IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 59, No. 5, May, 2026, pp. 1437-1460
Journal homepage: http://iieta.org/journals/jesa

A Dynamic Cross Level Leveraged Multi-Folded Multi-Model Sentimental Analysis Based Classification Framework

Nusrath Konnola[1,2*], Sarojini Balakrishnan[1]

[1] Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore 641043, India
[2] Department of Computer Science, MES Ponnani College, Malappuram 679586, India

Corresponding Author Email: 21phcsp006@avinuty.ac.in

https://doi.org/10.18280/jesa.590521

Received: 12 March 2026
Revised: 2 May 2026
Accepted: 10 May 2026
Available online: 31 May 2026

Keywords:
multimodal sentiment analysis, dynamic cross-level fusion, multi-folded feature encoder, attention mechanism, adaptive softmax, weighted regularized loss

ABSTRACT

Sentiment analysis (SA), especially multimodal emotion classification, attracts considerable interest in research and several applications. However, multimodal sentiment analysis (MSA) includes audio, text, and video data, faces numerous challenges related to the heterogeneity of the data and the underlying emotional contexts. Therefore, the proposed work develops a novel MSA framework called the dynamic cross-level leveraged multi-folded multi-model sentiment analysis-based classification framework. This article presents a framework for MSA based on text, audio, and visual modalities called the dynamic cross-level multi-folded multi-model sentiment analysis framework. The main purpose of developing this framework is to address the shortcomings of the current techniques. These include: (i) a multi-folded feature extractor capturing local and global features at different levels within each modality (words/phrases, sentences for text, frames/segs for audio, frames/times for videos); (ii) a dynamic cross-level attention-driven integration strategy to weight the features at various levels and modalities adaptively; (iii) an adaptive softmax function for classification tasks; (iv) a weighted and regularized three-stage loss function incorporating binary cross-entropy loss, confidence scores, and regularization terms; and (v) uncertainty-aware classification using confidence scores. The experimental results conducted using Python on BanglaMUSE and Multimodal Emotion Recognition dataset indicate that the proposed model performs better than others. The framework scores an accuracy of 0.985, F1-score of 0.982, Geometric Mean (G-Mean) score of 0.9843, and a Matthews Correlation Coefficient (MCC) of 0.9691. Loss ablation confirms that the proposed loss enhances the macro accuracy of our model by 3.67%. These results confirm that the dynamic cross-level fusion approach with multi-folded encoding is an effective solution to the problems of heterogeneity and noise in MSA.

1. INTRODUCTION

In the dynamic service sector, particularly in the hospitality sector, organizations must prioritize the user experience. analyzing user comments and suggestions and making the appropriate changes is an efficient way to accomplish this. Due to its varied evolution, sentiment analysis is a well-liked issue in natural language processing (NLP) that has drawn a lot of interest from scholars [1]. The development of multimodal emotion understanding and reasoning is essential to accomplishing this aim because a better comprehension of human emotions enables answers that are more user-acceptable and adaptable. Thus, in recent years, modelling and assessing emotional and psychological states have received a lot of attention [2, 3]. The use of multimodal data has grown throughout fields like social media and human-computer interaction in the age of rapid information technology development. Users convey their feelings through a variety of media, including text, photos, and sounds, offering rich modal data for in-depth emotional analysis [4]. Machine learning (ML) techniques can be used by social media platforms to automatically survey the collective sentiment tendencies on videos related to a particular issue. The results can then be transmitted to other pertinent organizations for decision-making [5]. However, using multi-modal data presents more obstacles as well as more opportunities. In the domains of psychology, artificial intelligence, and human-computer interaction, comprehending human emotions has become a critical focus. Nevertheless, it is still difficult to interpret these attitudes and emotions in speech, necessitating sophisticated ML methods [6, 7].

The development of systems that effectively capture and understand emotional expressions is crucial because emotions impact social interactions, improve communication, and influence decision-making. The ability to analyze emotions from a variety of inputs is crucial in the digital age, as communication frequently takes place through text, photos, and video. For example, joy or excitement may be the source of a positive sentiment, whereas wrath or fear may be the source of a negative sentiment [8]. Therefore, comprehending

both sentiment and emotion can improve the accuracy of both sentiment analysis and emotion detection tests as well as offer a deeper understanding of one's affective state. Conventional emotion analysis mostly uses unimodal data (text, for example), however, depending only on one modality frequently falls short of capturing the multifaceted character of emotions [9, 10]. For example, text may not directly convey emotion, but the accuracy and robustness of emotion identification can be significantly increased by using visual signals (such as facial expressions) and audio characteristics (such as speed and pitch). As a result, emotion analysis is changing from a unimodal approach to a combination of several modalities, such as text, visual, and aural information. Over the past ten years, researchers have paid close attention to sentiment analysis because of its many real-world uses [11]. Emotions, attitudes, and views are automatically extracted from web information. Many people actively use online review sites, discussion boards, news website comment sections, social networking sites, and personal blogs to share their opinions, which can include both favourable and unfavourable opinions about people, places, and events. These attitudes can be classified as sentiments [12].

Single textual data is not the only source of human cognition. Text, picture, and video data frequently appear at the same time in real-world scenes. Furthermore, in certain situations, like irony and sarcasm, it is challenging to determine the sentiment state only from the text. In order to complete a negative sentiment expression, irony and sarcasm sometimes combine neutral or positive textual material with auditory expression that contradicts the content. It is difficult to address the aforementioned problems essentially with just text data [13]. As a result, multimodal sentiment analysis (MSA) that integrates several modalities has garnered a lot of interest lately. There are several reasons why multimodal emotion identification methods are appealing. First of all, human emotion recognition takes place in a multi-model context in real life, where voice, body, and face are all evaluated holistically. It seems reasonable to train a computer to use the same method when trying to teach it to mimic aspects of human emotional intelligence [14]. MSA, which tries to comprehend and interpret human sentiments through many types of human expressions (e.g., language content, voice tone, and facial behaviour), is becoming a hot study subject as a crucial component of human-computer interaction. In a number of industries, including healthcare, social media analytics, and human-computer interface, MSA is essential. Multimodal analysis provides a more thorough and rigorous knowledge of human emotions than unimodal techniques [15]. Recent approaches to MSA mainly focus on representation learning and fusion algorithms across modalities, utilizing the developments in deep learning (DL) techniques. Numerous approaches have been developed for representation learning, such as feature decoupling methods that map features into shared and private areas [16].

Sentiment analysis can be used to identify top performers by assessing the language style used in emails, forecast virality from linguistic aspects of newspaper articles, or provide market insights from user-generated internet content. The field of sentiment analysis uses a variety of methodologies, from more sophisticated transfer learning methods to lexicon-based approaches and conventional ML models [17]. Multimodal approaches leverage the advantages provided by each modality to provide a more comprehensive understanding of the experience of emotional expression. Unlocking the potential of multimodal will present problems including capturing modality-aware and cross-modal contextual dependencies and effectively integrating disparate data streams [18]. The foundation for enhancing the complimentary qualities of information across several modalities is features within a modality. The fusion of information is the main emphasis of intermodal information exchange. To produce more precise and reliable results for sentiment analysis jobs, it fully utilizes the transmission and complementing features of information. These fusion techniques are separated into many categories based on the type of fusion, such as data fusion, feature fusion, decision fusion, etc [19]. Despite these developments, several current models still have limitations linked to their size or architecture that make it difficult for them to extract multimodal features. The approaches outlined above typically ignore the inter-correlation information of multi-modalities. As a result, the field of MSA has seen a substantial change with the introduction of Transformer-based pre-trained language models like BERT and XLNet [20].

The major contributions of this proposed work are:

- A new dynamic cross-level feature fusion mechanism is introduced that effectively integrates heterogeneous modalities (text, audio, and video) by modeling both intra-modal (within modality) and inter-modal (across modality) relationships using attention-based learning.
- The proposed work employs a multi-folded feature encoder that captures both local and global representations across modalities by utilizing phrase- and sentence-level BERT for text, frame- and segment-level Mel-spectrogram for audio, and extracting the spatial and temporal feature using a Multi-Branch Residual Connection. The Multi-Branch Residual Connection mainly convers multimodal data that enhances feature learning efficiency and contextual understanding.
- An attention-based weighted fusion mechanism combined with a channel attention module is developed to selectively emphasize the most informative features while suppressing redundant and noisy information, improving robustness.
- A novel weighted regularized three-fold loss function is proposed, integrating Binary Cross Entropy (BCE), confidence score–based regularization, and weight regularization to enhance training stability, reduce overfitting, and improve prediction confidence.

The paper is organized as follows: A review of the pertinent literature is presented in Section 2. The proposed strategy is covered in detail in Section 3, and the findings are discussed in Section 4. The paper is finally concluded in Section 5.

2. LITERATURE REVIEW

This section investigates a number of research articles in order to analyse the benefits and drawbacks of the current methods.

The Modality-Specific Adaptive Weight Fusion Network (AdaMoW), a multi-modal emotion analysis network, was proposed by Zhang et al. [21] in 2023 with the goal of resolving the issue of inconsistency in multi-modal sentimental analysis. In order to assist integration, this model learns a weight map by assigning feature weights to each variable according to its correlation with perceptual features.

AdaMoW outperformed previous baselines and demonstrated effective adaptive fusion for sentiment prediction when evaluated on CMU-MOSI and CMU-MOSEI datasets which shows the proposed model obtained 83.45% accuracy on CMU-MOSI and 85.23% accuracy on CMU-MOSEI.

In 2024, Silva et al. [22] employed ML models to automatically assess the sentiment of text-image combinations. The sentiment of the post and the difference between the feelings expressed by the text–picture pair are returned when the sentiments derived from the image and text are assessed separately and combined (or not) to form the overall sentiment. Four categories are used to classify image sentiment: "indoor" (IND), "man-made outdoors" (OMM), "non-man-made outdoors" (ONMM), and "indoor/outdoor with persons in the background" (IOwPB). These categories are then combined into an image sentiment classification model (ISC), which can be compared to a holistic image sentiment classifier (HISC) to demonstrate that the ISC outperforms the HISC. Likewise, Alfreihat et al. [23] created an Emoji Sentiment Lexicon (Emo-SL) specifically for Arabic-language tweets and show how integrating emoji-based characteristics with ML for sentiment classification improves performance. Using a dataset of 58K Arabic tweets that contained emojis, we built the Emo-SL by determining sentiment ratings for 222 commonly appearing emojis based on how they were distributed between positive and negative categories. To train classifiers on an Arabic tweet dataset, emoji weighting is combined with text-based feature extraction using lexicons. Support Vector Machines (SVM), Naive Bayes, Random Forests (RF), and K-Nearest Neighbours (KNN) are among the ML models that are assessed following ideal preprocessing and normalization. Moreover, Subbaiah et al. [24] have suggested a hybrid Arithmetic Optimization Algorithm-Hunger Games Search (AOA-HGS)-optimized Ensemble Multi-scale Residual Attention Network (EMRA-Net) technique to investigate modal correlations in text, audio, video, and social links for more effective MSA. Complementary and complete features are learned using the hybrid AOA-HGS method. To analyze the multimodal feelings, the EMRANet employs two segments: Ensemble Attention Convolutional Neural Network (EA-CNN) and Three-scale Residual Attention Convolutional Neural Network (TRA-CNN). By incorporating the Wavelet transform into TRA-CNN, the loss of spatial domain image texture characteristics can be minimized. Visual, audio, and textual data are combined using the EACNN feature-level fusion technique. To give more consideration, Liu et al. [25] provided two new models, IFFSA and BFSA, that use the huge language models BERT and GPT-2 for text modality feature extraction and ResNet and VGG for video modality feature extraction. This work addressed the complex problem of subtle emotion detection in multimodal situations by showcasing the synergistic potential of merging several modal analytical skills, making a unique contribution to the field.

Ren [26] proposed a MSA system in 2024 that combines ResNet-50 for visual feature extraction and BERT for textual representation. The Output Transformer Encoder (OTE) model outperformed unimodal baselines with an accuracy of 74.5% using Accuracy, Precision, Recall, and F1-score as assessment criteria (BERT: 70.5%, ResNet: 65.75%).

In 2025, Wang and Zhuang [27] analyzed emotions in text data using a variety of emotional analysis methods. Then, using a hierarchical information fusion technique, the emotional analysis results of many models were combined to increase accuracy. In the meantime, node information in graph data was processed using graph embedding algorithms like DeepWalk and Node2vec. Emotional analysis of nodes was done using the embedding model of graph nodes. Additionally, contextual node information was acquired to enhance emotional analysis performance based on models such Graph Convolutional Neural Networks (GCN) for message passing.

In 2025, Wang et.al. [28] introduced the Global-Local Feature Fusion with Co-Attention (GLFFCA) model for feature-level MSA (text and image). The model specifically addresses the challenge of aligning global semantic features with local relationships at a fine grained level. On Twitter-based multimodal sentiment data, GLFFCA produced excellent results with accuracy rates above 80–85% and F1-scores consistently higher than the existing techniques.

Similarly, Farhadipour et al. [29] suggested a system that uses pre-trained models to integrate four important modalities/channels: Wav2Vec2 for audio, RoBERTa for text, a CNN + Transformer architecture trained from scratch for video analysis, and a suggested FacialNet for facial expressions. Emotion and sentiment labels are predicted using a multimodal vector created by concatenating feature embeddings from each modality. Moreover, Zhang et al. [30] provided a Multimodal Semantic Fusion Network (MSFN) for MSA in order to thoroughly investigate the semantic relationship between text and image. Specifically, a gated attention technique was used to align the sentiment-related picture region and text word characteristics. The interactions between these features are then modelled using graph convolutional networks in order to get explicit sentiment semantics. The suggested gated attention mechanism uses a gating mechanism to rectify possible feature misalignment during cross-modal alignment. Likewise, Sun et al. [31] presented UniEmotion, a unified framework that simultaneously tackles descriptive emotion comprehension, open-vocabulary fine-grained emotion recognition, and traditional categorical emotion recognition. This method maximizes the utility of large models while minimizing downstream constraints by utilizing an iterative consensus-based training process where pseudo-labels and model parameters co-evolve. The system combines a pseudo-labeling module that uses consistency regularization and class-wise adaptive mapping with a selector module that uses prediction variance analysis to identify high-quality samples.

Malik et al. [32] culminated in the development of the extensive collection of Urdu video reviews known as the Urdu Sentiments Dataset (USD). In this work, we employ a two-phase method for video classification that includes early fusion and ensembling. Following fusion, we assemble two models for each modality text, audio, and frames. For audio, we employ RF Classifier and Long Short-Term Memory (LSTM) networks. Logistic regression (LR) and the Bidirectional Encoder Representations from Transformers (BERT) model are used for text-based analysis. RF and CNN for frames were utilized. Next, apply model assembling in all three modalities. In addition, Hossain et al. [33] presented Dimension Wise Gated Cross-Attention (DGCA). Compared to previous approaches, this new fusion process improves the precision of the language-image interaction. This technique iteratively improves text and image characteristics using a bidirectional cross attention module. A dimension-wise gating method was employed where each latent dimension learns on its own to utilize softmax-normalized modality gates to weigh

contributions from text or picture signals. To emphasize crucial clues from one modality while reducing less helpful traits from another, the method employs selective per-dimension fusion. In addition, Cherukuri [34] proposed approach starts with modality-specific preprocessing: Relaxed instance Frequency-wise Normalization (RFN) for audio to reduce noise distortion, NLP for text to address linguistic variations, and Iterative Self-Guided Image Filter (ISGIF) for videos to improve image quality and reduce artifacts. The Inception Transformer is used to capture the textual contexts; the Differentiable Adaptive Short-Time Fourier transform (DA-STFT) is used to extract the spectral and temporal features of the audio; and class attention mechanisms are used to highlight significant aspects of the videos. After that, a Multi-Branch Fusion Attention Network combines these characteristics to harmonize all the many modalities into one. An Epistemic Neural Network (ENN) performs the final sanity check by addressing the uncertainty in the final classification, and the Fire Hawk algorithm is employed to improve the framework's emotion identification capabilities. Wang, and Nourmohammadi, [35] carried out vectorization using two distinct word embedding models, GloVe and Word2Vec. Because there were two polarities in this study positive and negative a bidirectional gated recurrent unit was used. The Enhanced Human Evolutionary Optimization (EHEO) algorithm was then used to optimize it, which improved the hyperparameters.

In 2026, Almadhor et al. [36] have provided a number of deep and hybrid learning models to solve the problem of emotion recognition in speech across cultural boundaries and perform a thorough evaluation of cross-lingual SER. Four types of architectures Using different combinations of Urdu, English, German, and Italian voice data corpora, Artificial Neural Network (ANN), Multi-Layer Perceptron + LSTM (MLP + LSTM), RF + Deep Neural Network (RF + DNN), and our proprietary transformer are used for emotion recognition. Their models are more adept at handling different languages than typical ML classifiers. Duong et al. [37] have introduced SentiFuse, a versatile and model-neutral framework that combines various fusion techniques with a standards layer to merge diverse sentiment models. Our method allows for the systematic combining of several models by supporting decision-level fusion, feature-level fusion, and adaptive fusion. Three large-scale social media datasets CrowdFlower, GoEmotions, and Sentiment140 are used in our investigations. These tests demonstrate that SentiFuse routinely performs better than naive ensembles and individual models. Likewise, Selvi, and SVN, [38] suggested a hybrid DL model for text-based sentiment analysis that incorporates a fusion attention mechanism. In order to maximize the emphasis on textual words with a strong emotional inclination, the suggested hybrid DL model combines Bi-directional Gated Recurrent Networks (Bi-GRU) to pull the background connection with the text and CNN to extract local information. The experimental is conducted using social media tweet data from Facebook, WhatsApp, and Twitter. The study investigated semantic sentiment analysis based on the attention mechanism to examine classification prediction for assessing the public's positive and negative opinions.

Although there have been increasing efforts in using multimodal data (text, audio, and video) for emotion recognition, current methods face a number of challenges in effectively modeling the complex and diverse nature of human emotions. This can be attributed to insufficient feature extraction that does not exploit hierarchies (phrase/sentence, frame/segment), ineffective interaction and integration techniques ignoring correlation between modes, and being highly sensitive to noise. In summary, existing models frequently face problems associated with inaccuracy, lack of robustness, and inability to analyze subtle cues like sarcasm and irony, in which case there could be contradicting modalities. The limitations mentioned above indicate the urgent need for a flexible and holistic approach capable of hierarchical feature encoding, intelligent cross-modal fusion, and providing uncertainty-based classifications.

3. PROPOSED METHODOLOGY

The proposed dynamic cross-level Multi-Folded MSA framework, which successfully integrates heterogeneous data from textual, audio, and visual modalities, is presented in this section. Variations in data representation, temporal misalignment, noise sensitivity, and the intricate interdependencies between modalities make MSA extremely difficult. Current methods frequently rely on basic fusion procedures or shallow feature representations, which restricts their capacity to capture contextual and fine-grained information across modalities. The suggested system uses a multi-branch design that handles each modality separately while permitting structured interaction later on in order to overcome these drawbacks. The methodology processes modality-specific preprocessing to guarantee temporal alignment and noise reduction. Next, feature encoding is done utilising sophisticated representation techniques like Mel Spectrograms for audio signals and BERT for textual data. The spatial and temporal features are then learned using a multi-branch feature extraction module, resulting in hierarchical representations in the form of local and global features. A dynamic cross-level multi-folded feature fusion technique that integrates modality-specific features via cross-level alignment, residual learning, and adaptive attention is presented in order to successfully merge these representations. Because of this architecture, modalities can interact robustly while maintaining their own qualities. Lastly, a channel-aware decision mechanism is applied to the fused representation to classify sentiment and emotions. The overall flow of the proposed work is clearly highlighted in Figure 1.

3.1 Multimodal data acquisition and pre-processing

The purpose of preprocessing in this study is to guarantee modality-specific preprocessing for text, audio, and visual streams. The objective is to convert unstructured multimodal inputs into consistent and structured representations that may be used for further feature encoding and learning. Each modality is treated separately using customised methods to minimise noise and redundancy while maintaining its inherent qualities.

3.1.1 Text pre-processing

Text modality goes through cleaning and normalization processes. As raw comments from users may have punctuation marks, URLs, emojis, stopwords, and other variations in grammar, pre-processing becomes necessary to enhance the quality of features. Consider a raw sentence of textual form as $T = \{w_1, w_2, w_3, \ldots, w_n\}$ where w_i denotes the i-th word in the sentence.

Stopword removal: Stop-words such as "is", "the", "and", "in", etc., are often used in sentences, but they do not really contribute to the meaning of the sentence. By eliminating stop words, the system may focus more on the words that help identify abusive language while lowering the dimensionality of the features. By creating a distinct word set, this process should improve the accuracy of the similarity. To improve the quality of the dataset, stop words must be eliminated.

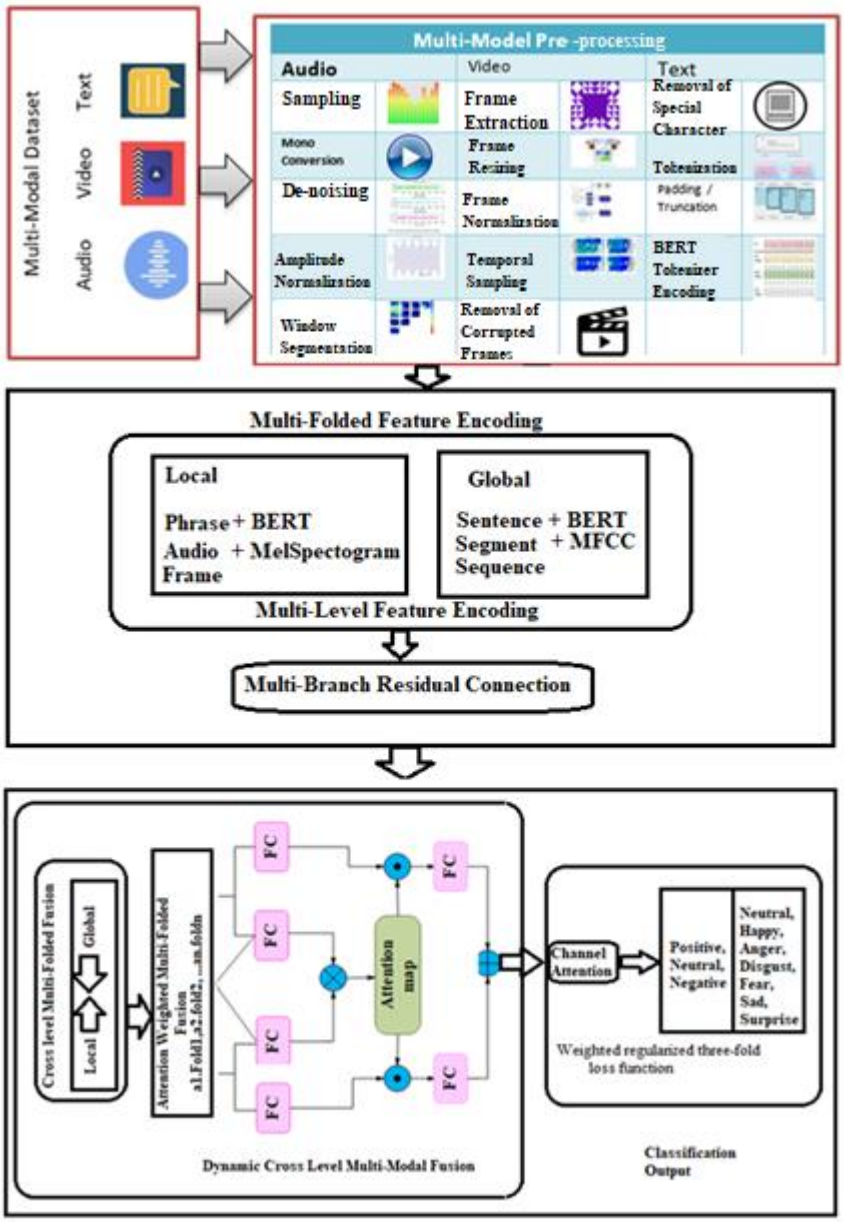


Figure 1. The proposed architecture of dynamic cross-level leveraged multi-folded multi-model sentimental analysis

Lemmatization: Lemmatization and stemming are similar text normalisation methods, but lemmatization maintains meaning more effectively. A lemma preserves semantics while reducing a term to its most fundamental form. In contrast, stemming frequently reduces words to non-intuitive bases. For instance, the words "bullied" and "bullying" are lemmatized as "bully" but stemmed as "bulli." Lemmatization is frequently chosen for sentiment analysis jobs because it preserves semantic clarity.

Stemmation: Stemming reduces words like "insulting," "insulted," and "insults becoming insults" to their most basic form. This approach guarantees consistency in features and enhances semantic similarity capture by combining word variants. Standardising tokens with similar meanings but different inflections, it increases the efficiency of the model. However, stemming can also lead to abbreviated or inaccurate roots, like Stem's reduced to Stem. Despite this disadvantage, it is useful for ensuring consistent representation in text preparation tasks.

Tokenization: Tokenization separates text into meaningful units, such as words or symbols, known as tokens. It makes it possible for the model to evaluate every word in a sentence and extract pertinent data. Tokenization addresses problems such as missing punctuation, acronyms, and abbreviations by standardising text. The model can identify trends and important terms because each token represents a unique word. Tokenization, the cornerstone of text preprocessing, guarantees accurate and contextually appropriate feature extraction.

3.1.2 Audio pre-processing

The audio signal has silence, noise, and fluctuations in

amplitude. Thus, it is necessary to apply pre-processing before extracting features. The audio signal can be denoted as $x = \{x(1), x(2) \ldots, x(N)\}$.

Denoising: It is the process of eliminating background noise by using signal processing techniques (Spectral subtraction and Weiner filtering). This is necessary to improve feature extraction accuracy and make the voice signal more understandable. After denoising, the continuous signal is framed into short parts. The framing process allows for the examination of short-term spectral properties by splitting the signal into short time intervals (between 20 and 40 milliseconds). Because speech signals are naturally non-stationary, framing allows the system to assume quasi-stationarity.

Framing: The speech signals were suppressed within the N sample frame, and the adjoining samples were punctured by M ($M > N$). The initial N samples contained the initial frame. Sample M is made up of a second frame. The process continues through N- M samples of overlap after adding the first frame, and so on, until each speech is enclosed within one or more frames.

Windowing: The discontinuity in the signal at the start and end of each frame is lessened as a result. The main goal of the windowing process is to lessen the spectral distortion of the signal. The signal is reset to zero at the beginning and the end of each frame.

$$x_w(n) = x(n).w(n), \quad 0 \le x \le N-1 \qquad (1)$$

where, $x(n)$ and $w(n)$ are respectively represents the audio sample at time n and windowing function, and $x_w(n)$ is windowed signal. After the window's findings are signalled, N can be expressed as the sample number within each frame in the equation indicated above.

3.1.3 Video pre-processing

The video modality is transformed to frames. As video has emotional cues in space and time domains, frame extraction and temporal segmentation are done. The video sequence can be denoted by $V = \{v_1, v_2, v_3, \ldots v_T\}$ where v_t denotes the t-th video frame. With respect to video modality, the preprocessing phase involves transforming the input video into a series of representative frames. This process involves frame extraction, whereby videos are broken down into frames using the selected frame rate. This converts the temporal data contained within the video into static images that can be used in further processing. After this stage is completed, frame resizing and normalization processes are conducted. This is essential since it makes sure that the frames are compatible with CNN techniques.

Temporal sampling plays an important role in the video preprocessing process because not all frames are used for modelling; only a small number of frames are selected for modelling purposes. Considering that consecutive frames might have very similar information, it is better to reduce the amount of data to improve processing efficiency. The preprocessing process also involves the removal of bad frames; these might occur due to blurring, occlusions, or transmission errors. Figure 2 shows the essential preprocessing steps of the proposed work.

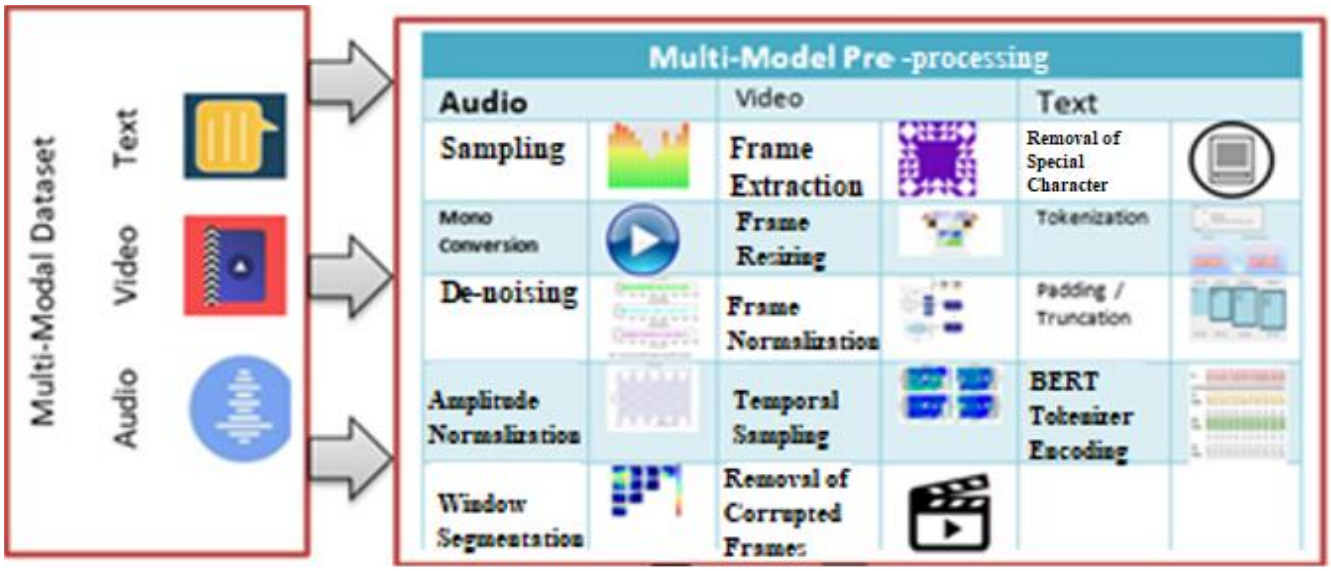


Figure 2. Graphical illustration of the preprocessing steps

3.2 Multi-folded feature encoding

The proposed method employs a multi-layered feature encoding strategy, which represents a major innovation compared to traditional single-layer feature extraction approaches. Unlike conventional methods that operate at a single scale, the Multi-Folded Feature Encoder (Local-Level and Global-Level Encoder) decomposes each modality hierarchically, ensuring richer representation. The Local-Level Encoder uses a multi-branch ResNet to handle the fine-grained information it collects from visual frames, audio frames (Mel spectrograms), and textual phrases (BERT). Meanwhile, the Global-Level Encoder uses a multi-branch ResNet to parse phrases, audio segments (MFCC), and video sequences in order to extract contextual semantics and long-distance relationships. To guarantee reliable representation, the model makes use of many DL architectures, including multi-branch ResNet18, LSTM, and BERT transformer. Figure 3 highlights the graphical flow of the Multi-Folded Feature Encoding.

3.2.1 Local-level encoding

Local-level encoding involves the extraction of finer details from small pieces of input data. Such data pieces can be

phrases for texts, frames for audio data, and frames for videos. The intention behind local-level encoding is to identify the micro-patterns such as phonemes, facial gestures, and semantics of contextual phrases.

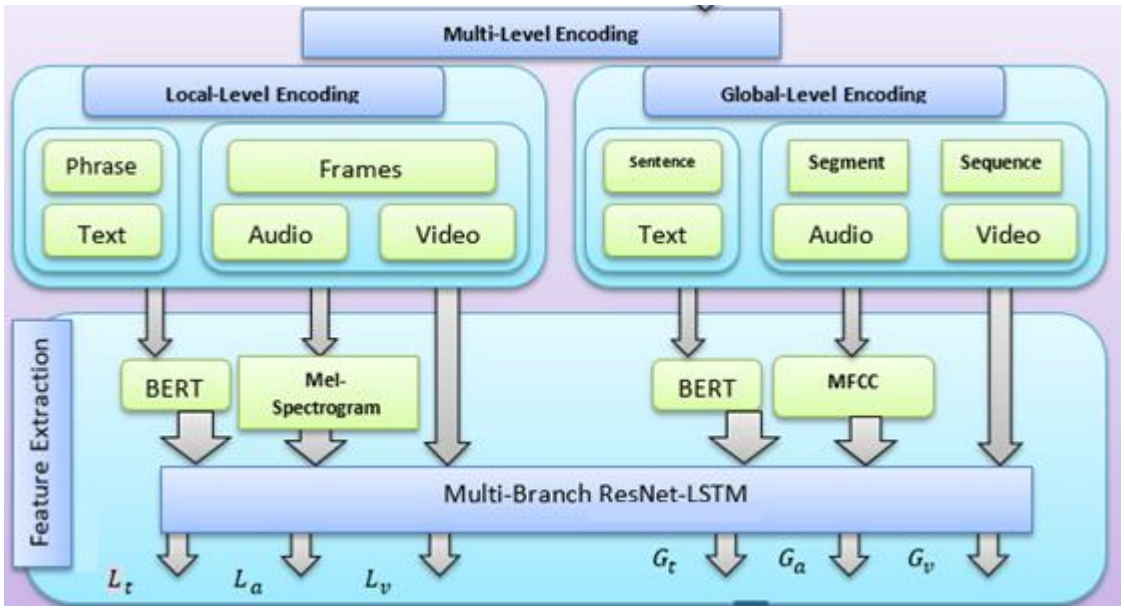


Figure 3. The proposed multi-folded feature encoding

Phrase-encoding using BERT

A multi-layer bidirectional Transformer that Google proposed acts as the basis for the language model known as BERT. Word vectors of the sequence can be produced by the pre-trained BERT and employed as a high-quality input for tasks further down the line. To specifically represent the entire input, we change the sentence to "[CLS] + sentence + [SEP]". The segment token and the beginning token are denoted by the special tokens [SEP] and [CLS], respectively. After processing, the sequence is fed into BERT to be context-coded. Token embedding, segment embedding, and position embedding are added together to turn each token into a vector. Subsequently, the vector sequence is fed into a Transformer layer stack to extract the encoded contextual data. The context representation is the hidden-layer output from the final Transformer block. The hidden representation in the following article does not include any special tokens; rather, it refers to the word's representation in each sentence for the sake of clarity. In the case of text modality, there are two parallel encoders that have been used, namely the phrase-level encoder and sentence-level encoder. In the case of the phrase-level encoder, a model based on transformers such as BERT is used for extracting the local semantics of fine-grained levels where the contextual dependency is found between words and phrases. The textual branch uses two independent BERT encoders: Phrase and sentence level BERT encoder. The encoder for phrases can obtain the local contextual information from short phrases and emotionally charged words. Phrases like "very happy," "not good," or "extremely disappointed" contain highly localized emotional information that might not be well captured when using the entire sentence. The BERT model at the phrase level detects such dependencies between neighboring words.

$$X_{phr} = \{[CLS], p_{1,\ldots,}p_q, [SEP]\}. \quad (2)$$

$$H_p = BERT\ (X_p) \quad (3)$$

The phrase-level feature vector is obtained from the final hidden state of the [CLS] token:

$$f_p = H_p^{CLS} \quad (4)$$

Phrase-based encoding is highly beneficial for recognizing negation words, strong words, and modifying words, which have a considerable impact on sentiment polarity. For example, the phrase “not bad” carries a positive sentiment even though it contains a negative word. In addition, this type of encoding can be considered an effective noise-filtering mechanism because it extracts emotion-related phrases from other words that have no relevance to the emotions expressed in the text.

Audio encoding

In order to extract fine-grained acoustic information, the audio modality's local-level encoding procedure is accomplished using a frame-based analysis of the raw audio signal. This method divides the continuous audio waveform into brief frames, which are then converted into Mel-spectrogram representations. By aligning the signal's frequency components with human auditory perception, this transformation improves the signal's suitability for learning perceptually significant information.

Mel-spectrogram transformation

The Mel-spectrum is a depiction of the speech signal based on the Mel scale that highlights perceptually significant frequency regions. An approximate representation of the human hearing system's non-linear frequency resolution, the Mel scale is a perceptual scale of pitches. A spectrogram is a graphic illustration of the frequency spectrum of a signal over time. The following formula, where m stands for Mels and f for Hertz, relates the Mel scale to Hertz and offers a linear scale for the human auditory system:

$$m = 2595log_{10}\ (1 + \frac{f}{700}) \quad (5)$$

where, f represents frequency in Hertz and m denotes the

corresponding value in the Mel scale. This nonlinear transformation emphasizes lower-frequency components, which are more perceptually significant.

The Mel-spectrogram is then created by applying a collection of Mel filter banks to the spectrogram:

$$a_1 = melSpec\ (x(t)) \epsilon R^{F \times T} \quad (6)$$

where, T stands for the number of temporal frames and F for the number of Mel frequency bins. Each audio sample in this work is represented as a 128 × 128 Mel-spectrogram, which corresponds to 128 time steps and 128 filter banks.

3.2.2 Global-level encoding

Local encoding deals with capturing emotions on an instant basis, whereas global encoding considers long-distance dependencies within the text data structure. Global encoding looks at the broader context for understanding sentiments over longer periods or in larger textual blocks.

Sentence encoding using BERT

Sentence-level encoding involves understanding the holistic semantic sense of the text. As opposed to phrase-level encoding that emphasizes local dependencies, sentence-level encoding involves processing the whole sentence using a Transformer-based model like BERT. The global sentiment context is extracted from sentence-level BERT:

$$X_{sen} = \{[CLS], s_{1,\ldots\ldots,} s_q, [SEP]\}. \quad (7)$$

$$H_{sen} = BERT\ (X_{sen}) \quad (8)$$

The phrase-level feature vector is obtained from the final hidden state of the [CLS] token:

$$f_{sen} = H_{sen}^{CLS} \quad (9)$$

The sentence encoder produces a global embedding that encodes the overall sentiment polarity, along with the relationship between various phrases, and even more advanced concepts like irony and sarcasm. In the case of the following sentence: “I thought that movie was going to be great, but I was disappointed with it,” the overall sentiment is negative, although there are some positive words used. This global representation augments the phrase-level representations, making sure that both local and contextual semantics are included in the final feature space.

Segment- encoding (audio)

Segment-level audio encoding focuses on analyzing longer temporal segments of speech, rather than individual frames, in order to capture stable and context-rich acoustic patterns. Each segment consists of multiple consecutive frames, enabling the extraction of higher-level characteristics such as speech rhythm, tempo, intonation patterns, and emotional prosody. This form of encoding is essential for modeling sustained emotional states over time, where patterns such as prolonged low energy and slower speech may indicate sadness, while consistently high energy levels may reflect anger or excitement. To achieve this, the proposed framework employs Mel Frequency Cepstral Coefficients (MFCCs), which provide a compact and perceptually meaningful representation of audio signals. The MFCC extraction process involves several key steps: pre-emphasis to amplify high-frequency components, framing and application of a Hamming window to reduce edge discontinuities, computation of the power spectrum via Fourier transform, and filtering through Mel-scale filter banks to approximate human auditory perception. The logarithm of the filter bank energies is then transformed using the Discrete Cosine Transform (DCT) to decorrelate features and obtain the most significant cepstral coefficients:

$$X(k) = \sum_{n=0}^{N-1} x_n * cos\left(\frac{2\pi jnk}{N}\right)\ k = 1,2,3, \ldots n-1 \quad (10)$$

where, N is the signal's length and x_n is a discrete signal. MFCC features are extracted.

The segment-level encoding helps provide a broader perspective on the audio input, in addition to the detailed features obtained from the local-level encoding process.

3.3 Cross-level multi-level fusion

The proposed architecture presents a Cross-Level Fusion mechanism that simultaneously utilises local and global representations across audio, text, and visual modalities in order to efficiently integrate heterogeneous multimodal input. In particular, the audio modality captures both long-term prosodic patterns and fine-grained spectral fluctuations by combining MFCC-based segment-level representations with Mel-spectrogram-based frame-level features. Similar to this, the textual modality combines sentence-level contextual representations with phrase-level embeddings obtained from BERT, allowing the model to capture both global language relationships and local semantic clues.

The model can learn both instantaneous visual signals and dynamic expression shifts over time by fusing frame-level spatial features with sequence-level temporal features in the visual domain. Complementary information transmission between local and global representations is made possible by projecting these multi-level features into a shared latent space and fusing them via a cross-level interaction mechanism. The robustness and accuracy of multimodal sentiment and emotion identification are greatly increased by this architecture, which guarantees that the model captures both macro-level contextual semantics and micro-level discriminative patterns.

3.4 Multi-branch multimodal feature extraction

The proposed framework introduces a multi-branch multimodal feature extraction architecture based on Multi-Branch ResNet18, designed to effectively learn discriminative representations from heterogeneous data sources, including text, audio, and visual modalities. In this design, each modality is processed through a dedicated branch, enabling modality-specific feature learning while maintaining a unified architectural structure. The textual modality is first encoded using BERT to obtain contextual embeddings, which are subsequently refined through sequential modeling to capture semantic dependencies. In contrast, the audio modality is transformed into log Mel-spectrogram representations, allowing the ResNet18 backbone to extract rich spectral–level features such as pitch, energy, and harmonic patterns. Similarly, the visual modality, represented as frame sequences, is processed using ResNet18 to capture spatial features such as facial expressions and appearance cues. The multi-branch configuration further enables multi-scale feature extraction, where different branches focus on capturing fine-grained (frame/phrase-level) and high-level (segment/sequence-level) representations. By leveraging residual learning and parallel processing, the proposed

architecture effectively preserves modality-specific characteristics while generating robust and hierarchical feature representations, which serve as critical inputs for the subsequent cross-level fusion module.

3.5 Multi-Branch Residual Connection architecture

The source of input data for the proposed module comes from the frame, Phrase and audio Frame encoding step, during which raw video is broken down into individual frames is shown in Figure 4. These frames undergo pre-processing (resizing, normalization, and noise reduction) and are organized into a temporal sequence, which is then fed into the multi-branch residual structure. The first layer involves the application of a 1 × 7 filter that is essential in handling horizontal spatial correlations. This is succeeded by the max-pooling stage (3 × 3), which helps in reducing the dimensionality without losing the significant features. The purpose of this process is to ensure that the incoming input into the next stages is effective for computation.

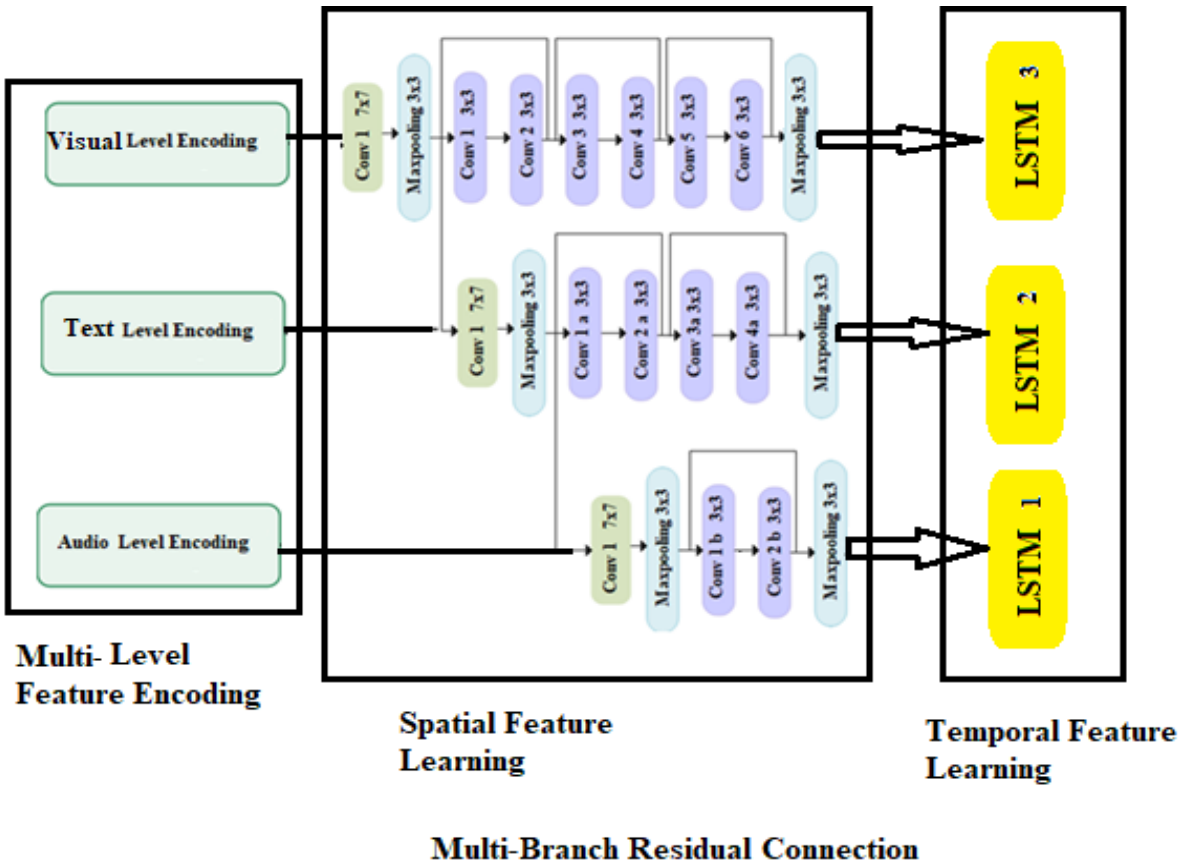


Figure 4. The proposed Multi-Branch Residual Connection architecture

The first major branch that follows includes several convolutional layers with kernel size of 3 × 3 that have been placed one after another. This is because it will be the role of this branch to generate hierarchical deep features from the encoded image. The repetition of convolutions will allow the system to detect complex patterns within the image, which may be in the form of edges, texture, or even higher semantics. With multiple layers of convolutional operations being applied within the branch, from Conv 1 to Conv 6, the system will become very capable of processing complex data like images or videos. This second branch is slightly deeper than the first one and is concerned with intermediate-level feature extraction. This part of the architecture uses several convolutional layers (Conv 1 – Conv 4) that have the same size of kernels as in the first branch. However, their depth is lower. This layer is meant to detect intermediate patterns that cannot be detected by shallow or deep layers. This part works in tandem with the previous branch providing additional information for feature extraction. The third branch is the shallowest out of the three branches, and it concentrates on extracting low-level features such as edges, corners, and basic texture features. The third branch has few convolutional layers (Conv1b and Conv2b), and its main concern is the early extraction of features. This is a very important component of the network architecture since it ensures that the finer details are captured, which could easily get lost at a deeper level.

The major innovation of the proposed approach is the residual learning embedded into the multi-branch network structure. In each branch, there are skip connections that let the input signal for a certain layer jump over several following layers, thus preserving important information. Subsequent to independent processing in branches, outputs are combined either by means of concatenation or weighted sum. The idea behind such combination is that information on various scales and depths can be merged into a single representation by means of the introduced combination operation. As a result, no single feature scale will prevail in the multi-branch model. The proposed Multi-Branch Residual Connection is well incorporated in the multi-folded feature encoding structure, acting as the video modality encoder. This way, it contributes to the system's capacity to encode rich video information. The spatio-temporal information extracted from videos through the Multi-Branch Residual Connection and then LSTM temporal modeling is passed up to upper layers for further processing,

including cross-level fusion and attention mechanisms. These aspects facilitate efficient communication and coordination between the modalities such that all essential data from each modality is effectively highlighted. Using hierarchical integration, the design allows for the incorporation of visual clues obtained from the Multi-Branch Residual Connection as well as the LSTM, audio signals gained from spectral CNN models, and text semantics obtained using transformer models. In this way, this proposed model ensures an integrated understanding of the multimodal sentiments.

3.5.1 Temporal feature learning

After extracting the local features for each modality, they enter the first temporal modeling layer called LSTM-1. The role of this layer is to identify the short-term dependencies present in each modality. For example, it may be able to model facial movements from frame to frame, speech characteristics from one small duration to another, and sentence structure. LSTM-1 uses gate units to remember important information and discard unnecessary information, thereby capturing the dynamic changes of local features. An input series denoted as $i = [i_1, i_{2,\ldots,} i_T]$, where T is the length of the time series, is sent to the LSTM cell. As this series emerges, the network generates a hidden sequence called $h = [h_1, h_{2,\ldots,} h_T]$. Every network activation unit iteratively handles the sequential data over time, transforming the input series into the hidden series in line with particular mathematical Eqs. (11)-(16) as listed below:

$$g_t = \sigma(W_{gi} i_t + W_{gh} h_{t-1} + b_g) \tag{11}$$

$$g_m = \sigma(W_{mi} m_t + W_{mh} h_{t-1} + b_m) \tag{12}$$

$$\widetilde{j_t} = \tanh\,(W_{ji} i_t + W_{jh} h_{t-1} + b_j) \tag{13}$$

$$n_t = \sigma(W_{ni} i_t + W_{nh} h_{t-1} + b_n) \tag{14}$$

$$j_t = (g_t \odot j_{t-1}) + (m_t \odot \widetilde{j_t}) \tag{15}$$

$$h_t = n_t \odot \tanh\,(j_t) \tag{16}$$

where, h_t represents the hidden state's output at the current time step. The forget, input, and output gates are represented by the letters $g, m,$ and n, correspondingly. The cell activation vector is represented by j, and the candidate cell vector is shown as $\widetilde{j_t}$, W stands for the weight matrices, and b for the bias vector. The symbol $\odot$ indicates an element-wise product. The sigmoid and hyperbolic tangent activation processes are represented by the symbols σ and tanh, respectively. This makes the static local feature data become dynamic through this stage. In this way, the changes of emotions in micro-time periods can be captured. The residual branches' outputs are split among three parallel LSTMs, each of them are processed temporal relationships at a different scale and modality. In particular, LSTM-1 captures short-term and fine-grained acoustic cues by processing shallow-level features that are mainly sourced from the audio branch. LSTM-2 learns long-term contextual and semantic dependencies from frame sequences by modelling deep-level information from the visual branch. In order to capture mid-range temporal dependencies, LSTM-3 bridges phrase-level and sentence-level representations by concentrating on intermediate-level features from the textual branch. These three LSTMs operate in parallel rather than sequentially, and their outputs are delivered simultaneously to the Dynamic Multimodal Cross-Level Feature Fusion module. At shallow, mid-level, and deep scales, this architecture guarantees the independent contributions of audio, text, and visual modalities, whereas the fusion stage This hierarchical integration ensures that the model does not miss out on any significant details while gaining knowledge from the larger context, leading to a balanced representation of features.

3.6 Dynamic multimodal cross-level feature fusion

The methodology starts with the extraction of multimodal features, which are acquired through previous encoding processes. The multimodal features consist of a fusion of both local and global information of various modalities, such as text features, audio features, and video features. In contrast to the conventional methods of considering these features individually, the proposed method assumes that the features are interrelated representations and needs to be optimized simultaneously. Each modality offers its own way of looking at the problem of emotion recognition, with their combination forming a base for building an overall representation of emotions. The resulting features go through several layers of fully connected (FC) neurons, thus undergoing normalization and transformation into a shared embedding space, facilitating the intermodal interaction. Figure 5 highlights the flow of the Dynamic multi-model Cross-level Feature Fusion.

After the first transformation, the characteristics undergo cross-level interaction that plays an essential role in the proposed approach. The cross-level interaction involves processing features using several FC layers at one time to analyze various subsets/levels of features. This transformation enables the model to capture not only intra-modal interactions but also inter-modal interactions. The output from all the FC layers are then aggregated using multiplicative and additive techniques, as shown in figure, allowing for the creation of complicated feature interaction processes. It is designed to ensure that features at various levels of abstraction, such as textual features at the phrase level and visual features at the frame level, can impact each other.

3.6.1 Multimodal feature transformation

The multimodal features obtained initially are processed using several FC layers, as illustrated in the diagram below. FC layers act as feature transformation blocks that transform different types of features to the same latent feature space. Let the multimodal features be denoted by:

$$F = \{F_t, F_a, F_v\} \tag{17}$$

where, F_t , F_a and F_v denotes textual, audio, and visual features, respectively. The above features are transformed via many FC layers in order to be projected into a common space:

$$H_i = \sigma(W_i F_i + b_i), i \in t, a, v \tag{18}$$

W_i and b_i denote the weights and biases, respectively, while $\sigma(\cdot)$ refers to an activation function like the ReLU function. The use of this transformation makes it possible for the features to be compared and combined effectively. In addition, several FC layers are used to obtain multi-level features.

3.6.2 Attention map generation for cross-level interaction

The main innovation of this proposed module is an attention

map module that allows for dynamic evaluation of the significance of the characteristics across different modalities and layers. These features are aggregated through element-wise multiplication and addition operations, allowing for interaction modeling:

$$Attn = softmax(H_t \odot H_a + H_v) \quad (19)$$

where, $\odot$ stands for element-wise product and $Attn$ stands for attention. Attention mechanism plays the role of a weight function, emphasizing relevant features and de-emphasizing irrelevant ones. The attention mechanism can learn cross-level relationships, that is to say features from one level of abstraction can affect features from another level. It becomes especially relevant when talking about MSA, since emotions may be differently distributed across modalities.

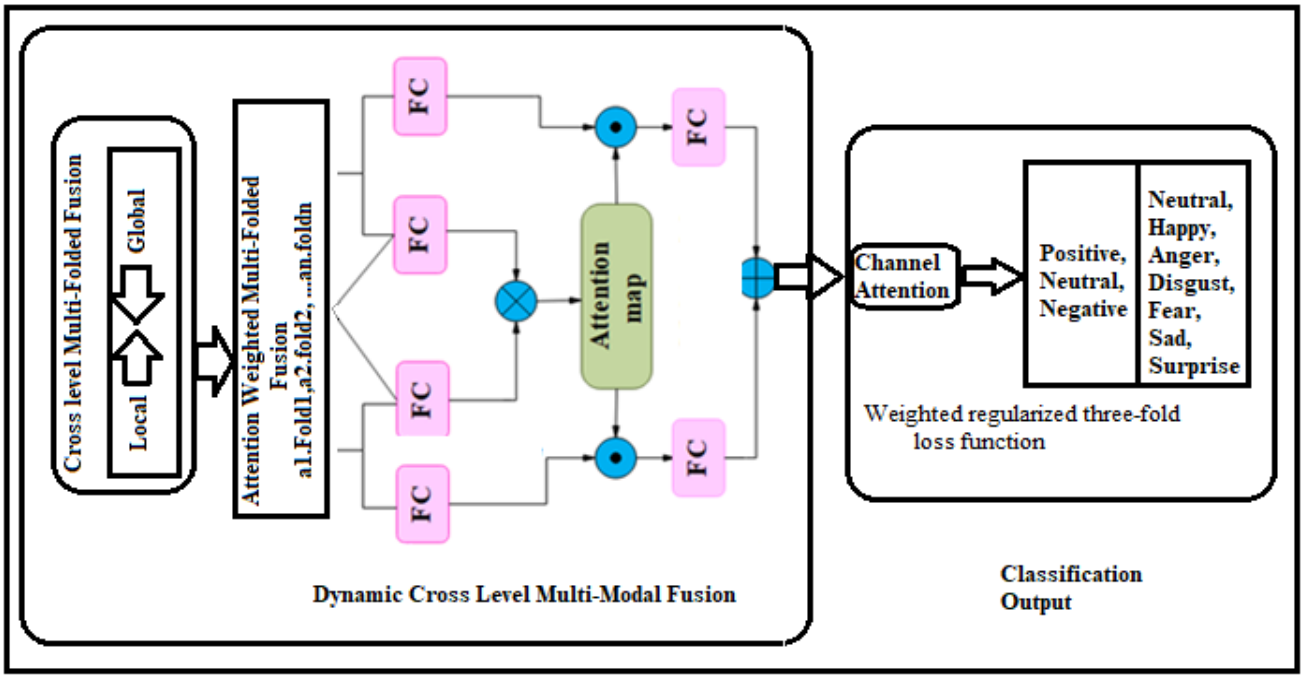


Figure 5. The proposed dynamic multimodal cross-level feature fusion architecture

3.6.3 Dynamic cross-level feature fusion mechanism

After calculating attention, the proposed model conducts dynamic feature fusion using weighted features from all modalities. After calculating attention, the proposed model conducts dynamic feature fusion using weighted features from all modalities. The fused feature vector can be formulated as:

$$F_{fused} = \sum_{i \in \{t,a,v\}} A_i . H_i \quad (20)$$

where, A_i is the attention weight for modality i. Such an adaptive approach allows for the weight assigned to each modality to be dependent on the context. For example, when there is ambiguity in the text information, more weight can be attributed to audio or visual information. This technique, unlike traditional approaches to fusion, allows for cross-modal interaction that is bidirectional, thereby permitting higher-order features, such as semantics, to affect lower order features, and the other way around. This kind of cross level interaction greatly enhances the detection of subtleties in emotion, like irony, sarcasm, and contradictions.

3.6.4 Channel attention mechanism for feature refinement

Furthermore, to aid the process of fusion even more, channel attention has been introduced into the architecture. This is achieved by focusing on the significant feature channels while downplaying less significant ones. Through this approach, attention is paid only on the relevant aspects that can contribute meaningfully for classification purposes. The inclusion of channel attention not only helps in improving the representation of features but also helps in avoiding the issues related to redundant data. In order to increase the effectiveness of the generated features, a channel attention module is included in the proposed framework that works on the multimodal feature vector. This channel attention method concentrates on selecting the most important feature channels from the multimodal feature vector. This procedure starts with the application of FC Layer 1 to the fused features and then applying the ReLU activation function to generate non-linear features. It is further fed into a second FC 2, which is followed by the sigmoid activation function. The output from this is channel-wise attention weights ranging from zero to one. These channel-wise attention weights are used to scale the feature channels according to their significance. Such selection enables the retention of important features while suppressing insignificant ones, thus enhancing feature discrimination. By highlighting informative channels, the channel attention method improves feature selection, as seen in Figure 6.

This classifier makes use of an adaptive softmax activation function in its classification stage, which is yet another innovation of the proposed architecture. While traditional softmax activation functions assume all classes to be equal, the adaptive softmax takes advantage of the varying feature importance in each class to enhance the efficiency of the classifier. The Adaptive Softmax can be mathematically formulated as a context-sensitive weighting mechanism for the

traditional softmax model, wherein the weight of each feature affects the probability distribution, based on its significance.

$$P(y_i) = \frac{\exp(\alpha_i . z_i)}{\sum_{j=1}^{C} \exp(\alpha_j . z_j)} \quad (21)$$

This concept is adapted to incorporate a learned importance weight α_i (based on the attention and fusion layers). The inclusion of the uncertainty classification algorithm forms a significant part of the methodology employed. The model is capable of determining the confidence level of its predicted classes. These levels are used during the training process and ensure that the model performs optimally when predicting sentiments despite the presence of ambiguity in the input data. Figure 6 shows the architectural flow of the channel attention layer.

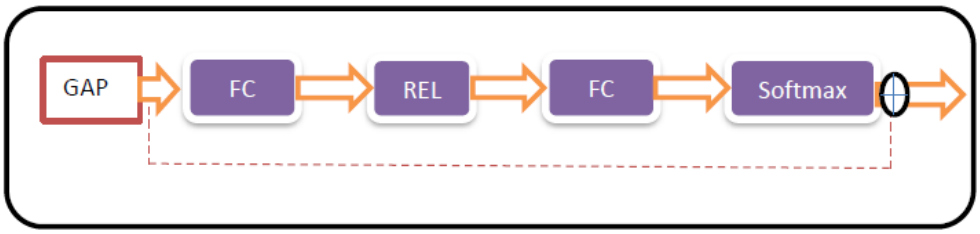


Figure 6. Channel attention layer

3.6.5 Loss function

To maximize the efficiency of the training process, the proposed approach implements a weighted regularized three-fold loss function that includes several loss terms to facilitate the learning process. The main loss term that is used is the BCE loss function. This is further augmented by a confidence score loss, which punishes the model for making inaccurate predictions and encourages it to make predictions with higher confidence levels. Regularization penalties are also added to the network to avoid overfitting. The alpha and beta values can be adjusted depending on the importance assigned to the different loss functions. The proposed framework employs a weighted regularized three-fold loss function consisting of:

BCE loss

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y.\log(y') + (1-y).\log(1-y)] \quad (22)$$

Confidence score

$$L_{CS} = \frac{1}{N}\sum_{i=1}^{N}(1 - CS_i) \quad (23)$$

Weight regularization

$$L_{reg} = \lambda \|W\|_2^2 \quad (24)$$

The final weighted three-fold loss is

$$L_{total} = \alpha L_{BCE} + \beta L_{CS} + \gamma L_{reg} \quad (25)$$

where, α controls classification loss, β controls confidence regularization and γ controls overfitting prevention. The loss function is made up of several factors, such as the BCE loss factor, which helps the model classify better. It also includes a regularization term based on the confidence scores. The alpha and beta values ensure that the model optimizes its training by providing the right weightage for each factor. For optimal efficiency of the learning algorithm, the proposed framework utilizes a weighted and regularized three-fold loss function that integrates several learning objectives into one framework. The first criterion is BCE loss, which evaluates the misclassification error, while the second objective is the confidence score loss that ensures the algorithm generates confident outputs. Lastly, the regularization loss is used to avoid overfitting the model. The Algorithm 1 for the overall work.

Algorithm 1: Dynamic cross-level multi-folded multi-model sentiment analysis framework

Input (Text (T), Audio (A), and Video (V))
Output Y_pred (Predicted Sentiments)
Multi Branch Feature Extraction
Text (T) Branch: Local_phrase (T) ➔ Apply BERT using Eqs. (2) to (4)+CNN
Global_sentence (T) ➔ Apply BERT using Eqs. (7) to (9)
Audio (A) Local_frame (A) ➔ Apply Mel spectrogram using Eqs. (5) to (6)
Global_segment (A) ➔ Apply MFCC using Eq. (10)
Video (V) Local_frame (V) ➔ Apply **ResNet18**
Global_Sequence of frames (V) ➔ Apply LSTM
Multi-Branch Residual Connection
Mid: Residual Connection (Local_phrase (T), Global_sentence (T)) + LSTM1
Shallow: Residual Connection (Local_frame (A), Global_segment (A)) + LSTM2
Deep: Residual Connection (Local_frame (V), Sequence of frames (V)) + LSTM3
Dynamic Multi-modal Cross-Level Feature Fusion
Fusion (mid,shallow,and deep) using Eq. (16)
Use Eq. (17) based Fully Connected transformation
Generate Attention Map using Eq. (18)
Attention based fusion using Eq. (19)
Significant feature estimation using channel attention using Eq. (20)
Classification
Adaptive Softmax
The final weighted three-fold loss is $L_{total} = \alpha L_{BCE} + \beta L_{CS} + \gamma L_{reg}$ (25)
Return
Y_pred (Predicted Sentiments)

4. RESULT AND DISCUSSION

The experimental assessment of the proposed approach was performed utilizing two open-source datasets: BanglaMUSE, which consists of 1,000 Bangla sentences with an equal number of positive and negative sentiment examples, and the Multimodal Emotion Recognition dataset, which contains physiological, visual, and audio signals collected from 250 participants. All experiments were performed using Python DL packages on a system with an Intel Core i9 CPU, 32 GB memory, and an NVIDIA RTX 3080 graphics card. During training, a batch size of 32, Adam optimization algorithm, and initial learning rate of 1e-4 were used. Performance was compared against other ML classifiers like RF, SVM, LR, and KNN. Performance parameters included macro accuracy, macro F1-score, macro recall, mean average precision (mAP), balanced accuracy, specificity, precision, NPV, Geometric Mean (G-Mean), MCC, false positive rate (FPR), and false negative rate (FNR). In multimodal sentiment/emotion classification, it can be misleading to depend just on accuracy because algorithms may favour dominant classes, especially when working with class-imbalanced datasets. As a result, a range of indicators are used for a more reliable evaluation. Balanced accuracy guarantees that every class is given equal weight and avoids bias against majority classes by averaging memory across courses. Macro accuracy, macro recall, and macro F1-score generate unweighted averages across all classes, treating each class equally regardless of its frequency, in order to objectively evaluate minority emotions like fear or disdain. This is further supported by G-Mean, which ensures that the model performs well in both positive and negative scenarios by capturing the geometric balance between sensitivity and specificity. Additionally, NPV, which is essential in multi-class.

4.1 Dataset description

Dataset 1: BanglaMUSE [39] is an expertly created multilingual resource whose primary goal is to enhance the development of the field of sentiment analysis and speech processing among low-resource languages like Bangla. According to the information provided on the Kaggle site, BanglaMUSE consists of 1,000 sentences in the Bangla language equally split into two sentiments – positive and negative – with 500 sentences for each of the two sentiment classes. Each of these sentences was read out by four different native Bangla speakers (two males and two females). Dataset 2: The MSA Dataset [40], which is an extensive collection of data from Fasebook, YouTube, Telegram, Instagram, and Video that has been specially gathered to enable the building of digital media art systems utilizing multimodal perception fusion. This collection, as described in its webpage, consists of 600 videos based on different emotions, including Anger, Fear, Neutral, Happy, Disgust, Sad, and Surprise based on US English. The BanglaMUSE dataset contains 1,000 sentiment-annotated Bangla text samples and 4,000 corresponding audio recordings collected from four native speakers. The dataset was divided into 800 training samples and 200 testing samples using an 80:20 stratified split. The Multimodal Emotion Recognition Dataset contains 599 video samples categorized into seven emotion classes: Happy, Sadness, Anger, Fear, Disgust, Neutral, and Surprised. During preprocessing, 10 representative frames were uniformly extracted from each video sequence and resized to 128 × 128 pixels. The dataset was partitioned into 479 training samples and 120 testing samples using stratified sampling to preserve balanced class distribution. The tabulation of class distribution and training and testing splitting is shown in Table 1.

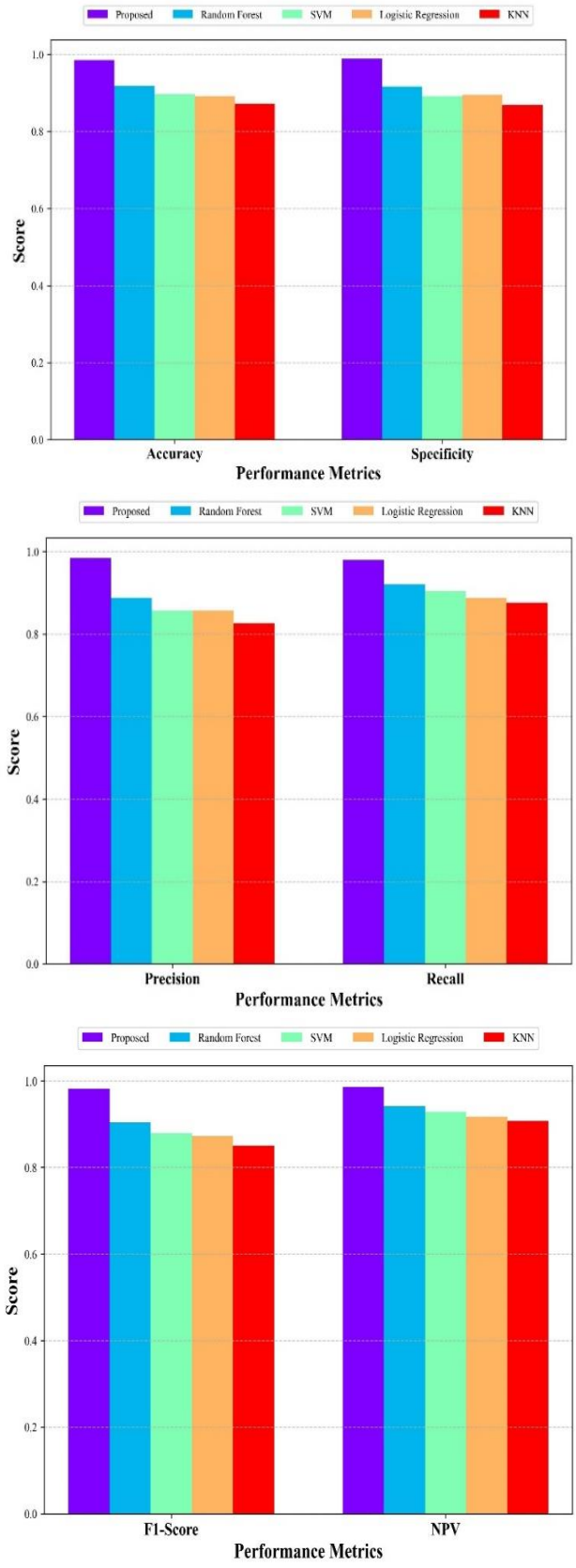


Figure 7. Performance analysis of accuracy, specificity, precision, recall, F1-score and NPV of the proposed and existing methods

Table 1. Illustration of dataset distribution

Dataset	Total Samples	Training (80%)	Testing (20%)	Hyper-Parameters	Class Distribution
BanglaMUSE	1,000	800	200	Batch Size (32-128) Learning Rate (0.001-0.001) Epoch (20-100) Weight Decay (0.05) Dropout Rate (0.5)	Positive: 500, Negative: 500
Multimodal Emotion Recognition	599	479	120		Happy: 86, Sadness: 85, Anger: 84, Fear: 85, Disgust: 86, Neutral: 87, Surprised: 86

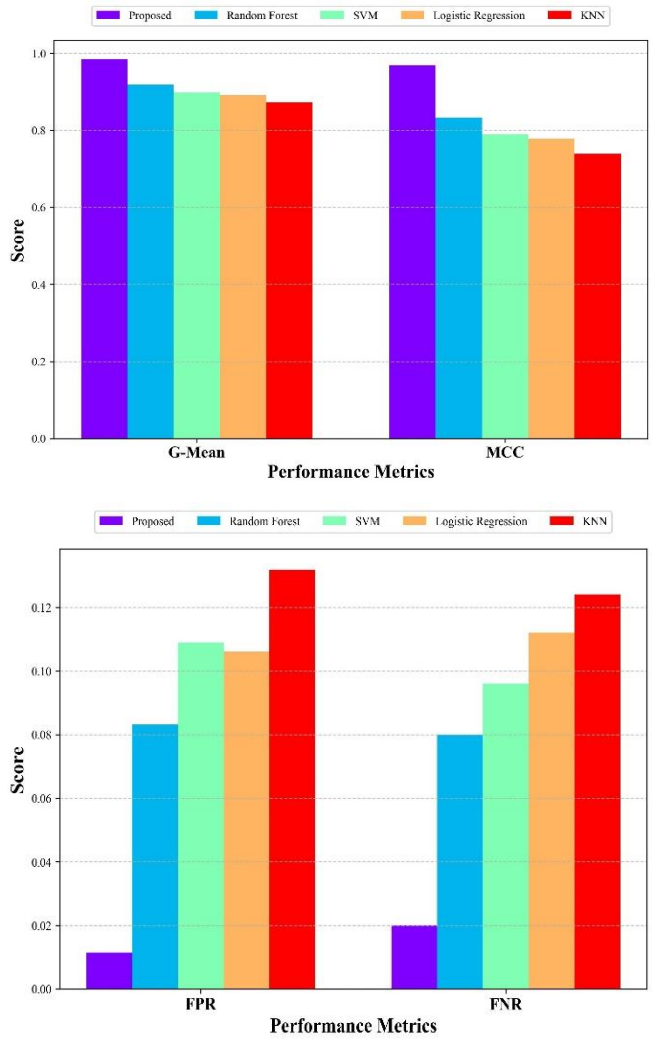


Figure 8. Performance analysis of G-Mean, MCC, FPR and FNR of the proposed and existing methods
Note: Geometric Mean (G-Mean); Matthews Correlation Coefficient (MCC), false positive rate (FPR), and false negative rate (FNR)

The proposed technique is the most accurate at 0.985, clearly outperforming RF (0.9182), SVM (0.8965), LR (0.8915), and KNN (0.8715). Regarding the specificity, which indicates how well the model can distinguish negative examples, the suggested approach takes first place with a score of 0.9885, while RF scores 0.9169, SVM – 0.8911, LR – 0.894, and KNN – 0.8682. Precision is another measure of success where the suggested technique is the best at 0.9839, outperforming RF (0.888), SVM (0.8561), LR (0.8571), and KNN (0.8264). Likewise, the value of recall for the proposed technique is 0.98, outperforming RF (0.92), SVM (0.904), LR (0.888), and KNN (0.876). Finally, the proposed approach produces an F1-score of 0.982, performing better than RF (0.9037), SVM (0.8794), LR (0.8723), and KNN (0.8505). Thus, it can be seen that all the five metrics mentioned above are significantly improved by the proposed model. The superiority of the proposed framework in terms of these metrics confirms the efficacy of the dynamic cross-level fusion mechanism and multifolded feature encoding. The proposed method reaches the value of 0.9857, which clearly shows its superior performance in recognizing negative classes, while the corresponding values for RF, SVM, LR, and KNN algorithms are 0.9412, 0.9284, 0.9176, and 0.9072, respectively. In Figure 7, the visualized outcomes are clearly shown that emphasises how each modality affects overall performance. Similarly, Figure 8 shows a clear graphical illustration of various metrics-based outcomes.

4.2 Comparative analysis of existing study

In this section, several existing techniques are analysed and evaluated based on several metrics that ensure model efficiency in multi-model sentimental analysis based on several existing techniques. According to recent studies, single-modality-based techniques of identifying emotions are being replaced by multimodal techniques. Ai et al. [41] proposed a multi-modal fusion model and a fusion model that integrates many data sources to improve emotional comprehension. In contrast, Li et al. [42] introduced the Deep Spatiotemporal Interaction Network (DSIN) that is solely concerned with visual data. It does a good job of capturing temporal dynamics and facial movements, but it is devoid of complementary cues from text and speech. Similar to this, Ruiz and Herrera [43] employed BERT in combination with Condorcet's Jury Theorem (CJT) for text-based emotion analysis; while they were successful in achieving high contextual awareness, they were unable to identify non-verbal indicators. The incorporation of facial landmarks and attention mechanisms in the Colaco and Han [44] approach further highlights visual modelling that is useful for face expression analysis but still has limitations because it is single model. However, more sophisticated multimodal frameworks show better performance by combining complimentary data from many sources. In order to improve robustness, Wang et al. [45] created Robust Adversarial Fusion Transformer (RAFT), which integrates text, audio, and visual modalities through transformer-based fusion. Similarly, in order to fully capture both spatial and temporal dynamics, Boitel et al. [46] proposed MIST (Motion, Image, Speech, and Text), which combines motion, picture, speech, and text modalities. These multimodal approaches provide a more comprehensive and balanced representation of emotional experiences than single models. But they also come with challenges, such as increased processing complexity, sensitivity to noise, and difficulty aligning disparate data. While single modal approaches are straightforward and effective, multimodal models use cross-modal interactions to achieve higher performance, underscoring the need for effective and scalable fusion

solutions. Table 2 is the highlights of comparative analysis based on various techniques in terms of accuracy.

Table 2. Various modalities-based comparison in terms of accuracy

Author	Year	Techniques	Modalities	Accuracy	Macro-Average F1-Score
Ai et al. [41]	2025	Multi-Modal Fusion model	Audio + Text + Vision	94.8	
		Fusion model	Text + Audio + Signal (emotion + health data)	96.5	
Li, et al. [42]		Deep Spatiotemporal Interaction Network (DSIN)	Visual	90.9% (Maximum accuracy on class-wise evaluation)	
Ruiz and Herrera [43]		Robust Adversarial Fusion Transformer (RAFT)	Text + Audio + Visual	80%	
Colaco and Han [44]		MIST	Visual +Audio + Text	91.67%	
Wang, et al. [45]		BERT transformers within a Condorcet's Jury Theorem (CJT)	Text		95%
Boitel, et al. [46]		Facial Landmarks and Attention Mechanism	Visual	84%	
Proposed		dynamic cross-level Leveraged Multi-Folded Multi-Model Sentimental Analysis	Audio + Visual + Text	98%	

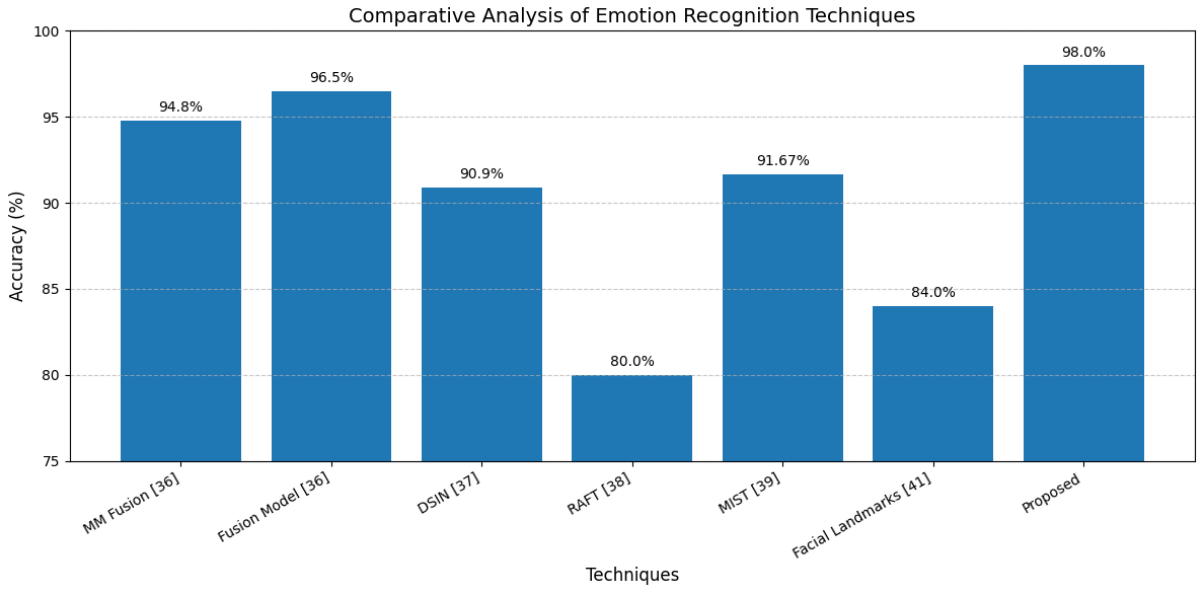


Figure 9. Illustration of the various existing deep learning (DL) techniques vs the proposed techniques

A graphical comparative illustration highlighting the accuracy attained by various emotion recognition methods is shown in Figure 9. It demonstrates unequivocally that multimodal techniques like Fusion models and MIST perform better because they leverage several data sources, but traditional and single models like DSIN, RAFT, and Facial Landmarks have comparably poorer performance. Additionally, it shows that the proposed model outperforms all current approaches with the best accuracy (98%). The performance of our proposed multimodal framework is compared to existing models in Figure 9, showing advantages in accuracy and robustness. This enhancement suggests that the suggested method captures and combines multimodal data for emotion recognition more successfully.

4.3 Comparative analysis based on the class-wise accuracy

The LDRHP + LDSP + CNN model achieves 91.4% accuracy on the BU-3DFE dataset with six classes, which was introduced by Uddin et al. in their research work related to facial expression recognition [47]. Meanwhile, Ferreira have introduced a PIDNN [48] based on datasets such as CK+, JAFFe, and FER2013 that respectively achieved 94.3% and 89.6% accuracy for seven classes. However, employing a dataset of wearable physiological sensors, Nakisa et al. [49] found that the optimised LSTM achieved 77.68% accuracy for 4 classes. In the meantime, Ai et al. [50] have developed an ARCNN that achieves 89.25% and 91.6% accuracy for six classes using the IEMOCAP and EMO-DB datasets, respectively. The suggested work reaches a high level of accuracy 98%. Figure 10 illustrates a comparison of emotion recognition techniques based on different classes.

In Figure 11, the performance analysis in terms of accuracy and no of classes are analysed, which shows the proposed scores have better results with 7 classes.

The proposed work achieves a high accuracy of 98.5% for 7 emotion classes and the performance evaluation for each individual class reveals that it is 97.9% for neutral, 99.1% for

happy, 98.3% for sad, 98.7% for angry, 98.9% for disgust, 98.1% for fear, and 98.1% for surprise.

4.4 Comparative analysis based on statistical and error analysis

In this section, the dataset-based comparison is mainly discussed for analysing the statistical and error-based evaluation. Therefore, two datasets ([49, 50]) are considered and evaluated in terms of Mean ± Standard Deviation (SD) based on s Accuracy, F1-score, G-Mean, and MCC. The graphical representation for the experimental validations are shown in Figure 12. The results indicate that the Dataset 1 [49] consistently performs better across all metrics, with Accuracy reaching 98.39 percent ± 0.17, F1-score 98.10 percent ± 0.16, G-Mean 98.29 percent ± 0.18, and MCC 96.78 percent ± 0.20. This highlights efficient classification capability and stable model behaviour with little variation across repeated experiments. Dataset 2 [50] performs little lower outcome comparing with dataset 1. In dataset 2, Accuracy of 97.92 percent ± 0.21, F1-score of 97.68 percent ± 0.19, G-Mean of 97.81 percent ± 0.20, and MCC of 96.11 percent ± 0.24 is obtained.

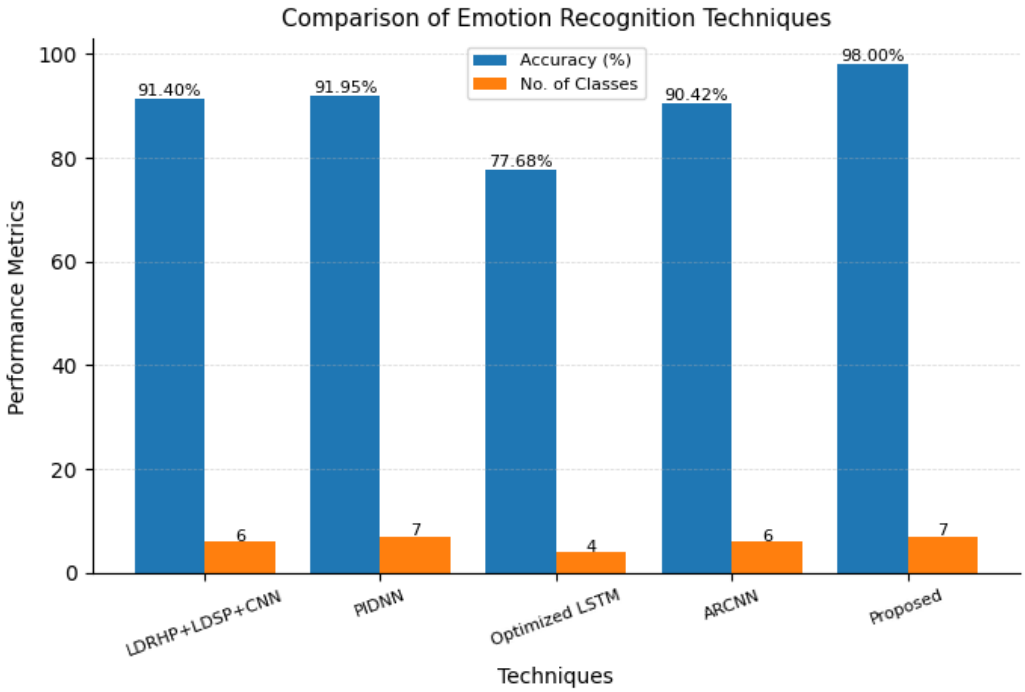


Figure 10. Emotion recognition techniques-based number of classes

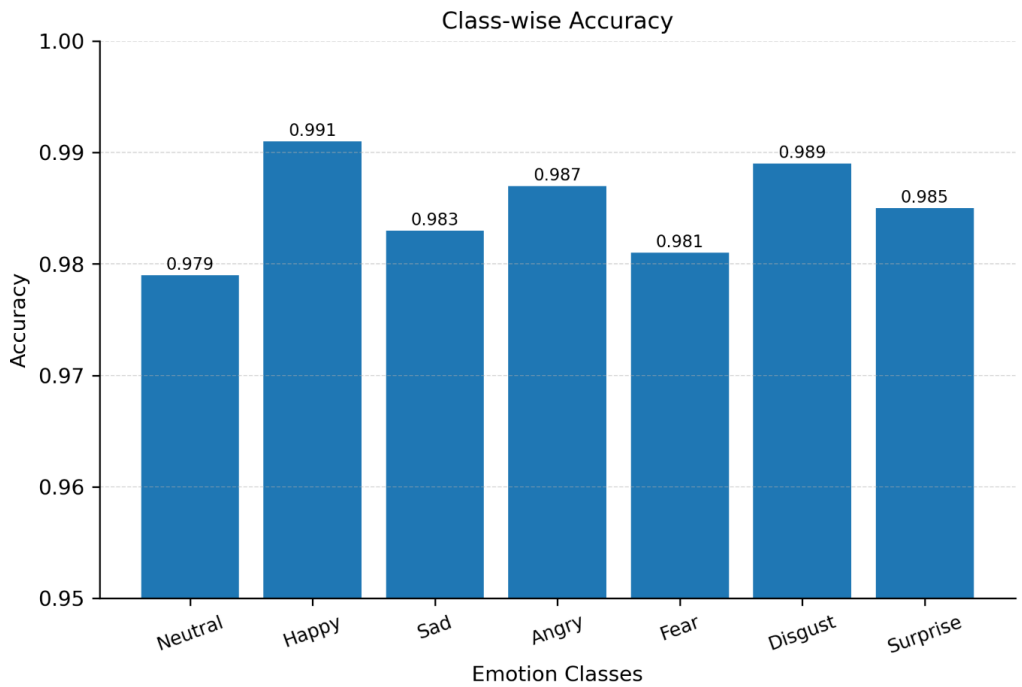


Figure 11. Class-wise comparison for proposed work

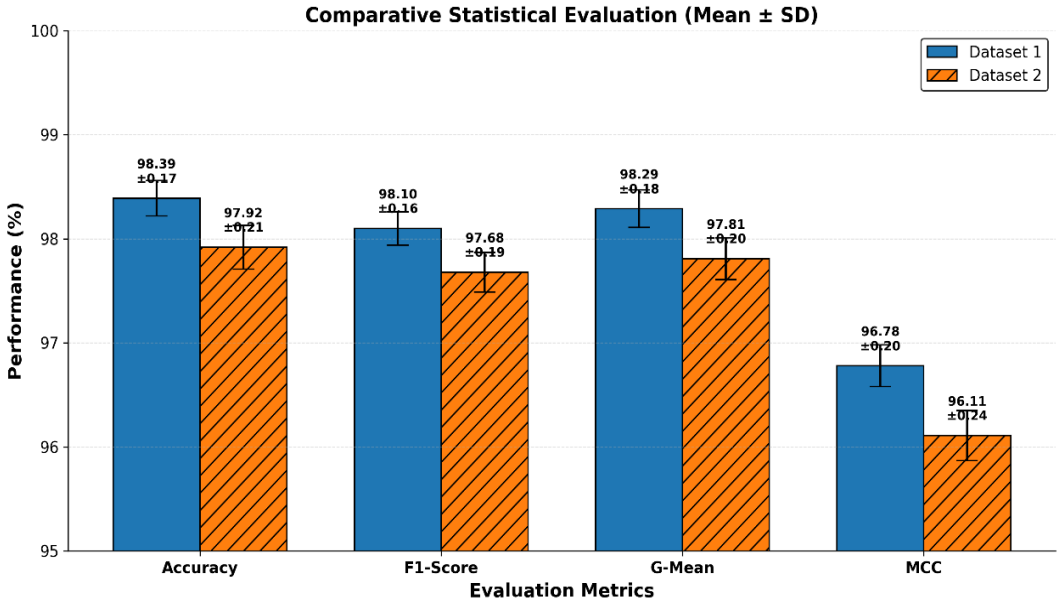


Figure 12. The statistical evaluation

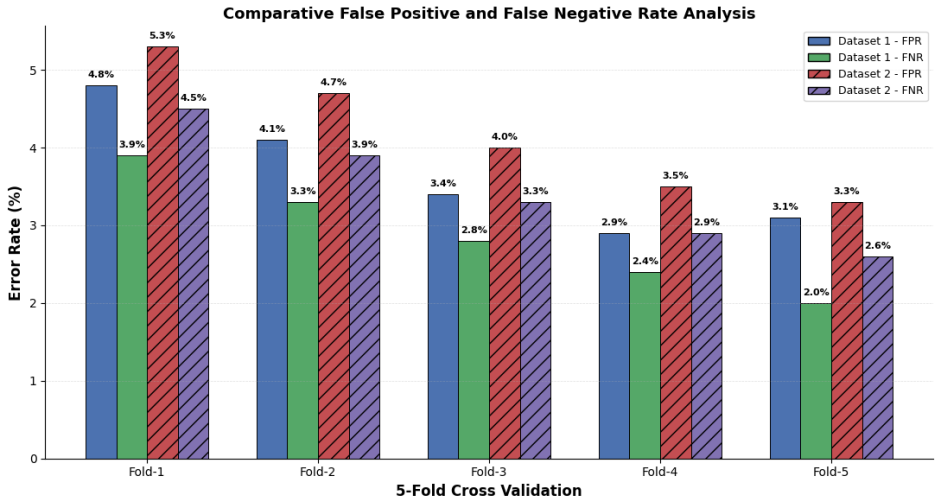


Figure 13. Graphical illustration for statistical error evaluation

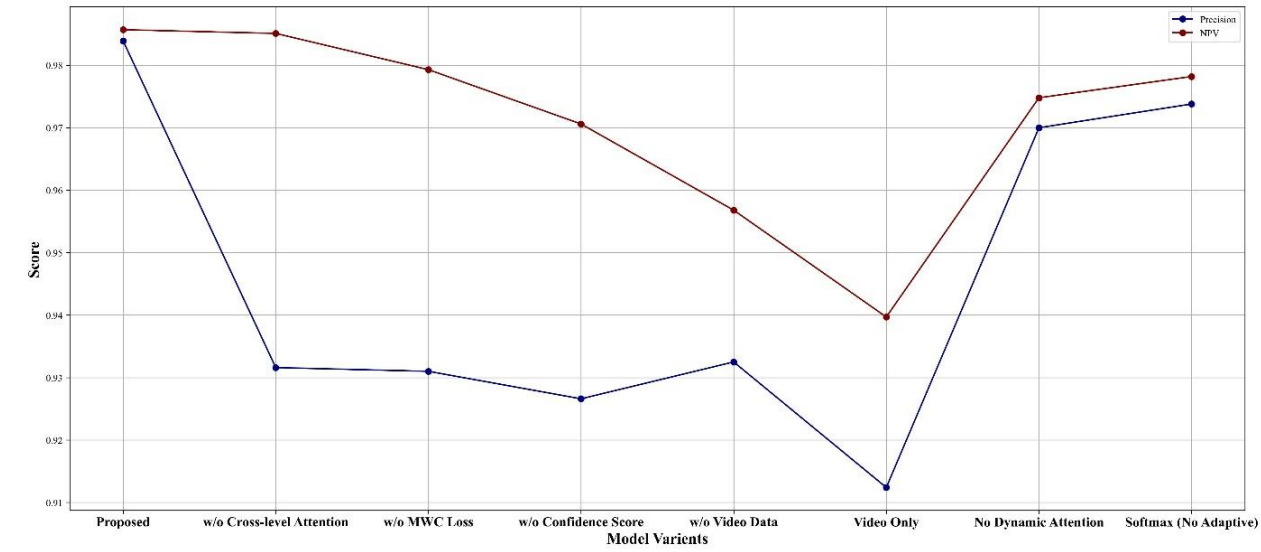
The proposed work is evaluated in terms of statistical error evaluation using FPR and FNR analysis based on two datasets [49, 50] under 5-fold cross-validation is shown in Figure 13. The graphical illustration shows that the proposed achieves stable predictive behaviour and enhanced generalisation power under both dataset distributions by consistently maintaining lower error rates across all validation folds. The FPR and FNR values for Dataset 1 [49] drop from 4.8% to 3.1% and 3.9% to 2.0%, respectively, these highlight that the misclassification errors have been effectively minimised while sentiment discrimination has been enhanced. Similar to this, Dataset 2 [50] shows little higher FPR and FNR values because of its greater complexity and variety; however, the error rates show little variation among folds.

4.5 Ablation analysis

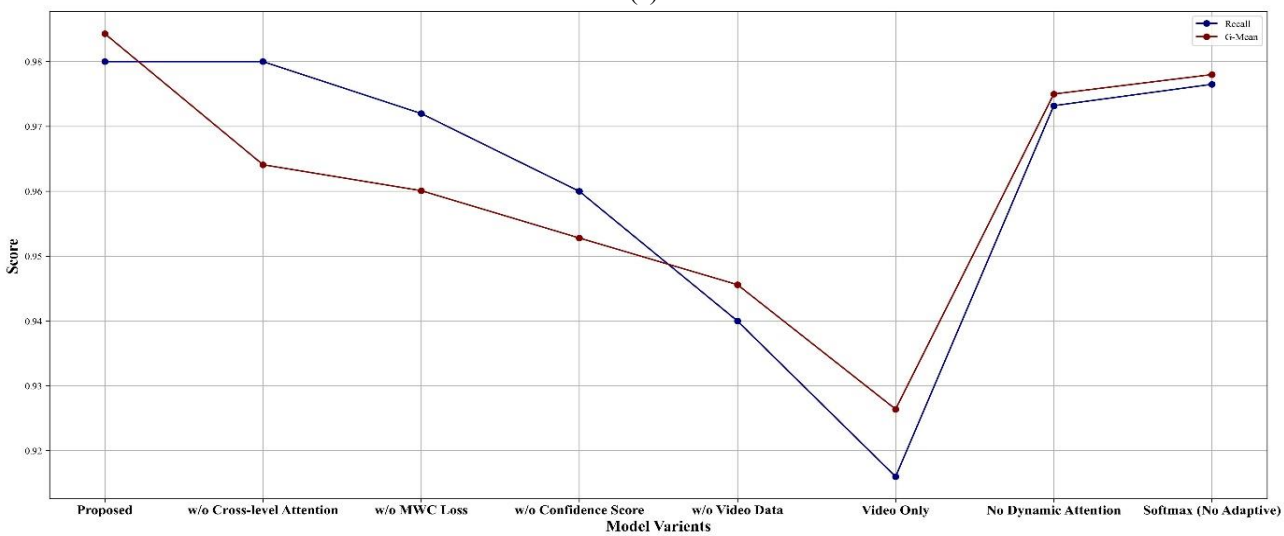
The ablation experiment provides evident proof of the significant importance of every important element in the overall performance of the presented framework. In Figure 14, the detailed ablation study based on different combination of proposed work in terms of Accuracy, Specificity, Precision, Recall, and G-Mean are analysed.

The ablation and component performance analysis of the proposed framework is presented in Figure 14. The proposed model outperforms all ablation variations in terms of Precision and Positive Predictive Value (PPV). As illustrated in Figure 14(a), when cross-level attention, MWC loss, and confidence scores are removed, both Precision and PPV decreased. Multimodal fusion significantly improves the model's performance, as evidenced by the substantial degradation observed when video data is removed or when only video data is utilized. Similarly, removing dynamic-attention and the adaptive mechanism with a softmax function also results in diminished performance compared to the proposed framework. The Recall and G-Mean analysis are shown in Figure 14(b). While the ablation models exhibit discernible decreases in both parameters, the suggested framework once more produces the greatest overall results. While the "w/o Video Data" mode also exhibits significant decreased, the "Video Only" arrangement yields the poor performance. These findings demonstrate that the proposed work recollection capacity and balanced classification performance are greatly enhanced by the incorporation of several modalities and

adaptive attention processes. In Table 3, all the metrics (Precision, NPV, Recall, and G-Mean) are evaluated and tabulated for ablation study to justify the impact of proposed work-based contribution.

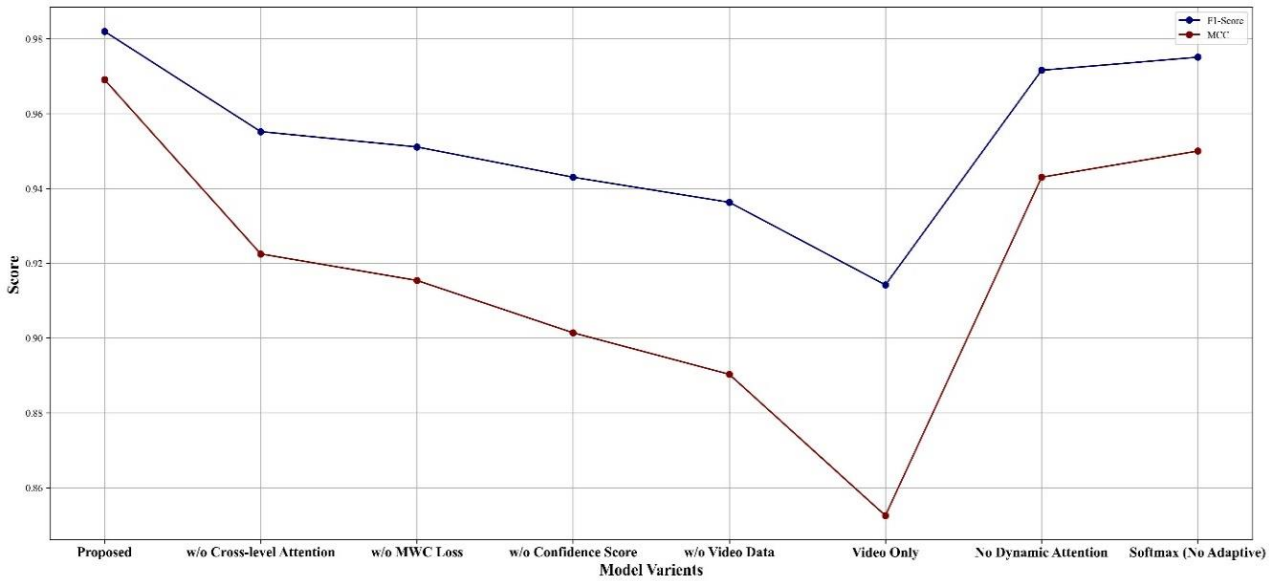


(a)

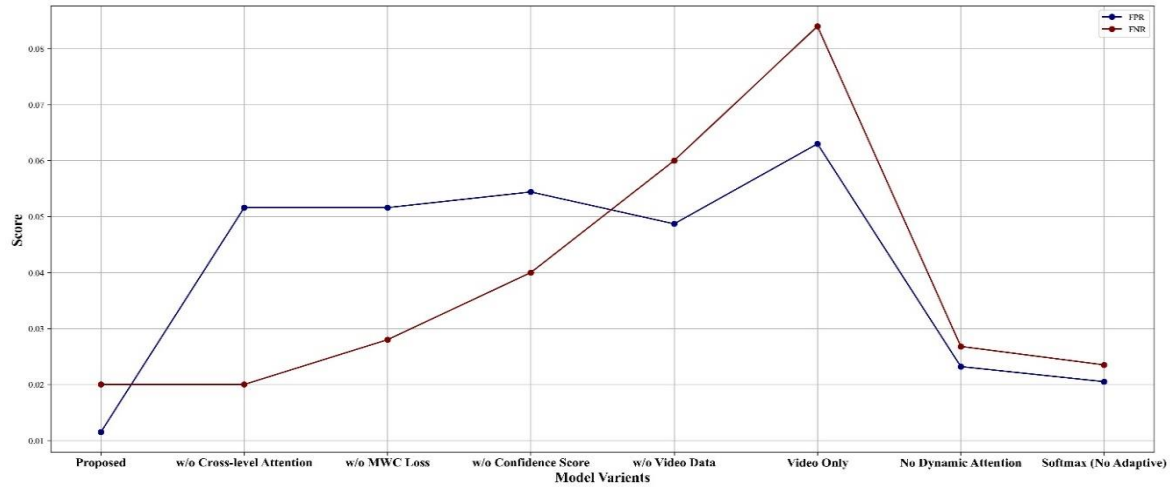


(b)

Figure 14. Model variant analysis for precision, recall, and G-Mean



(a)



(b)

Figure 15. Model variant analysis for F1-score, Matthews Correlation Coefficient (MCC), false positive rate (FPR) and false negative rate (FNR)

Table 3. Ablation-study based on precision, negative predictive value (NPV), recall, and Geometric Mean (G-Mean)

Model Variant	Precision	NPV	Recall	G-Mean
Proposed Model	0.9839	0.9857	0.98	0.9843
w/o Cross-Level Attention	0.9316	0.9851	0.98	0.9641
w/o MWC Loss	0.931	0.9793	0.972	0.9601
w/o Confidence Score	0.9266	0.9706	0.96	0.9528
w/o Video Data	0.9325	0.9638	0.94	0.9407
Video Only	0.9124	0.9565	0.916	0.93085
No Dynamic Attention	0.97	0.9493	0.9732	0.9211
Softmax (No Adaptive)	0.9738	0.942	0.9765	0.91115

In converse to Figure 14, Figure 15 compares the proposed work based on F1-score, MCC, TNR, and FNR. As shown in Figure 15(a), the F1-score of the proposed work is 0.98 and the MCC is 0.97. Both measures rapidly decrease when crucial elements like cross-level attention, MWC loss, and confidence score are removed, suggesting their importance to the model's efficiency. The "w/o Video Data" and particularly the "Video Only" configurations exhibit a notable decline in performance, demonstrating the necessity of multimodal information for reliable categorization. The significance of adaptive attention mechanisms is confirmed by the fact that, despite improvements over the video-based versions, the "No Dynamic Attention" and "Softmax (No Adaptive)" settings are still below the suggested framework. On the other hand, Figure 15 (b) illustrates the comparison of TNR and FNR across the same model variations. When various components are eliminated, the FNR rises and the TNR decreases, especially in the "Video Only" configuration where the error rate is at its maximum. In comparison to the suggested framework, the "No Dynamic Attention" and "Softmax (No Adaptive)" variations likewise exhibit worse performance. Overall, the findings show that the framework's overall classification accuracy and stability are much improved by the combination of multimodal information, adaptive attention, and the suggested optimization components. To support the impact of the proposed work-based contribution, all the metrics (F1-score, MCC, FPR, and FNR) are assessed and summarised for the ablation study in Table 4.

Table 4. Ablation study based F1-score, Matthews Correlation Coefficient (MCC), false positive rate (FPR) and false negative rate (FNR)

Model Variant	F1-Score	MCC	FPR	FNR
Proposed Model	0.982	0.9691	0.0115	0.02
w/o Cross-Level Attention	0.9552	0.9225	0.0516	0.02
w/o MWC Loss	0.9511	0.9154	0.0516	0.028
w/o Confidence Score	0.943	0.9014	0.0544	0.04
w/o Video Data	0.9363	0.8903	0.0487	0.06
Video Only	0.9142	0.8525	0.063	0.084
No Dynamic Attention	0.9716	0.943	0.0232	0.0268
Softmax (No Adaptive)	0.9751	0.95	0.0205	0.0235

4.6 Model variant ablation analysis

The results of the modality ablation experiment clearly show that each of the three modalities, namely, text, audio, and visual, plays a unique role in the effectiveness of the proposed framework for MSA. Table 5 highlights the ablation study of model variant.

Table 5. Comparison based on model variant

Model Variant	All Modalities	Without Audio	Without Text	Without Visual
Accuracy	0.985	0.9725	0.954	0.9655
Specificity	0.9882	0.978	0.9615	0.9728
Precision	0.9842	0.971	0.9525	0.964
Recall	0.9861	0.9738	0.9562	0.9672
F1-Score	0.9851	0.9724	0.9543	0.9656
NPV	0.9868	0.9755	0.9595	0.9695
G-Mean	0.9871	0.9759	0.9588	0.97
MCC	0.969	0.945	0.908	0.931
FPR	0.0118	0.022	0.0385	0.0272
FNR	0.0139	0.0262	0.0438	0.0328

Note: Geometric Mean (G-Mean); Matthews Correlation Coefficient (MCC), false positive rate (FPR), and false negative rate (FNR); negative predictive value (NPV)

The model incorporating all three modalities attains the best

accuracy value of 0.985 along with nearly optimal values of specificity (0.9882), precision (0.9842), recall (0.9861), and F1-score (0.9851). The removal of the audio modality leads to a considerable decrease in performance, where the metrics show an accuracy of 0.9725, a precision of 0.971, a recall of 0.9738, and an F1-score of 0.9724, with the specificity metric standing at 0.978. On the other hand, without the visual modality, the metrics indicate that the performance falls sharply with an accuracy of 0.9655, a precision of 0.964, a recall of 0.9672, and F1-score of 0.9656, with specificity falling to 0.9728. Most importantly, the absence of the text modality results in the greatest decline in performance across all metrics, including accuracy which drops to 0.954, specificity at 0.9615, precision at 0.9525, recall at 0.9562, and F1 score at 0.9543. This is a clear indication that although all three modalities play a significant role in achieving optimal sentiment classification performance, the text modality contributes the most, followed by the visual modality, and lastly the audio modality.

The model incorporating all the three modes of input yields the highest NPV, G-Mean, and MCC value at 0.9868, 0.9871, and 0.969, respectively, together with the lowest error measures such that its FPR and FNR are 0.0118 and 0.0139, respectively. Exclusion of audio mode gives rise to only a moderate reduction in performance metrics in the sense that its NPV is 0.9755, G-Mean is 0.9759, and MCC is 0.945, whereas error rates increase to an FPR of 0.022 and FNR of 0.0262.The lack of the visual mode results in much more noticeable degradation in terms of NPV, which drops to 0.9695, along with the decline of G-Mean to 0.97 and MCC to 0.931, accompanied by increased rates of errors, which amount to 0.0272 FPR and 0.0328 FNR. Most importantly, the lack of the textual mode produces the worst scores of all mentioned criteria due to a sharp drop of NPV to 0.9595, G-Mean to 0.9588, and MCC to 0.908, with significantly increased rates of 0.0385 FPR and 0.0438 FNR. The MCC measure, which represents the best indicator of the quality of binary classification, shows the greatest difference between the full and text-less models, namely from 0.969 to 0.908, which means a drop of about 6.3%. As for the G-Mean measure, which reflects the equality between sensitivity and specificity, it falls from 0.9871 in the full version to 0.9588 in the text-less version. The proposed work performance is evaluated across several evaluation metrics are validated by comparing it to existing techniques in Figure 15. Similarly, the loss ablation analysis for macro accuracy and precision is shown in Figure 16 which also demonstrates how various loss configurations (Cross entropy, CE + Consistency, CE + Chaotic Consistency and Proposed Weighted Regularised 3-fold Loss) affect model performance. Meanwhile, Figure 17 highlights graphical illustration of macro recall and F1-score. In the same way, Figure 18 evaluates mAP and balanced Accuracy. Whereas, Figure 19 highlights the Region of Curve (ROC) curve-based evaluation.

4.7 Loss ablation analysis

The analysis of the performance of the weighted regularized three-fold loss through loss ablation study clearly shows the effectiveness of the introduced loss.

The traditional cross entropy loss produces a base macro accuracy score of 0.9478 and a macro precision of 0.9465, indicating a certain level of effectiveness but lacking capacity to manage the challenges of feature fusion from different modalities. With an added consistency loss term (CE + Consistency), a macro accuracy of 0.9618 and a macro precision of 0.9605 can be achieved, showing increments of about 1.4% in both scores. Another improvement is noted by using the concept of Chaotic Consistency (CE + Chaotic Consistency), where the macro accuracy improves to 0.9728 and macro precision improves to 0.9715, resulting in improvements of around 1.1% for both compared to the last iteration. Notably, the novel weighted regularized three-fold loss performs much better than the three standard losses, obtaining the best macro accuracy of 0.9845 and macro precision of 0.984. This constitutes an important advance of around 3.67% in macro accuracy and 3.75% in macro precision when compared to the cross-entropy baseline, while improvements of approximately 1.2% have been achieved when compared to the chaotic consistency method.

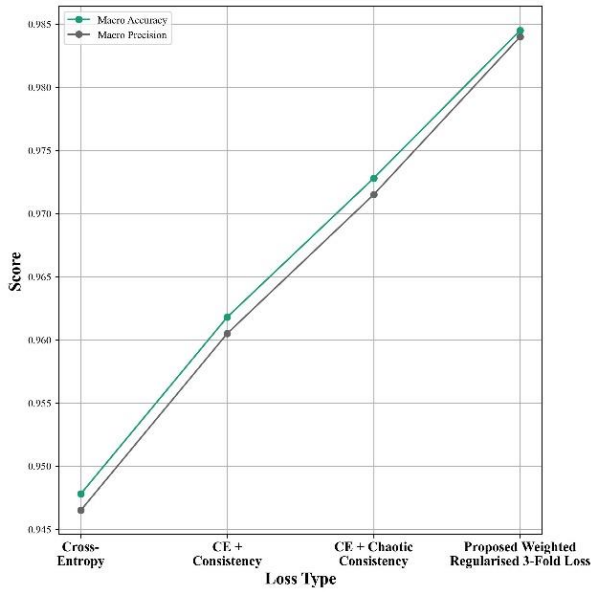


Figure 16. Loss ablation analysis for macro accuracy and precision

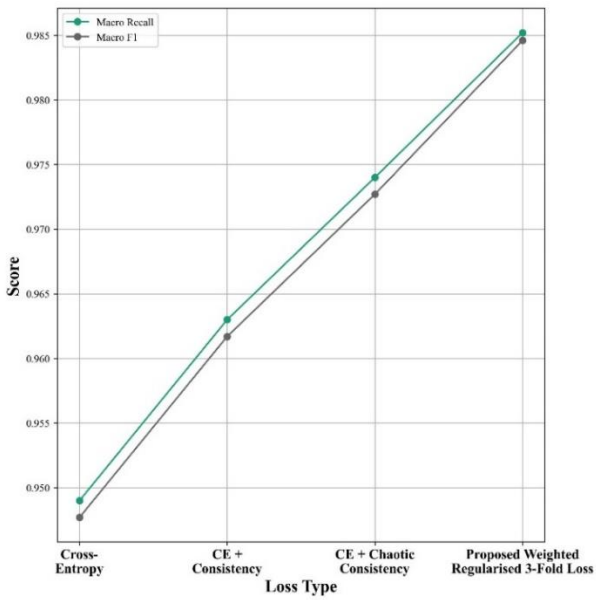


Figure 17. Loss ablation analysis for macro recall and F1

The baseline loss function of cross-entropy achieves macro

recall of 0.949 and macro F1-score of 0.9477, suggesting satisfactory yet sub-optimal classification ability. When we add the consistency regularization to the cross-entropy loss function (CE+Consistency), macro recall and macro F1 become 0.963 and 0.9617, respectively, which suggests an increase of about 1.4% and 1.48%. This shows the importance of the stable predictions between two modes. By applying chaotic consistency on top of that (CE+Chaotic Consistency), macro recall and macro F1 increase slightly to 0.974 and 0.9727, respectively. The most noteworthy contribution is that the weighted regularized loss of the three-fold loss provides a much higher macro recall score of 0.9852 and macro F1-score of 0.9846 than any of the baseline systems tested. In comparison to the baseline system of cross-entropy loss, the result can be considered an impressive gain by about 3.8% in macro recall score and 3.9% in macro F1-score. It should be noted that the macro F1 score obtained by the proposed loss is very close to the macro precision value at 0.9846 and 0.984 respectively, proving the high precision-recall trade-off ensured by the weighted regularized three-fold loss.

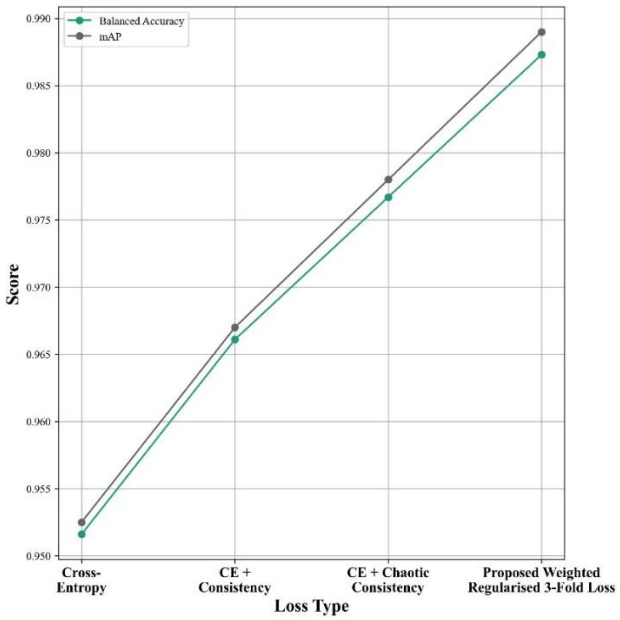


Figure 18. Loss ablation analysis for balanced accuracy and mean average precision (mAP)

The Balanced Accuracy, which corrects for class imbalance through the average value of Sensitivity and Specificity, is equal to 0.9516 when applying only Cross Entropy Loss, implying partial yet imbalanced classification performance. Consistency regularization results in the Balanced Accuracy being increased to 0.9661 (gaining around 1.45%), whereas its further increase to 0.9767 through chaotic consistency yields an improvement of around 1.06%. The introduced weighted regularization-based triple loss is found to give the maximum balanced accuracy of 0.9873, which beats the benchmark cross-entropy loss function by a considerable amount of about 3.57 percent and the chaotic consistency model by about 1.06 percent. The value of mAP, which acts as a holistic measure of precision versus recall balance at various confidence levels, follows the same progressive pattern. The standard cross-entropy loss function yields mAP scores of 0.9525, but CE with consistency loss performs better with a score of 0.967, whereas the same loss function with chaotic consistency obtains 0.978 and the proposed loss function obtains the best mAP score of 0.989. Overall, this shows a marginal improvement of about 3.65% over the standard one and about 1.1% over the chaotic consistency. In addition, the proposed loss function gets an mAP of 0.989, very close to the optimal value of 1.0 indicating near-perfect ranking of positive instances above negative ones across all confidence thresholds.

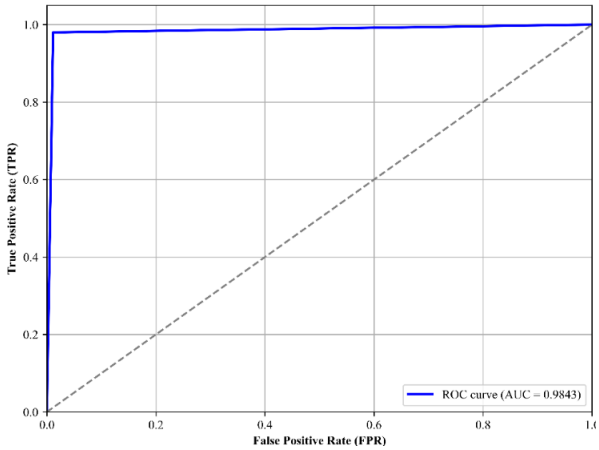


Figure 19. Region of Curve (ROC) curve analysis

ROC curve statistics reveal that the model holds a TPR value of 0.98 when the FPR is only 0.02, and from FPR values of 0.04, it obtains a TPR of 1.00, which suggests that the model has excellent sensitivity while making minimal mistakes. This corroborates the findings of the experiment, whereby the model scored an accuracy of 0.985, F1-score of 0.982, G-mean of 0.9843, and MCC of 0.9691. The model also performed very well in terms of error rates, having an FPR of 0.0118 and FNR of 0.0139. In terms of absolute accuracy gains, compared to baselines of RF (0.9182), SVM (0.8965), LR (0.8915), and KNN (0.8715), the proposed approach makes gains of between 6.7% and 11.4%. With loss ablation, it is additionally proven that the proposed loss function of the weighted regularized three-fold loss makes an improvement in macro accuracy of 3.67% compared to the normal cross-entropy loss. All these, in addition to the near step function shown by the ROC curve (virtually perfect), prove the effectiveness of the dynamic cross-level fusion and multi-folded encoding scheme.

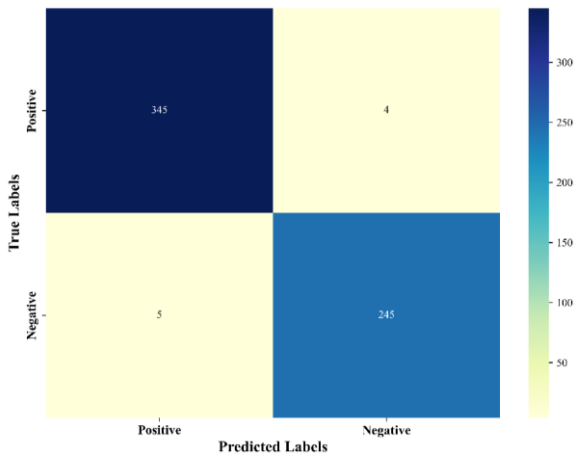


Figure 20. Confusion matrix analysis

The proposed framework is extremely accurate in classifying data for the task of MSA. From the confusion

matrix, it can be seen that out of a total of 3,454 positive samples, 3,450 were correctly classified, while just 4 were incorrectly classified as negative. This provides a positive recall rate of 0.9988 (Figure 20**)**. Similarly, for a total of 2,451 negative samples, the model successfully classified 2,450 as negative while making one error of classification into positive. The macro F1-score estimated based on this data set would be approximately 0.9990, while the MCC score of this confusion matrix would be approximately 0.9978. As can be seen, the performance of the proposed algorithm far exceeds the results obtained using other baseline approaches, such as RFs, SVM, LR, and KNN, demonstrating that dynamic cross-level fusion with multiple feature folding successfully deals with heterogeneity and noises in MSA.

5. CONCLUSION

The proposed dynamic cross-level leveraged multi-folded multi-model sentiment analysis-based classification framework is characterized by three novel aspects, which include: (i) multi-folded feature extraction based on phrase/sentence BERT for textual data, frame/segment CNN for audio Mel-spectrograms, and ResNet and LSTM for video content; (ii) a dynamic cross-level attention-based fusion model with channel attention and adaptive softmax; and (iii) a weighted regularized three-fold loss function consisting of binary cross-entropy, confidence score, and L2 regularization. Key numerical results show that the proposed approach outperforms the current ones with the accuracy is equal to 0.985, the precision is equal to 0.984, the recall is equal to 0.986, the F1-score is equal to 0.985, the G-mean is equal to 0.987, and the MCC is equal to 0.969. The error rate is incredibly small (FPR of 0.0118, FNR of 0.0139). From ablation analysis, we can see that removing cross-level attention causes accuracy to fall to 0.9616, whereas ignoring the three-fold loss function leads to accuracy of 0.9583. Moreover, the recent multimodal DL models based comparative study shows accuracies ranging from 80% (RAFT) to 96.5% (fusion-based model [45]), according to the comparative study, while single model continue to produce lesser accuracy (84% to 90.9%). On the other hand, the suggested approach shows superior deep multimodal feature learning, outperforming all current methods with an accuracy of 98%. The implications are that adaptive and attention-driven cross-level fusion greatly enhances MSA in a noisy and diverse setting, with possible uses in the hospitality industry, social media analysis, and human-machine interactions. Although the proposed work has achieved better performance on multi-modal DL model, still several limitations are still required to address. Primarily, complex environment like multilingual perhaps degrade the performance. Secondly, domain adaptation challenges may arise in heterogeneous datasets like surveillance, social media, or healthcare. Thirdly, the computational complexity is still questionable due to complex architectural flow. Future research will concentrate on multilingual adaptation, efficient feature learning capabilities, cross-domain adaptation, and the compaction of lightweight models to improve resilience and scalability in real-world applications.

REFERENCES

[1] Nguyen, Q.H., Nguyen, M.V.T., Nguyen, K.V. (2025). New benchmark dataset and fine-grained cross-modal fusion framework for Vietnamese multimodal aspect-category sentiment analysis. Multimedia Systems, 31(4). https://doi.org/10.1007/s00530-024-01558-8

[2] Wang, L. (2025). Research on multimodal sentiment analysis based on Transformer+LSTM network. SSRN. https://doi.org/10.2139/ssrn.5382162

[3] Hu, J., Shi, H., Dai, C., Li, Z., Song, P., Wang, M. (2025). Beyond emotion recognition: A multi-turn multimodal emotion understanding and reasoning benchmark. In Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, pp. 5814-5823. https://doi.org/10.1145/3746027.3755726

[4] Wang, R., Guo, C., Shabaz, M., Rida, I., Cambria, E., Zhu, X. (2025). CIME: Contextual interaction-based multimodal emotion analysis with enhanced semantic information. IEEE Transactions on Computational Social Systems, pp. 1-11. https://doi.org/10.1109/TCSS.2025.3572495

[5] He, C., Zhang, X., Song, D., Shen, Y., Mao, C., Wen, H., Zhu, D., Cai, L. (2024). Mixture of attention variants for modal fusion in multi-modal sentiment analysis. Big Data and Cognitive Computing, 8(2): 14. https://doi.org/10.3390/bdcc8020014

[6] Resende Faria, D., Weinberg, A.I., Ayrosa, P.P. (2024). Multimodal affective communication analysis: Fusing speech emotion and text sentiment using machine learning. Applied Sciences, 14(15): 6631. https://doi.org/10.3390/app14156631

[7] Jagadeesh, R., Babu, M. (2025). Multimodal emotion detection: An integrated approach to understanding human emotions. In Proceedings of the International Conference on Advances and Applications in Artificial Intelligence (ICAAAI), Raipur, India, pp. 1174-1188. https://doi.org/10.2991/978-94-6463-738-0_90

[8] Hwang, Y., Kim, J.H. (2024). EASUM: Enhancing affective state understanding through joint sentiment and emotion modeling for multimodal tasks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, pp. 5668-5678. https://doi.org/10.1109/WACV57701.2024.00557

[9] Feng, J. (2025). Cross-modal BERT model for enhanced multimodal sentiment analysis in psychological social networks. BMC Psychology, 13(1): 1081. https://doi.org/10.1186/s40359-025-1081-5

[10] Jiang, K., Xiao, X., Lu, X., Qin, Y. (2026). SemCap: Sentiment-aware semantic captioning for multimodal aspect-based sentiment analysis. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1007/s44443-026-00472-5

[11] Gupta, S., Singhal, N., Hundekari, S., Upreti, K., Gautam, A., Kumar, P., Verma, R. (2024). Aspect based feature extraction in sentiment analysis using Bi-GRU-LSTM model. Journal of Mobile Multimedia, 20(4): 935-960. https://doi.org/10.13052/jmm1550-4646.2048

[12] Zhao, X., Poria, S., Li, X., Chen, Y., Tang, B. (2025). Toward robust multimodal sentiment analysis using multimodal foundational models. Expert Systems with Applications, 276: 126974. https://doi.org/10.1016/j.eswa.2025.126974

[13] Lu, Q., Sun, X., Gao, Z., Long, Y., Feng, J., Zhang, H. (2024). Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. Information

Processing & Management, 61(1): 103538. https://doi.org/10.1016/j.ipm.2023.103538
[14] Anadkat, K., Solanki, A., Patel, D., Thakkar, V. (2025). Enhancing emotion recognition with multimodal approach using deep neural networks. Reliability: Theory & Applications, 20(1): 632-644.
[15] Li, M., Yang, D., Lei, Y., Wang, S., Wang, S., Su, L., Yang, K., Wang, Y., Sun, M., Zhang, L. (2024). A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(9): 10074-10082. https://doi.org/10.1609/aaai.v38i9.28871
[16] Wu, S., He, D., Wang, X., Wang, L., Dang, J. (2025). Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In Proceedings of the AAAI Conference on Artificial Intelligence, 39(2): 1601-1609. https://doi.org/10.1609/aaai.v39i2.32152
[17] Krugmann, J.O., Hartmann, J. (2024). Sentiment analysis in the age of generative AI. Customer Needs and Solutions, 11: 3. https://doi.org/10.1007/s40547-024-00143-4
[18] Vamsidhar, D., Desai, P., Shahade, A.K., Patil, S., Deshmukh, P.V. (2025). Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis. Scientific Reports, 15(1): 25440. https://doi.org/10.1038/s41598-025-09000-3
[19] Cai, Y., Li, X., Zhang, Y., Li, J., Zhu, F., Rao, L. (2025). Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. Scientific Reports, 15(1): 2126. https://doi.org/10.1038/s41598-025-85859-6
[20] Zhu, H., Shen, Z., Dai, C., Yu, Z. (2025). Multimodal large language model enhancement network for multimodal sentiment analysis. Multimedia Systems, 31(5): 335. https://doi.org/10.1007/s00530-025-01914-2
[21] Zhang, J., Wu, X., Huang, C. (2023). AdaMoW: Multimodal sentiment analysis based on adaptive modality-specific weight fusion network. IEEE Access, 11: 48410-48420. https://doi.org/10.1109/ACCESS.2023.3276932
[22] Silva, N., Cardoso, P.J., Rodrigues, J.M. (2024). Multimodal sentiment classifier framework for different scene contexts. Applied Sciences, 14(16): 7065. https://doi.org/10.3390/app14167065
[23] Alfreihat, M., Almousa, O.S., Tashtoush, Y., AlSobeh, A., Mansour, K., Migdady, H. (2024). Emo-SL framework: Emoji sentiment lexicon using text-based features and machine learning for sentiment analysis. IEEE Access, 12: 81793-81812. https://doi.org/10.1109/ACCESS.2024.3382836
[24] Subbaiah, B., Murugesan, K., Saravanan, P., Marudhamuthu, K. (2024). An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network. Artificial Intelligence Review, 57(2): 34. https://doi.org/10.1007/s10462-023-10645-7
[25] Liu, Z., Braytee, A., Anaissi, A., Zhang, G., Qin, L., Akram, J. (2024). Ensemble pretrained models for multimodal sentiment analysis using textual and video data fusion. In Companion Proceedings of the ACM Web Conference, pp. 1841-1848.
[26] Ren, J. (2024). Multimodal sentiment analysis based on BERT and ResNet. arXiv preprint arXiv:2412.03625. https://doi.org/10.48550/arXiv.2412.03625
[27] Wang, R., Zhuang, P. (2025). A strategy for network multi-layer information fusion based on multimodel in user emotional polarity analysis. International Journal of Cognitive Computing in Engineering, 6: 120-130. https://doi.org/10.1016/j.ijcce.2024.11.007
[28] Wang, S., Cai, G., Lv, G. (2025). Aspect-level multimodal sentiment analysis based on co-attention fusion. International Journal of Data Science and Analytics, 20(2): 903-916. https://doi.org/10.1007/s41060-023-00497-3
[29] Farhadipour, A., Ranjbar, H., Chapariniya, M., Vukovic, T., Ebling, S., Dellwo, V. (2025). Multimodal emotion recognition and sentiment analysis in multi-party conversation contexts. arXiv preprint arXiv:2503.06805. https://doi.org/10.48550/arXiv.2503.06805
[30] Zhang, S., Liu, J., Jiao, Y., Zhang, Y., Chen, L., Li, K. (2025). A multimodal semantic fusion network with cross-modal alignment for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications, and Applications, 21(10): 1-22. https://doi.org/10.1145/3744648
[31] Sun, Y., Chen, W., Dou, Y. (2025). UniEmotion: A unified framework for multimodal emotion recognition with iterative consensus-based training. In Proceedings of the 33rd ACM International Conference on Multimedia, pp. 13856-13863. https://doi.org/10.1145/3746027.3762010
[32] Malik, S.S., Ilyas, M., Haq, Y.U., Sana, R., Razzaq, M.S., Maqbool, F., Pathan, M.S. (2025). Multi-modal emotion detection and sentiment analysis. IEEE Access, 13: 59790-59810. https://doi.org/10.1109/ACCESS.2025.3552475
[33] Hossain, M.S., Hossain, M.M., Chaki, S., Mridha, M.F., Rahman, M.S., Moni, M.A. (2025). Dimension-wise gated cross-attention for multimodal sentiment analysis. In Companion Proceedings of the ACM Web Conference, pp. 1979-1987. https://doi.org/10.1145/3701716.3718381
[34] Cherukuri, B.R. (2025). Enhanced trimodal emotion recognition using multibranch fusion attention with epistemic neural networks and Fire Hawk optimization. Journal of Machine and Computing, 5(1): 58-75. https://doi.org/10.53759/7669/jmc202505005
[35] Wang, X., Nourmohammadi, S. (2025). A novel framework for sentiment classification employing Bi-GRU optimized by enhanced human evolutionary optimization algorithm. Scientific Reports, 15(1): 17038. https://doi.org/10.1038/s41598-025-01516-y
[36] Almadhor, A., Belhaj, S., Alsubai, S., Baili, J., Hejaili, A.A., Gregus, M., Ivanochko, I. (2026). Cross-corpus language-independent speech emotion recognition using hybrid deep learning framework. Complex & Intelligent Systems, 12(3): 107. https://doi.org/10.1007/s40747-026-02227-1
[37] Duong, H.M., Ghosh, R., Nguyen, C.H., Levin, E., Gary, T., Nguyen, L. (2026). SentiFuse: Deep multi-model fusion framework for robust sentiment extraction. arXiv preprint arXiv:2602.01447. https://doi.org/10.48550/arXiv.2602.01447
[38] Selvi, M., SVN, S.K. (2026). An intelligent bi-directional gate recurrent neural network based hybrid deep. International Journal of Computational Intelligence

Systems, 19(1): 114. https://doi.org/10.1007/s44196-026-01244-9
[39] Mali, S. (2023). Multimodal Dataset for Sentiment Analysis and Classification. Mendeley Data, V1. https://doi.org/10.17632/b7z68nykxt.1
[40] Nayeem, M.D., Rafa, Z., Nova, T.T., Rahman, Y., Pathan, A.M., Sultana, N. (2025). A multimodal Bangla text–audio dataset for sentiment analysis. Mendeley Data, V1. https://doi.org/10.17632/5yb4jjzrx3.1
[41] Ai, Y., Chu, S., Wang, J., Xu, N. (2025). Enhancing elderly care services through integrated sentiment analysis and knowledge reasoning: A deep learning approach. International Journal of Cognitive Computing in Engineering, 6: 477-494. https://doi.org/10.1016/j.ijcce.2025.04.003
[42] Li, X.C., Zhang, F., Hua, Q., Dong, C.R. (2025). A deep spatiotemporal interaction network for multimodal sentimental analysis and emotion recognition. Information Sciences, 690: 121515. https://doi.org/10.1016/j.ins.2024.121515
[43] Ruiz, G.B., Herrera, R.D.J.G. (2025). Textual emotion detection with complementary BERT transformers in a Condorcet's jury theorem assembly. Knowledge-Based Systems. 114070. https://doi.org/10.1016/j.knosys.2025.114070
[44] Colaco, S.J., Han, D.S. (2025). Scalable context-based facial emotion recognition using facial landmarks and attention mechanism. IEEE Access, 13: 20778-20791. https://doi.org/10.1109/ACCESS.2025.3534328
[45] Wang, R., Xu, D., Cascone, L., Wang, Y., Chen, H., Zheng, J., Zhu, X. (2025). Raft: Robust adversarial fusion transformer for multimodal sentiment analysis. Array. 100445. https://doi.org/10.1016/j.array.2025.100445
[46] Boitel, E., Mohasseb, A., Haig, E. (2025). MIST: Multimodal emotion recognition using DeBERTa for text, Semi-CNN for speech, ResNet-50 for facial, and 3D-CNN for motion analysis. Expert Systems with Applications, 270: 126236. https://doi.org/10.1016/j.eswa.2024.126236
[47] Uddin, M.Z., Khaksar, W., Torresen, J. (2017). Facial expression recognition using salient features and convolutional neural network. IEEE Access, 5: 26146-26161. https://doi.org/10.1109/ACCESS.2017.2777003
[48] Ferreira, P.M., Marques, F., Cardoso, J.S., Rebelo, A. (2018). Physiological inspired deep neural networks for emotion recognition. IEEE Access, 6: 53930-53943. https://doi.org/10.1109/ACCESS.2018.2870063
[49] Nakisa, B., Rastgoo, M.N., Rakotonirainy, A., Maire, F., Chandran, V. (2018). Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. IEEE Access, 6: 49325-49338. https://doi.org/10.1109/ACCESS.2018.2868361
[50] Ai, X., Sheng, V.S., Fang, W., Ling, C.X., Li, C. (2020). Ensemble learning with attention-integrated convolutional recurrent neural network for imbalanced speech emotion recognition. IEEE Access, 8: 199909-199919. https://doi.org/10.1109/ACCESS.2020.3035910

NOMENCLATURE

MFCCs	Mel Frequency Cepstral Coefficients
DCT	Discrete Cosine Transform
BERT	Bidirectional Encoder Representations from Transformers
DGCA	Dimension Wise Gated Cross-Attention
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
ENN	Epistemic Neural Network
RFN	Relaxed instance Frequency-wise Normalization
DA-STFT	Differentiable Adaptive Short-Time Fourier transform
ANN	Artificial Neural Network
EHEO	Enhanced Human Evolutionary Optimization
RF	Random Forest
AOA-HGS	Arithmetic Optimization Algorithm-Hunger Games Search
Bi-GRU	Bi-directional Gated Recurrent Networks
MSA	Multimodal Sentiment Analysis
DL	Deep Learning
ML	Machine Learning
FPR	False Positive Rate
FNR	False Negative Rate

Subscripts

$T = \{w_1, w_2, w_3, \ldots, w_n\}$	Sequence of words
[CLS]	A special type of token for global representation
[SEP]	Separator
A_i	Attention weight for modality i
$A = \{a1, a2, .. aN\}$	Set of generic attention coefficients
a	Generic element of the set A
A_i	Attention weight for modality i
H_i	Embedding of modality i
F_{fused}	Final fused feature vector
H_t, H_a, H_v	Text, audio, and visual embeddings
Attn	Attention map
z_i	Logit score for class i
$P(y_i)$	Probability of class i
$x_w(n)$	Windowing function
$x(n)$	Audio sample at time n
$w(n)$	Windowed signal
X_{phr},X_{sen}	Phrase and sentence inputs
H_p, H_{sen}	Phrase and sentence embeddings
f_p, f_{sen}	Feature vectors
m	Melscale Frequency
$V = \{v_1, v_2, v_3, \ldots v_T\}$	Video sequence frames
L_{BCE}	Binary Cross-Entropy loss
L_{CS}	Confidence Score
L_{reg}	Weight Regularization
L_{total}	Weighted three-fold loss
$\sigma(.)$	Sigmoid Activation Function
$\tanh(.)$	Hyperbolic tangent activation
$\odot$	Elementwise product