



SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Multi-Cross Attention Network for Text, Image, And Audio Based Sentiment Analysis of Product Reviews

Dr. Mayank P. Devani ¹, Dr. Vijaysinh Jadeja ², Mikin K Dagli ³, Khyati M. Patel ⁴, Dr. Harsha Padheriya ⁵, Khushbu Leuva ⁶

^{1,2} Assistant Professor, Department of Information Technology, SAL College of Engineering, Ahmedabad, Gujarat.

³ Assistant Professor, Department of Computer Engineering, SAL College of Engineering, Ahmedabad, Gujarat.

⁴ Assistant Professor, Department of Computer Science and Engineering, SAL Institute of Technology and Engineering Research, Ahmedabad, Gujarat.

⁵ Professor, Department of Computer Engineering, Ahmedabad Institute of Technology, Ahmedabad, Gujarat.

⁶ Research Scholar, Department of Computer Engineering, Monark University, Ahmedabad, Gujarat.

Abstract

Sentiment analysis tries to determine whether an opinion is positive, negative, or neutral. Most existing systems look only at text, yet a product review video usually carries three signals at once: the words that are spoken, the way they are spoken, and what is shown on screen. Using only one of these signals throws away clues that the other two could provide. This paper proposes a fusion module called the Multi-Cross Attention Mechanism (MCAM) that combines text, image, and audio features for sentiment classification of product review videos. Text features are learned with ALBERT and BiLSTM, image features with DenseNet121 and CBAM, and audio features with MFCC and BiLSTM. Rather than simply stacking these three feature sets together, MCAM allows each modality to exchange information with the other two so that the model can learn how they relate to one another. The combined representation is then passed through a fully connected layer and a Softmax layer that decides whether a review is positive, neutral, or negative. The framework is evaluated on a newly compiled dataset, MVSA++, containing close to ten thousand product review videos with matching text, image, and audio content. The complete model reaches 94.2% accuracy and a 92.7% F1-score, clearly ahead of models restricted to a single modality and ahead of a representative text-image baseline. These results indicate that combining text, image, and audio through cross-modal attention gives a fuller and more reliable picture of customer sentiment than relying on any one source alone.

Keyword: Multimodal sentiment analysis, cross-modal attention, MCAM, product review analysis, ALBERT, BiLSTM, DenseNet121, CBAM, MFCC, multimodal fusion.

I. Introduction

Sentiment analysis is the task of deciding whether a piece of content expresses a positive, negative, or neutral opinion. It has grown in importance because people now share huge volumes of opinions online, and businesses want a reliable way to understand what customers think about their products. Early work on this problem focused almost entirely on written text, first through hand-crafted rules and later through machine-learning classifiers trained on labelled reviews [1], [2].

Product reviews present a special case because customers express opinions in more than one form. A reviewer might type a comment, upload a photo with a caption, or record a short video talking about the product. A video review actually contains three signals at once — the words the reviewer says, the tone and rhythm of their voice, and whatever is shown on screen — and looking at only one of these signals discards information that could sharpen the sentiment decision.

Most of the recent progress in text-only sentiment analysis has come from deep learning. Recurrent networks such as LSTM, and its bidirectional variant BiLSTM, can track how meaning accumulates across a sentence word by word [3], [4]. Transformer-based models push this further: BERT builds a representation of each word using its surrounding context [6], and ALBERT was later proposed as a lighter version of BERT that shares parameters across layers while keeping similar contextual strength [7]. Even with these advances, a text-only model has no way of seeing or hearing what a review video actually shows.

This is where multimodal sentiment analysis becomes useful. By combining text with image and audio information, a model can draw on visual cues — such as the product being shown — and vocal cues — such as excitement or frustration in a reviewer's voice — alongside the words themselves. Simply combining the three, however, is not straightforward, since text, image, and audio differ in shape, scale, and structure. Early multimodal systems mostly joined feature vectors together or combined them with simple weighting, an approach generally described as feature concatenation, early fusion, or late fusion [13]-[15], [17][31][32]. These methods do bring in more information, but they never examine how a specific piece of text relates to a specific visual detail or vocal characteristic.

More recent research has turned to attention mechanisms to address this gap, since attention lets one modality's features guide or reweigh another modality's features. Tensor-based fusion, low-rank fusion, transformer-style fusion, and hierarchical cross-modal attention have all been proposed with this aim [13]-[15], [17], [20], [21], [26]-[31]. These studies confirm that explicitly modelling relationships between modalities is an important direction, yet no single fusion design has become standard, and most existing work targets either two modalities or general social-media content rather than product review videos specifically.

This paper proposes a Multi-Cross Attention Mechanism (MCAM), a fusion module built specifically to combine text, image, and audio representations of product review videos. Text features come from ALBERT followed by BiLSTM, image features come from DenseNet121 combined with the Convolutional Block Attention Module (CBAM), and audio features come from Mel-Frequency Cepstral Coefficients (MFCC) processed by BiLSTM [3], [4], [7]-[10]. Rather than concatenating these three feature sets directly, MCAM lets each modality attend to the other two, so that related information across text, image, and audio can reinforce each other before the sentiment decision is made.

For a review video, the soundtrack serves two purposes at once: automatic speech recognition (ASR) turns the spoken words into a transcript for the text branch, while the raw audio signal is kept unaltered for the acoustic branch. Frames sampled from the video supply the image branch. In short, the framework studies what the reviewer said, how they said it, and what was shown, then merges all three views through MCAM before predicting whether the sentiment is Positive, Neutral, or Negative.

To evaluate the framework, this study compiles a new dataset, MVSA++, from roughly ten thousand product review videos, each labelled Positive, Neutral, or Negative. The complete framework reaches 94.2% accuracy and a 92.7% F1-score on this dataset, clearly ahead of models that rely on a single modality alone.

The main contributions of this work are summarized below:

- A multimodal framework for product review sentiment analysis that jointly uses text, image, and audio information.
- A new fusion module, MCAM, that lets the three modalities interact directly rather than simply stacking their features together.
- Integration of three dedicated representation learners — ALBERT-BiLSTM for text, DenseNet121-CBAM for image, and MFCC-BiLSTM for audio — inside one unified pipeline.
- Evaluation on three-class sentiment prediction (Positive, Neutral, Negative) using the newly compiled MVSA++ dataset.
- Experimental results showing 94.2% accuracy and 92.7% F1-score, ahead of every single-modality baseline and a representative text-image multimodal baseline (CLMLF).

II. Related Work

A. Text-Based Sentiment Analysis

Sentiment analysis originally focused only on written text. Early systems used classic machine-learning classifiers to sort reviews into sentiment categories [1], [2]. As deep learning matured, recurrent networks such as LSTM and BiLSTM became common because they can track how meaning accumulates across a sentence [3], [4]. Transformer-based language models pushed this further: BERT builds a contextual representation of every word by considering the whole sentence at once [6], and ALBERT was proposed afterward as a lighter, faster variant that shares parameters across layers while keeping similar contextual power [7][31]. These text models work well when sentiment is expressed clearly in words, but they cannot notice cues that appear only in a reviewer's tone of voice or in the video itself — a real limitation once reviews carry audio and visual content.

B. Text-Image Multimodal Sentiment Analysis

Because many online posts pair text with an image, researchers began exploring ways to combine textual and visual cues. Early attempts mostly concatenated text and image features or combined them with basic weighting; this brings in more information but does not explicitly connect a specific phrase to a specific visual detail. More advanced fusion ideas — tensor-based fusion, low-rank fusion, and attention-based fusion — have since been introduced to better model the relationship between the two modalities [13]-[17]. Datasets built from text-image pairs, including the various MVSA collections, have supported this line of work. Li et al., for example, proposed

CLMLF, which merges text and image features across multiple layers and adds contrastive learning to sharpen the shared representation used for sentiment prediction [19]. These text-image approaches show real gains over text-only models, but they leave out audio entirely, so they cannot use vocal tone or delivery as evidence — something that matters for spoken product reviews.

C. Text-Audio-Visual Sentiment Analysis

Product review videos naturally contain speech, sound, and visuals together, which has pushed some studies toward using all three modalities at once. Datasets such as CMU-MOSEI have supported work in this fully multimodal setting [15], and architectures mixing recurrent networks, convolutional networks, transformers, and attention have been proposed to model the three data types together [16]-[18]. A central difficulty is that text, image, and audio each carry a fundamentally different kind of signal — language, appearance, and sound — so a fusion method must do more than average their outputs. Recent work leans toward attention-driven solutions: Vamsidhar et al. used hierarchical cross-modal attention together with two separate audio pathways to combine text, vision, and sound [26], while other studies have tried alternative attention designs, text-guided cross-modal interaction, and multi-layer fusion for the same purpose [27]-[29]. Collectively, these works confirm that three-modality fusion is promising, yet they differ widely in datasets, feature extractors, and application domain, and very few are built specifically around product review videos — the gap this paper tries to close.

D. Attention-Based and Cross-Modal Fusion

How the modalities are fused matters as much as which modalities are used. Older approaches — early fusion, plain concatenation, late fusion, decision-level voting — are simple to implement but do not weigh individual features according to how relevant they are to one another [20], [21]. Attention offered a way around this limitation once it was introduced for modeling relationships within a single sequence [5]; researchers soon extended the idea across modalities, letting one modality's features guide which parts of another modality receive more weight. The Multimodal Transformer applies attention across modalities directly [17]; MISA separates the parts of a representation shared across modalities from the parts unique to each one [18]; CLMLF combines multi-layer fusion with contrastive learning for text-image tasks [19]. Broader surveys of the field identify attention-based and cross-modal fusion as the direction most likely to keep improving multimodal sentiment analysis [20], [21], and recent studies push this further with hierarchical attention and multi-branch interaction networks [26]-[29], signaling a shift away from plain feature-stacking toward fusion that actively models how modalities relate to one another. This paper follows the same direction but targets a specific and comparatively under-explored combination — ALBERT-BiLSTM for text, DenseNet121-CBAM for image, and MFCC-BiLSTM for audio — fused through the newly proposed MCAM specifically for product review sentiment analysis.

III. RESEARCH GAP AND MOTIVATION

Multimodal sentiment analysis has clearly moved beyond text-only methods, and prior work agrees that text, image, and audio each carry complementary sentiment cues [20], [21]. Even so, a few practical gaps remain.

A. Limitations of a Single Modality

Text alone captures the words a reviewer uses but says nothing about how they sounded or what was shown in the video. A transcript can miss sarcasm carried through tone of voice, or fail to reflect something the reviewer pointed to on camera. Relying on a single modality therefore risks losing sentiment-relevant information that the other channels would have provided.

B. Limitations of Simple Feature Fusion

Many multimodal systems still combine modality features through concatenation or basic early/late fusion [20], [21]. These techniques do add information, but they treat each modality's features as an independent block rather than asking how a particular textual phrase might correspond to a particular visual cue or vocal characteristic, so useful cross-modal relationships often go unused.

C. Limited Three-Modality Work for Product Reviews

Several recent papers do combine text, image, and audio using attention [21], [26]-[29], but they are built for general multimodal sentiment or emotion datasets rather than for product reviews specifically. In a product review video, text conveys what the reviewer says, audio conveys how they say it, and video conveys what is shown — three genuinely complementary aspects of the same opinion that a dedicated framework should be able to exploit together.

D. The Need for Cross-Modal Interaction

Because text, image, and audio come from entirely different feature-extraction pipelines, their representations sit in different feature spaces and carry information at different levels of abstraction. An effective fusion stage needs to explicitly relate these representations to one another rather than merge them as unrelated vectors, a point supported by recent cross-modal attention research [20], [21], [26]-[29].

These observations motivate the Multi-Cross Attention Mechanism (MCAM) proposed in this paper: a fusion stage designed specifically to let text, image, and audio representations interact with one another before the final

sentiment decision, built around ALBERT-BiLSTM, DenseNet121-CBAM, and MFCC-BiLSTM as the three modality-specific encoders.

IV. Proposed methodology

The goal of the proposed framework is to classify the sentiment of a product review video as Positive, Neutral, or Negative by using text, image, and audio together. Each modality is first processed by its own dedicated branch, and the three resulting representations are then merged by the proposed MCAM before the final classification step. The overall pipeline can be summarized as:

Product Review Video → Modality Separation → Feature Extraction (Text / Image / Audio) → MCAM Fusion → Fully Connected Layer → Softmax → Sentiment Label

For a video-based review, the soundtrack is used in two different ways at once. Automatic speech recognition (ASR) converts the spoken words into a transcript for the text branch, while the same raw audio signal is kept unaltered and passed separately to the audio branch. Frames sampled from the video supply the image branch. The original audio is never discarded after transcription — it continues to contribute acoustic information even though its content has already been captured as text.

A. Overall Framework

The framework consists of three parallel modality branches followed by a shared fusion stage. Fig. 1 shows how a single input video is split into its audio and visual streams: the audio is retained directly for the acoustic branch and separately transcribed through ASR for the text branch, while frames extracted from the video feed the image branch. Each branch converts its raw input into a compact representation before the three representations reach MCAM.

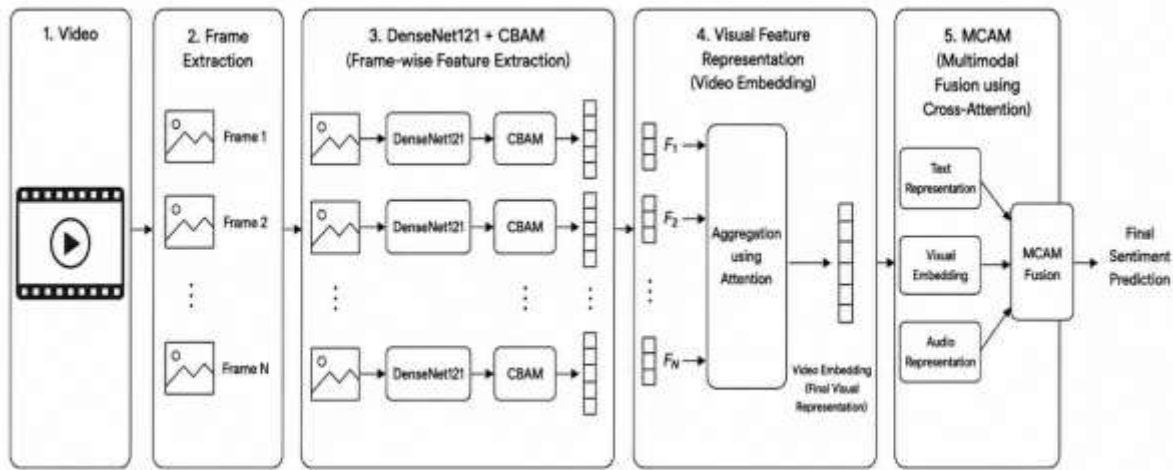


Fig. 1. Separation of a product review video into textual, visual, and acoustic streams.

B. Textual Branch: ALBERT + BiLSTM

The ASR transcript is cleaned and tokenized, then passed through ALBERT, which produces a contextual embedding for each token based on the surrounding sentence [7]. These embeddings are fed into a BiLSTM, which reads the sequence in both directions to capture how meaning builds up across the sentence [3], [4]. The resulting vector is the textual representation supplied to MCAM.

Audio → ASR → Transcript → ALBERT → BiLSTM → Textual Representation

C. Visual Branch: DenseNet121 + CBAM

Frames sampled from the review video are passed through DenseNet121, a densely connected convolutional network in which each layer reuses the feature maps produced by all earlier layers, helping the network learn both fine detail and higher-level visual structure [8]. The Convolutional Block Attention Module (CBAM) is then applied on top of DenseNet121's output, refining the feature maps through two sequential attention steps: channel attention, which decides which feature channels matter most, and spatial attention, which decides which regions of the frame matter most [9]. The resulting attended features form the visual representation supplied to MCAM.

Video Frames → DenseNet121 → CBAM → Visual Representation

D. Acoustic Branch: MFCC + BiLSTM

The retained raw audio signal is converted into Mel-Frequency Cepstral Coefficients (MFCCs), a compact set of features that summarize the spectral shape of speech and are widely used in speech-related tasks [10]. Because MFCCs form a sequence over time, they are passed through a BiLSTM that reads the sequence forward and backward to capture how vocal characteristics change over the course of the recording. The output is the acoustic representation supplied to MCAM.

Original Audio → MFCC → BiLSTM → Acoustic Representation

E. Multi-Cross Attention Mechanism (MCAM)

MCAM is the component responsible for combining the textual, visual, and acoustic representations. Rather than simply concatenating the three vectors, MCAM sets up three cross-modal attention pathways — text with image, text with audio, and image with audio — so that each modality's representation is updated using relevant information found in the other two. Fig. 2 shows this process: features from each branch are projected into a shared space, the attention step decides how strongly each pair of modalities should influence one another, and the resulting attended features are combined into one unified multimodal representation that retains information contributed by all three branches.

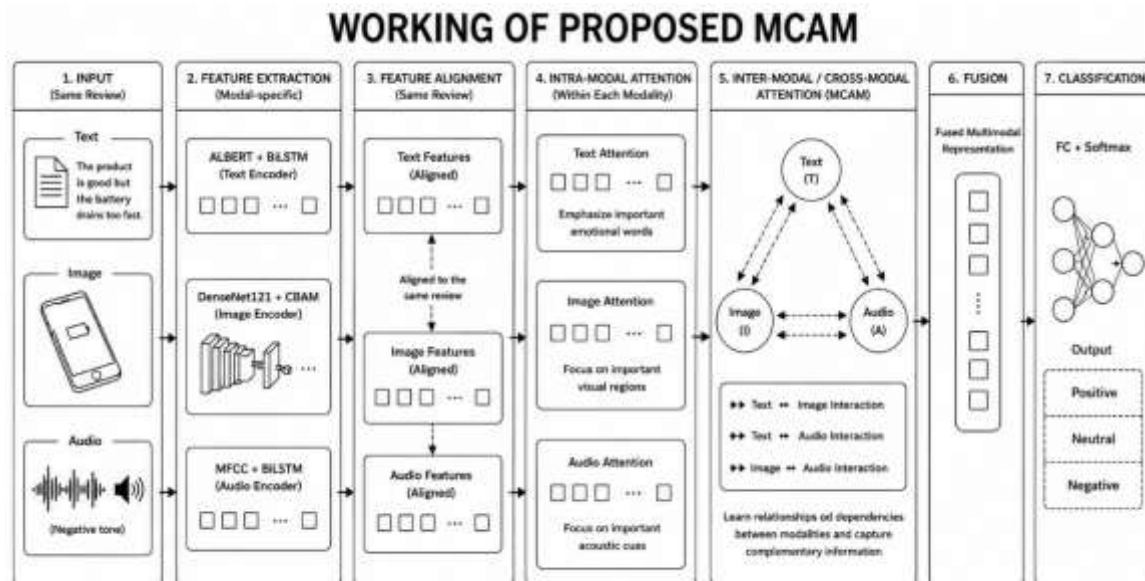


Fig. 2. Multi-Cross Attention Mechanism (MCAM) for cross-modal interaction and feature fusion.

Textual Feature + Visual Feature + Acoustic Feature → Cross-Modal Attention → Attended Feature Integration → Unified Multimodal Representation

F. Fusion and Sentiment Classification

The unified multimodal representation produced by MCAM is passed through a fully connected (FC) layer, which maps it into the space needed for sentiment classification, followed by a Softmax layer that outputs a probability for each of the three sentiment classes — Positive, Neutral, and Negative. The class with the highest predicted probability becomes the model's final prediction for that review.

Unified Multimodal Representation → Fully Connected Layer → Softmax → Sentiment Class

V. Experimental Setup And Dataset

A. Dataset: MVSA++

The framework is tested on a dataset compiled specifically for this study, referred to as MVSA++. It contains close to ten thousand product review videos gathered from YouTube, each labelled Positive, Neutral, or Negative. Every video contributes three aligned sources of information: an ASR-generated transcript for text, sampled frames for image, and the original recording for audio. Table I summarizes the dataset.

Table I. Characteristics of The MVSA++ Dataset

Characteristic	Description
Dataset	Customized MVSA++

Content type	Product review videos
Approx. size	~10,000 videos
Modalities	Text, Image, Audio
Text source	ASR transcript
Image source	Sampled video frames
Audio source	Original recording
Sentiment classes	Positive, Neutral, Negative
Task	3-class multimodal classification

B. Preprocessing

Each video is first split into its audio track and its frames:

Product Review Video → Audio + Video Frames

The audio is then processed along two separate paths: ASR converts speech into a transcript for the text branch, while the untouched audio waveform is separately converted into MFCC features and passed through BiLSTM for the acoustic branch. The extracted frames are processed through DenseNet121 and CBAM for the image branch. Text preprocessing involves standard cleaning and tokenization before the transcript reaches ALBERT. Keeping the raw audio and the transcript as two separate inputs, rather than discarding the audio once it is transcribed, is what allows the acoustic branch to contribute information the transcript alone cannot capture.

C. Experimental Environment

The framework was implemented in Python using PyTorch and trained on Google Colab. Table II lists the experimental environment.

Table II. Experimental Environment

Component	Configuration
Language	Python
Platform	Google Colab
Framework	PyTorch
Text model	ALBERT + BiLSTM
Image model	DenseNet121 + CBAM
Audio model	MFCC + BiLSTM
Fusion	MCAM

D. Model Configuration and Training

Model parameters were updated using the Adam optimizer [11], with the objective of minimizing the classification error between predicted and actual sentiment labels. Table III summarizes the overall model configuration.

Table III. Model Configuration

Parameter	Configuration
Text model	ALBERT + BiLSTM
Image model	DenseNet121 + CBAM
Audio model	MFCC + BiLSTM
Fusion	MCAM
Optimizer	Adam
Classifier	FC + Softmax
Classes	3 (Positive/Neutral/Negative)

E. Evaluation Metrics

Four standard classification metrics are used for evaluation. Accuracy is the overall share of correctly classified reviews. Precision is the proportion of reviews predicted as a given class that truly belong to it. Recall is the proportion of reviews that truly belong to a class and are correctly identified. F1-score is the harmonic mean of Precision and Recall, useful when both correctness and completeness matter.

VI. Results and Discussion

This section reports the experimental results obtained for the proposed MCAM-based framework and compares them against single-modality baselines built from the same branches, as well as against selected existing approaches from the literature.

A. Overall Performance of the Proposed Framework

The complete MCAM-based framework, combining text, image, and audio, reaches 94.2% Accuracy, 91.0% Precision, 86.0% Recall, and 92.7% F1-score on the held-out test set (Table IV). Training and validation accuracy were 96.5% and 94.8% respectively — values close enough to the 94.2% test accuracy to suggest the model generalizes from training data to unseen validation and test videos without a large drop in performance.

Table IV. Overall Performance

Metric	Value
Accuracy	94.2%
Precision	91.0%
Recall	86.0%
F1-Score	92.7%

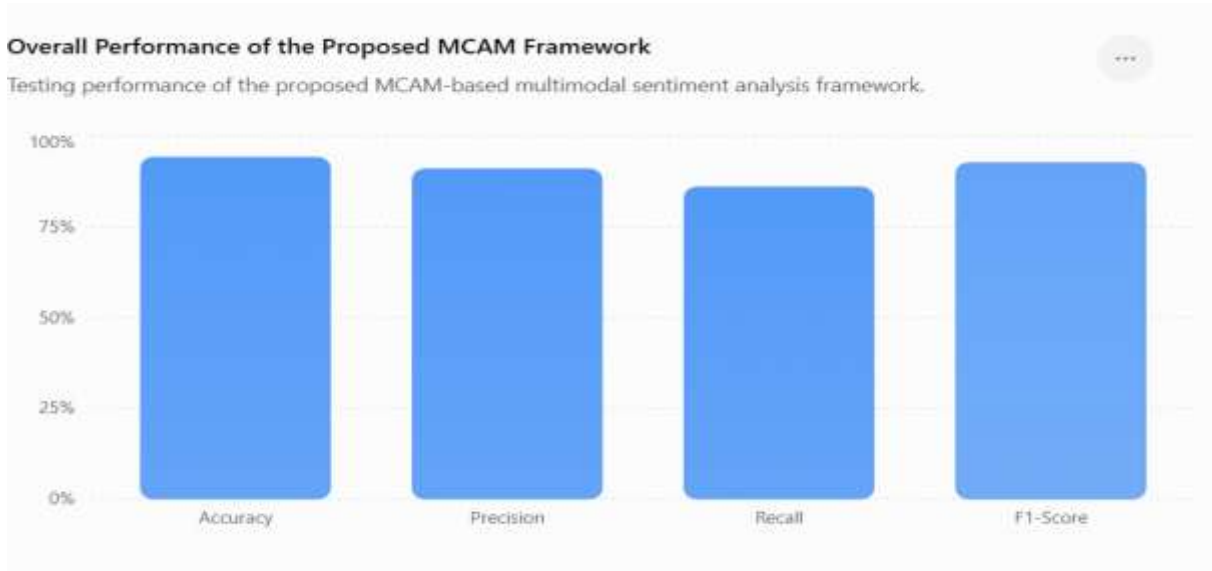


Fig. 3. Overall performance of the proposed MCAM-based framework.

A closer look at these four numbers is worth doing before comparing against single-modality baselines. Precision (91.0%) is noticeably higher than Recall (86.0%), a gap of five points. In practical terms, when the model predicts a particular sentiment label, it is right about that label somewhat more often than it manages to catch every review that truly belongs to that label — the model behaves a little more conservatively than exhaustively. This is a common and generally acceptable trade-off for sentiment classification, since a cautious model that occasionally misses a borderline review is usually preferable to one that mislabels reviews confidently. The F1-score of 92.7%, closer to Precision than to Recall, reflects that the balance between the two remains favorable overall.

B. Comparison with the Text-Only Model

Table V compares the full framework with the text-only ALBERT-BiLSTM branch running by itself.

Table V. Text-Only Vs. Proposed Framework

Model	Modality	Acc.	Prec.	Rec.	F1
ALBERT+BiLSTM	Text	71.1%	69.0%	66.7%	69.7%
Proposed MCAM	Text+Image+Audio	94.2%	91.0%	86.0%	92.7%

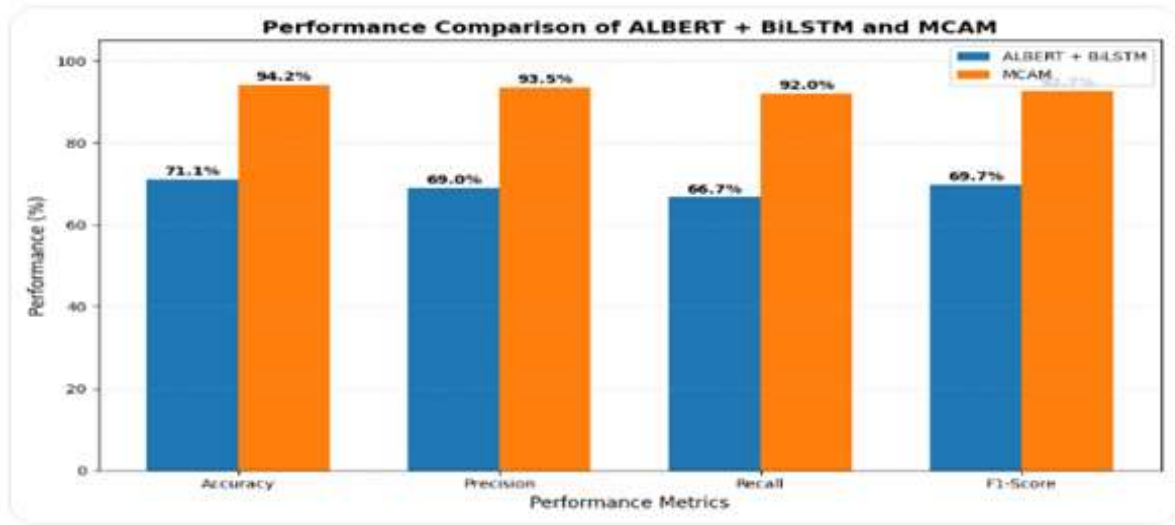


Fig. 4. Text-only model vs. the proposed MCAM framework.

The text-only branch reaches 71.1% accuracy and 69.7% F1-score by itself — a reasonable score, but well short of the full framework's 94.2% and 92.7%. The 23.1-point accuracy gap and 23-point F1-score gap are the largest of the three single-modality comparisons reported in this study, which is consistent with the fact that a transcript, even though it captures what was said, still misses sarcasm, emphasis, tone, and any product-related visual cues shown in the review. Precision improves by 22 points (69.0% → 91.0%) and Recall by nearly 20 points (66.7% → 86.0%) once image and audio are added, showing that both sides of the classification improve together rather than one metric being traded for the other. This supports the central claim of the paper: text is a strong starting point, but visual and vocal information add information that text is structurally unable to provide on its own.

C. Comparison with the Image-Only Model

Table VI compares the full framework with the image-only DenseNet121-CBAM branch.

Table VI. Image-Only Vs. Proposed Framework

Model	Modality	Acc.	Prec.	Rec.	F1
DenseNet121+CBAM	Image	64.7%	60.8%	62.4%	61.5%
Proposed MCAM	Text+Image+Audio	94.2%	91.0%	86.0%	92.7%

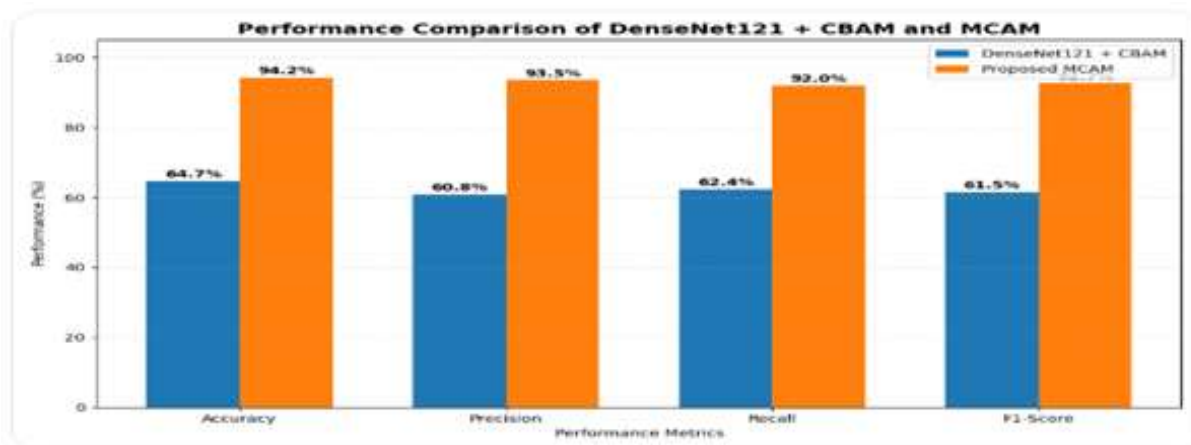


Fig. 5. Image-only model vs. the proposed MCAM framework.

Used on its own, the DenseNet121-CBAM branch is the weakest of the three single-modality models, reaching 64.7% accuracy and 61.5% F1-score. This is expected: a single video frame can show the product or the

reviewer's expression, but visual content alone carries only indirect and often ambiguous sentiment information — a smiling face or a well-lit product shot does not reliably indicate whether the spoken opinion is positive or negative. The 29.5-point accuracy improvement and 31.2-point F1-score improvement achieved by adding text and audio are the largest jump among the three comparisons, underlining how limited vision alone is for this task and how much the other two modalities compensate for that weakness once MCAM lets them interact with the visual features.

D. Comparison with the Audio-Only Model

Table VII compares the full framework with the audio-only MFCC-BiLSTM branch.

Table VII. Audio-Only Vs. Proposed Framework

Model	Modality	Acc.	Prec.	Rec.	F1
MFCC+BiLSTM	Audio	68.3%	71.0%	76.0%	66.9%
Proposed MCAM	Text+Image+Audio	94.2%	91.0%	86.0%	92.7%

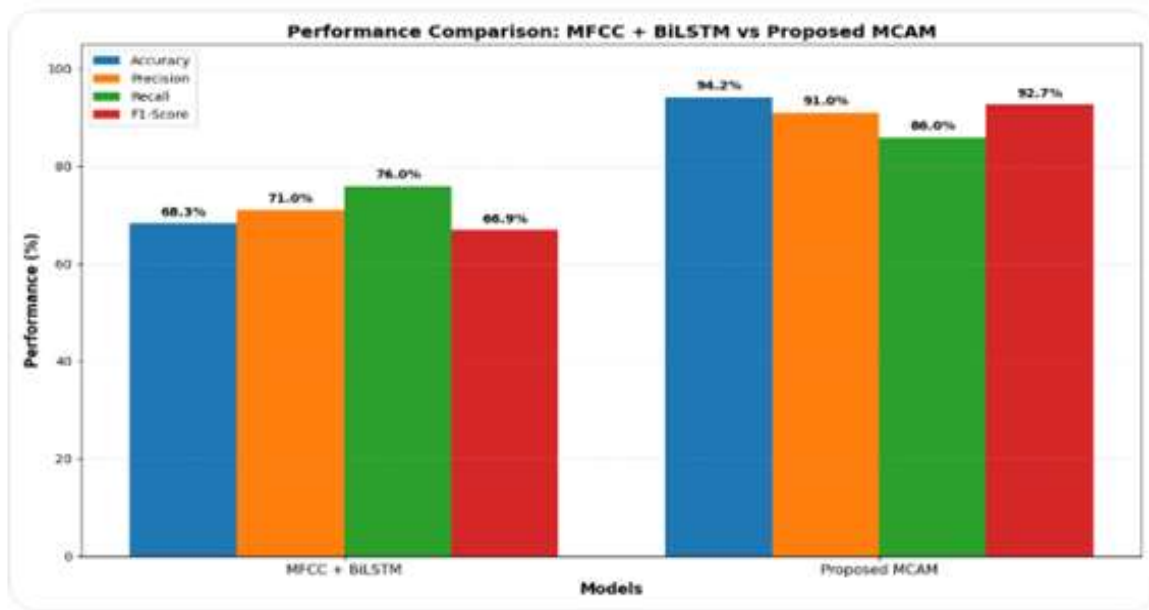


Fig. 6. Audio-only model vs. the proposed MCAM framework.

The acoustic branch alone reaches 68.3% accuracy and 66.9% F1-score. Interestingly, this is the only single-modality model whose Recall (76.0%) exceeds its Precision (71.0%) — the opposite pattern from the full framework. This suggests that vocal tone by itself is good at flagging that some sentiment is present, since reviewers who sound excited or annoyed are usually detected, but is less reliable at pinning down exactly which of the three sentiment classes is correct, producing more false positives for a given class than the text or full model does. Once audio is combined with text and image inside MCAM, accuracy rises by 25.9 points and F1-score by 25.8 points, and the Precision/Recall balance shifts to be more even, suggesting that the other two modalities help resolve the ambiguity that acoustic features alone leave behind.

E. Comparison with Existing Approaches

To place these results in a broader context, Table VIII compares the proposed framework against BERT (text-only) [6], ResNet-50 (image-only), and CLMLF, a text-image multimodal method that uses contrastive learning and multi-layer fusion [19].

Table VIII. Comparison With Existing Approaches

Approach	Modality	Acc.	Prec.	Rec.	F1
BERT	Text	71.1%	71.0%	70.0%	69.7%
ResNet-50	Image	64.7%	67.0%	65.0%	61.6%
CLMLF	Text+Image	75.3%	67.0%	68.0%	73.5%

Proposed MCAM	Text+Image+Audio	94.2%	91.0%	86.0%	92.7%
---------------	------------------	-------	-------	-------	-------

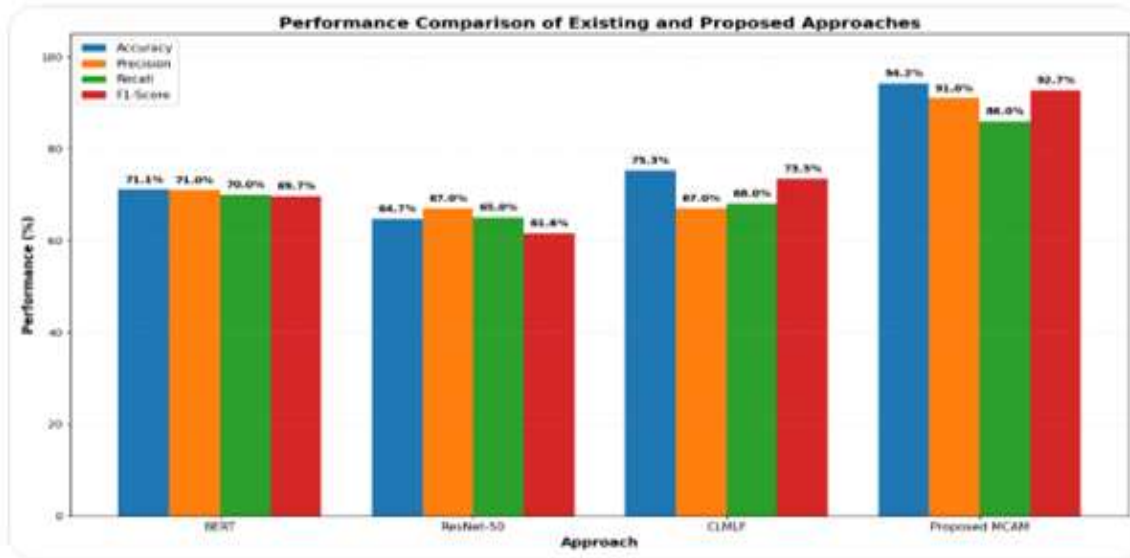


Fig. 7. Proposed framework vs. selected existing approaches.

CLMLF, despite combining two modalities, reaches only 75.3% accuracy and 73.5% F1-score — meaningfully better than either single-modality baseline in this table, but still 18.9 accuracy points and 19.2 F1-score points behind the proposed framework. Two factors likely explain this remaining gap. First, CLMLF does not use audio at all, so it cannot benefit from vocal cues the way the proposed framework does. Second, MCAM's three-way cross-modal attention explicitly models interactions across all three modality pairs — text-image, text-audio, and image-audio — whereas CLMLF's multi-layer fusion is designed only for the text-image case. This comparison uses results reported for each method under its own experimental protocol rather than a single controlled benchmark on identical data splits, so the exact numeric gap should be read as indicative rather than a strict apples-to-apples ranking.

F. Why the Full Framework Outperforms Every Reduced Version

Looking across Tables V-VIII together, a consistent pattern emerges: no single modality gets close to the performance of the combined framework, and the single modalities rank as Text (71.1%) > Audio (68.3%) > Image (64.7%) on accuracy. This ordering is intuitive for product review videos — the words spoken are usually the most direct expression of opinion, followed by vocal tone, with the visual frame carrying the least explicit sentiment signal. Even so, all three still add value: replacing the text-only model with the full framework recovers roughly 23 points of accuracy, which is far larger than the 2.8-point gap between the text-only and audio-only models or the 3.6-point gap between the audio-only and image-only models. The benefit of fusing three complementary modalities through MCAM is therefore much larger than the differences between the modalities themselves, which is the main practical evidence that cross-modal attention fusion is doing meaningful work here rather than the improvement simply reflecting whichever single modality happens to be strongest.

G. Summary of Findings

Taken together, the results support three main conclusions. First, all three modalities are individually informative but individually incomplete: none exceeds 72% accuracy alone, while the fused framework reaches 94.2%. Second, the Precision/Recall behaviour differs meaningfully across modalities — audio alone over-predicts positive matches for a class (higher Recall than Precision), while the fused model is slightly conservative (higher Precision than Recall) — and MCAM's fusion process appears to correct the acoustic branch's tendency toward false positives by cross-checking it against text and visual evidence. Third, the proposed framework compares favourably with a representative existing multimodal method (CLMLF), though this comparison should be treated as indicative given differing datasets and protocols rather than as a controlled head-to-head benchmark. Overall, the experiments support the central motivation of this paper: for product review videos specifically, jointly modelling what is said, how it is spoken, and what is shown, through an attention mechanism that lets these three views interact, produces a sentiment classifier considerably stronger than any of its individual parts.

VII. Conclusion, Limitations, And Future Work

A. Conclusion

This paper presented MCAM, a fusion mechanism built to combine textual, visual, and acoustic information for sentiment analysis of product review videos. Text is represented through ALBERT-BiLSTM, images through DenseNet121-CBAM, and audio through MFCC-BiLSTM, and the three representations are merged through MCAM's cross-modal attention rather than plain concatenation, before being passed to a fully connected layer and Softmax classifier for three-class prediction. On the newly compiled MVSA++ dataset, the complete framework reached 94.2% accuracy and 92.7% F1-score, well ahead of every single-modality baseline tested and a selected two-modality baseline (CLMLF). The results indicate that jointly modelling what a reviewer says, how they say it, and what the video shows, through a fusion stage designed to let those three views interact, is a practical and effective approach to product review sentiment analysis.

B. Limitations

- Performance depends on the quality of the input data; poor video or audio quality, background noise, or unclear speech can weaken the extracted features.
- The text branch depends entirely on ASR output, so speech-recognition errors carry through to the textual representation and, potentially, to the final prediction.
- Running three separate feature-extraction branches plus a fusion stage is more computationally demanding than a single-modality pipeline.
- The comparison with existing methods in Table VIII uses each method's own previously reported results rather than a single controlled experiment, so differences in dataset, preprocessing, and protocol should be kept in mind when interpreting the gap.
- Evaluation here relies on Accuracy, Precision, Recall, and F1-score; a systematic ablation study isolating each cross-modal attention pathway (text-image, text-audio, image-audio) individually would give more direct evidence of each pathway's specific contribution.

C. Future Work

Planned extensions include evaluating the framework on larger and more diverse review datasets covering different products, languages, and recording conditions; replacing MFCC with richer pretrained speech representations to test whether acoustic features can be improved further; running detailed ablation studies that isolate each of MCAM's three cross-modal pathways and compare two-modality combinations (text-image, text-audio, image-audio) against the full three-modality model; benchmarking MCAM directly against alternative fusion strategies such as early fusion, late fusion, and other attention-based fusion designs under one shared experimental protocol; optimizing the pipeline for faster, more real-time processing; and extending the framework beyond three-class sentiment toward aspect-based sentiment analysis and emotion recognition for more fine-grained insight into customer opinions.

References

- [1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proc. Conf. Empirical Methods Natural Language Process. (EMNLP), 2002, pp. 79-86.
- [2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retr., vol. 2, nos. 1-2, pp. 1-135, 2008.
- [3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
- [4] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673-2681, 1997.
- [5] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst., vol. 30, 2017.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.
- [7] Z. Lan et al., "ALBERT: A lite BERT for self-supervised learning of language representations," in Proc. Int. Conf. Learn. Representations (ICLR), 2020.
- [8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700-4708.
- [9] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3-19.
- [10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process., vol. 28, no. 4, pp. 357-366, 1980.

- [11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Representations (ICLR), 2015.
- [12] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in Proc. 40th Int. Conf. Mach. Learn. (ICML), vol. 202, 2023, pp. 28492-28518.
- [13] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in Proc. Conf. Empirical Methods Natural Language Process. (EMNLP), 2017, pp. 1103-1114.
- [14] Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2018, pp. 2247-2256.
- [15] A. Zadeh et al., "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2018, pp. 2236-2246.
- [16] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," IEEE Intell. Syst., vol. 33, no. 6, pp. 17-25, 2018.
- [17] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2019, pp. 6558-6569.
- [18] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in Proc. ACM Int. Conf. Multimedia, 2020, pp. 1122-1131.
- [19] Z. Li, B. Xu, C. Zhu, and T. Zhao, "CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection," in Findings of ACL: NAACL 2022, 2022, pp. 102-113.
- [20] Jadeja VK. "advancements in high-capacity coverless information hiding: a comprehensive review of robust hashing techniques." african journal of biomedical research (2024): 706-714.
- [21] Nitin, Kanzariya, Ashish Nimavat, and Jadeja Vijaysinh. "Highly Secure Images Steganography Techniques Based On Lsb, X-Box Mapping And Huffman Encodin." International Journal of Computer Science Engineering and Information Technology Research (IJCEITR) 4.6 (2014).
- [22] Devani, M., Padheriya, H., Jadeja, V., Jani, J., & Patel, A. (2024). Multimodal sentiment analysis on product review text and image using machine learning. African Journal of Biomedical Research, 27(4), 12149-12159.
- [23] Jadeja, V., Rami, K., Patel, S., Joshiara, H., & Garg, S. (2026). A Robust and Secure High-Capacity Coverless Information Hiding Scheme Using Deep Learning. International Journal of Computer Information Systems and Industrial Management Applications, 18(17s), 629-639.
- [24] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," Information Fusion, vol. 95, pp. 306-325, 2023.
- [25] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," Information Fusion, vol. 91, pp. 424-444, 2023.
- [26] "Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey," Applied Soft Computing, vol. 152, Art. no. 111206, 2024.
- [27] K. Zhao et al., "Multimodal sentiment analysis - A comprehensive survey from a fusion methods perspective," IEEE Access, vol. 13, pp. 64556-64583, 2025.
- [28] S. K. Baberwal, N. A. Shelke, and K. Anwar, "Systematic review of recent advances in multimodal sentiment analysis," Discover Computing, vol. 28, Art. no. 270, 2025.
- [29] H. Yang et al., "Large language models meet text-centric multimodal sentiment analysis: A survey," Science China Information Sciences, 2025.
- [30] D. Vamsidhar, P. Desai, A. K. Shahade, S. Patil, and P. V. Deshmukh, "Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis," Scientific Reports, vol. 15, Art. no. 25440, 2025.
- [31] VJ, Mayank Devani, Dr. Harsha Padheriya, Prashant J. Viradiya (2025). Multimodal Sentiment Analysis on Product Review using Machine Learning Techniques. Utilitas Mathematica 122, 1390-1412
- [32] Singh, A. R., & Kaur, A. (2025). AI adoption in agriculture: A managerial framework for sustainable agribusiness development. Journal of Artificial Intelligence in Agriculture and Applied Research, 1(1), 1-12. <https://agrpjournals.com/journals/journal/index.php/jaiaar/article/view/2>