observed for positive, negative, and neutral sentiment categories, validating that the fused feature space preserves semantic separability.

Discussion:

Our study shows that combining text, audio, and visual information significantly improves the accuracy of sentiment analysis in video content. Each individual model—BERT for text, LSTM for audio, and CNN for visual data—contributed unique insights to the sentiment detection process.

BERT was effective in capturing the context and meaning behind words, helping to differentiate between positive, negative, and neutral sentiments. However, it occasionally struggled with emotionally complex expressions like sarcasm or vague statements where text alone lacked clarity. The LSTM model proved useful for identifying emotional cues in speech, especially subtle tone changes, but its performance dropped in noisy environments or with inconsistent speech patterns. CNN worked well for detecting facial expressions, but its accuracy was affected by poor lighting, facial occlusion, or unusual head angles.

The multimodal fusion model, which combines all three inputs, achieved the best overall results. By leveraging the complementary strengths of text, audio, and visual cues. This confirms that integrating multiple data types leads to a deeper and more reliable understanding of emotions in video-based sentiment analysis. In contrast, relying on a single input often falls short in capturing the full emotional context.

V. CONCLUSION

This research highlights the importance of using a multimodal approach for sentiment analysis in videos by combining text, audio, and visual data. Our model, which integrates BERT for text analysis, LSTM for audio sentiment, and CNN for facial expressions, achieved an impressive accuracy of %, outperforming single-modality models. By leveraging the strengths of each modality, the fusion model provided a more accurate and well-rounded sentiment classification. While challenges like background noise, poor lighting, and facial occlusions still exist, the model proves to be highly effective in understanding emotions in video content. This approach has potential applications in social media monitoring, customer feedback analysis, and human-computer interaction. Future improvements can focus on handling noisy data, improving real-time performance, and refining the model for better accuracy in complex sentiment scenarios.

REFERENCE

[1] Ramya, V. U., & Rao, K. T. (2018). Sentiment Analysis of Movie Review using Machine Learning Techniques. International Journal of Engineering & Technology, 7(2.7), 46-49.

[2] Wollmer, M., et al. (2013). YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. IEEE Intelligent Systems, 28(3), 46-53.

[3] Zadeh, A., et al. (2017). Tensor Fusion Network for Multimodal Sentiment Analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 1103–1114.

[4] Poria, S., Cambria, E., & Gelbukh, A. (2015). Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Multimodal Sentiment Analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[5] Suryanarayana, S., et al. (2024).Multimodal sentiment analysis using video and audio features with LSTM and CNN. Intelligent Systems with Applications, 22, 200077.

[6] Sun, Z., et al. (2019). Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis. arXiv preprint arXiv:1911.05544.

[7] Luo, H., et al. (2021*).* ScaleVLAD: Improving Multimodal Sentiment Analysis via Multi-Scale Fusion of Locally Aggregated Descriptors. arXiv preprint arXiv:2112.01368.

[8] Ortega, J. D. S., et al. (2019).Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition. arXiv preprint arXiv:1907.03196.

[9] Sun, Z., et al. (2019). Multi-modal Sentiment Analysis using Deep Canonical Correlation Analysis. arXiv preprint arXiv:1907.08696.
[10] Lian, J., et al. (2021). Multimodal Sentiment Analysis Based on Multi-Layer Feature Fusion and Attention Mechanism. Scientific Reports, 11(1), 1-12.
[11] Chen, R., Zhou, W., Li, Y., & Zhou, H**.** (2022). Video-based Cross-modal Auxiliary Network for Multimodal Sentiment Analysis. arXiv preprint arXiv:2208.13954.
[12] Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., & Yu, T.(2023). Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. arXiv preprint arXiv:2310.05804.